Bayesian computation project

Charles Dufour

EPFL

Spring semester 2019

Table of contents

Framework implementations and limits

Optimization

- Gradient descent
- Linear search gd
- Wolfe cond gd
- Stochastic gd
- Newton gd (slow)

Framework implementations and limits

Optimization

- Gradient descent
- Linear search gd
- Wolfe cond gd
- Stochastic gd
- Newton gd (slow)

Approximation

- Laplace
- GVA (unstable)

Framework implementations and limits

Optimization

- Gradient descent
- Linear search gd
- Wolfe cond gd
- Stochastic gd
- Newton gd (slow)

Approximation

- Laplace
- GVA (unstable)

Sampling

- MH random walk
- MALA
- IS, RS
- ~~Gibbs~~
- MH within Gibbs

Structure

Figure: Hourly wage and features in the USA, May 1985

ED	SOUTH	NONWH	HISP	FE	MARR	MARRFE	EX	EXSQ	UNION	LNWAGE	AGE	MANUF	CONSTR	MANAG	SALES	CLER	SERV	PROF
10	0	0	0	0	1	0	27	729	0	2.1972	43	0	1	0	0	0	0	0
12	0	0	0	0	1	0	20	400	0	1.7047	38	0	0	0	1	0	0	0
12	0	0	0	1	0	0	4	16	0	1.3350	22	0	0	0	1	0	0	0
12	0	0	0	1	1	1	29	841	0	2.3514	47	0	0	0	0	1	0	0
12	0	0	0	0	1	0	40	1600	1	2.7080	58	0	1	0	0	0	0	0

Structure

Figure: Hourly wage and features in the USA, May 1985

ED	SOUTH	NONWH	HISP	FE	MARR	MARRFE	EX	EXSQ	UNION	LNWAGE	AGE	MANUF	CONSTR	MANAG	SALES	CLER	SERV	PROF
10	0	0	0	0	1	0	27	729	0	2.1972	43	0	1	0	0	0	0	0
12	0	0	0	0	1	0	20	400	0	1.7047	38	0	0	0	1	0	0	0
12	0	0	0	1	0	0	4	16	0	1.3350	22	0	0	0	1	0	0	0
12	0	0	0	1	1	1	29	841	0	2.3514	47	0	0	0	0	1	0	0
12	0	0	0	0	1	0	40	1600	1	2.7080	58	0	1	0	0	0	0	0

Purpose

- Predict exactly the revenue
- Predict if revenue above mean

Structure

Figure: Hourly wage and features in the USA, May 1985

ED	SOUTH	NONWH	HISP	FE	MARR	MARRFE	EX	EXSQ	UNION	LNWAGE	AGE	MANUF	CONSTR	MANAG	SALES	CLER	SERV	PROF
10	0	0	0	0	1	0	27	729	0	2.1972	43	0	1	0	0	0	0	0
12	0	0	0	0	1	0	20	400	0	1.7047	38	0	0	0	1	0	0	0
12	0	0	0	1	0	0	4	16	0	1.3350	22	0	0	0	1	0	0	0
12	0	0	0	1	1	1	29	841	0	2.3514	47	0	0	0	0	1	0	0
12	0	0	0	0	1	0	40	1600	1	2.7080	58	0	1	0	0	0	0	0

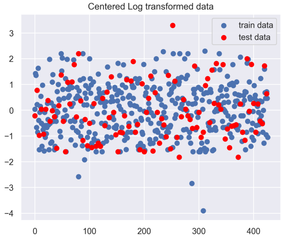
Purpose

- Predict exactly the revenue
- Predict if revenue above mean

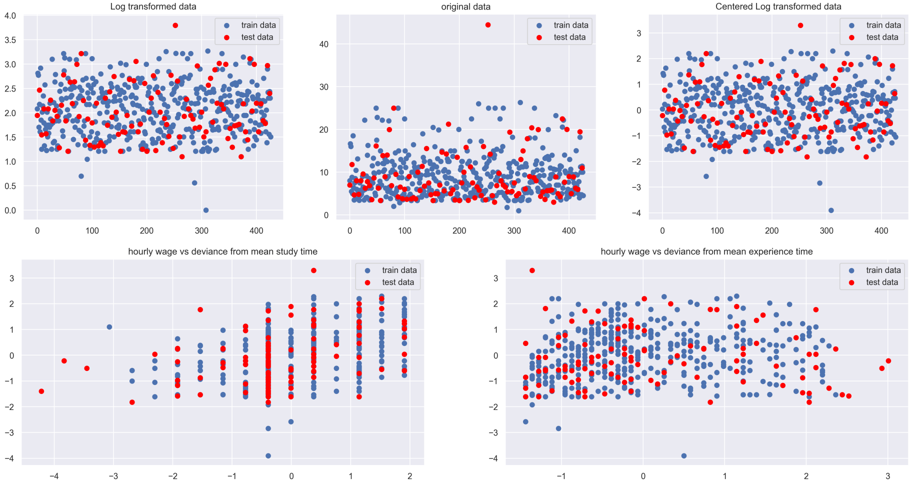
Features dropped due to high correlation

- AGE
- EXSQ

Visualization



Visualization



Models

3 models implemented:

- Gaussian model

$$Y|\beta,\sigma \sim \mathcal{N}\left(X\beta,\sigma^2\right) \qquad \beta \sim \mathcal{N}_d(\vec{0},3^2 I),\ \sigma \sim exp(2)$$

Models

3 models implemented:

- Gaussian model

$$Y|\beta,\sigma \sim \mathcal{N}\left(X\beta,\sigma^2\right) \qquad \beta \sim \mathcal{N}_d(\vec{0},3^2 I),\ \sigma \sim exp(2)$$

- Student model

$$Y|\beta,\nu \sim X\beta + t_\nu \qquad \beta \sim \mathcal{N}_d(\vec{0},3^2 I),\ \nu \sim \Gamma(2,4)$$

Models

3 models implemented:

- Gaussian model

$$Y|\beta,\sigma \sim \mathcal{N}\left(X\beta,\sigma^2\right) \qquad \beta \sim \mathcal{N}_d(\vec{0},3^2 I),\ \sigma \sim exp(2)$$

- Student model

$$Y|\beta,\nu \sim X\beta + t_\nu \qquad \beta \sim \mathcal{N}_d(\vec{0},3^2 I),\ \nu \sim \Gamma(2,4)$$

- Logistic regression

$$\mathbb{P}\left(Y=1|X,\beta\right) = \frac{e^{X^T\beta}}{1+e^{X^T\beta}}, \quad \beta \sim \mathcal{N}_d(0,3^2)$$

Laplace approximation

Laplace approximation

Fit a Gaussian approximation to the unormalized posterior:

- mean: $\theta^* = \text{argmax}_\theta \ \tilde{f}(\theta|D = d)$
- covariance matrix: $\Sigma = H_\psi(\theta^*)^{-1}$

with $\psi(\theta) = -\log\left(\tilde{f}(\theta|D = d)\right)$ which will be used in the computations.

Laplace approximation

Laplace approximation

Fit a Gaussian approximation to the unormalized posterior:

- mean: $\theta^* = \text{argmax}_\theta \; \tilde{f}(\theta|D = d)$
- covariance matrix: $\Sigma = H_\psi(\theta^*)^{-1}$

with $\psi(\theta) = -\log\left(\tilde{f}(\theta|D = d)\right)$ which will be used in the computations.

Optimization routines

- Vanilla gradient descent
- Stochastic gradient descent
- Line search backtracking gradient descent
- Wolfe condition checking gradient

MH with random walk

Theory

for $i = 1$ to N **do**
  draw $\eta \sim \mathcal{N}_d(0, 1)$
  $\theta_c = \theta_n + \varepsilon\eta$
  $R = f(\theta_c|d)/f(\theta_n|d)$
  if $U(0, 1) \leq R$ **then**
    $\theta_{n+1} = \theta_c$
  else
    $\theta_{n+1} = \theta_n$
  end if
end for

Practice

- Set ε such that the acceptance rate of the proposal is between 10 and 50 percent.
- Compute everything using `expsumlog`
- Check visually the chain to determine the burn-in
- Test different initialization to detect potential silent failure

MH with Langevin correction (MALA)

Theory

As for the random walk MH algorithm except for:

- Proposal:

$$\theta_c = \theta_n + \tau \nabla \log f(\theta_n|d) + \sqrt{2\tau}\eta$$

- Acceptance ratio:

$$R = \frac{f(\theta_c|d)q(\theta_n|\theta_c)}{f(\theta_n|d)q(\theta_c|\theta_n)}$$

Practice

- As before but be more careful with tuning the step size τ
- $$q(x, x') \propto \exp\left(\frac{||x' - x - \tau\nabla \log f(x|d)||_2^2}{-4\tau}\right)$$
- Biggest challenge: implement computation of gradient in efficient manner

Prediction

Marginalize out θ

$$f(y|d) = \int f(y|\theta)f(\theta|d)d\theta \approx \frac{1}{N}\sum_{i=1}^{N} f(y|\theta_i)$$

Where $\theta_i \sim f(\theta|d)$ comes from one of the sampling methods above.

Prediction

Marginalize out θ

$$f(y|d) = \int f(y|\theta)f(\theta|d)d\theta \approx \frac{1}{N}\sum_{i=1}^{N} f(y|\theta_i)$$

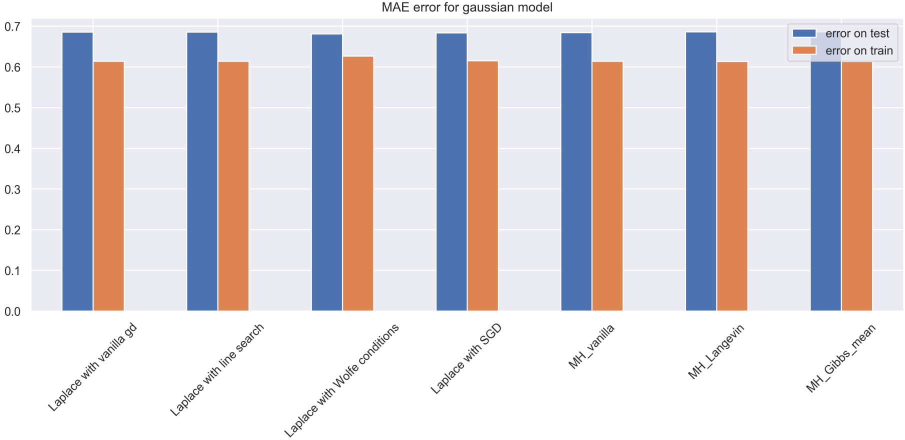
Where $\theta_i \sim f(\theta|d)$ comes from one of the sampling methods above.

First estimate θ then predict mean

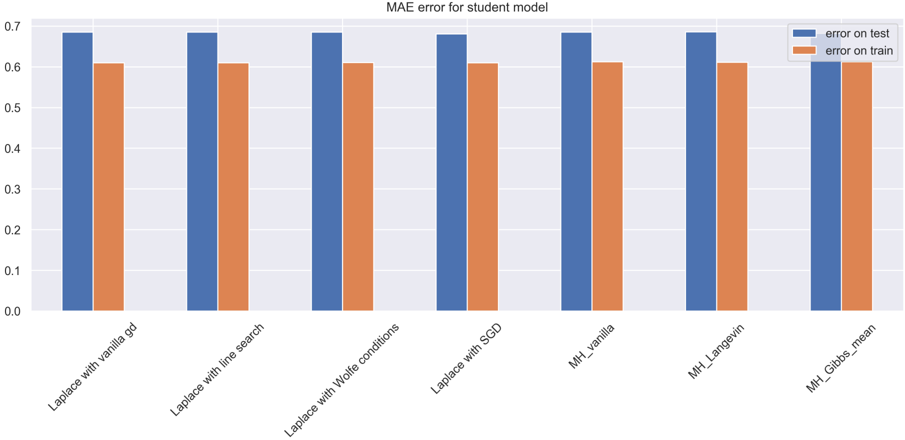
In the gaussian case:

$$\begin{aligned}
\mathbb{E}\left[f(y|d)\right] &= \iint yf(y|\theta)f(\theta|d)d\theta dy \\
&= \int \left(\int yf(y|\theta)dy\right) d\theta \\
&= \int x^T\theta \cdot f(\theta|d)d\theta = \mathbb{E}_{\theta\sim f(\theta|d)}\left[x^T\theta\right]
\end{aligned}$$

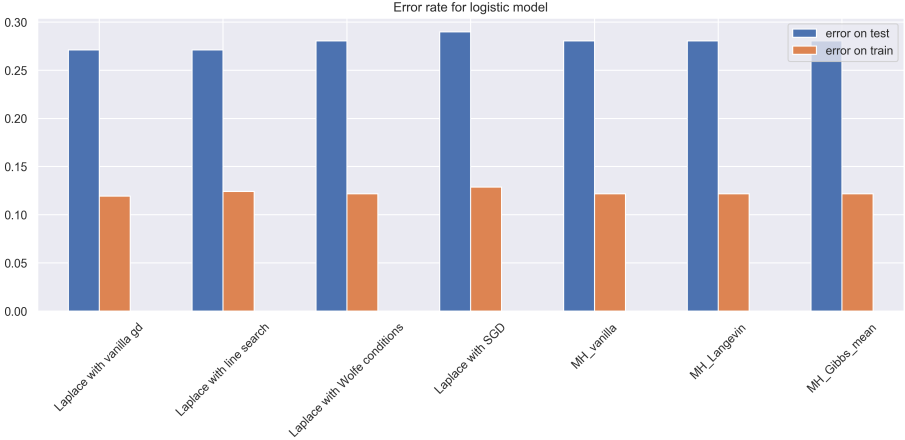
Accuracy of the Gaussian model



Accuracy of the Student model



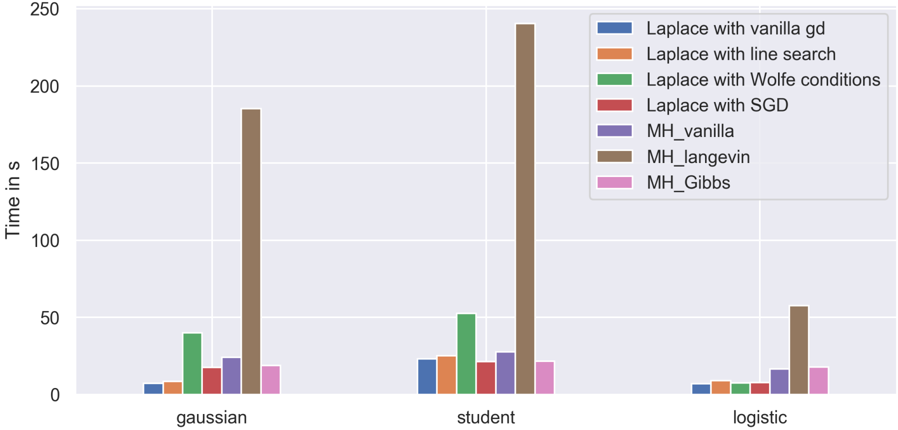
Accuracy of the Logistic model



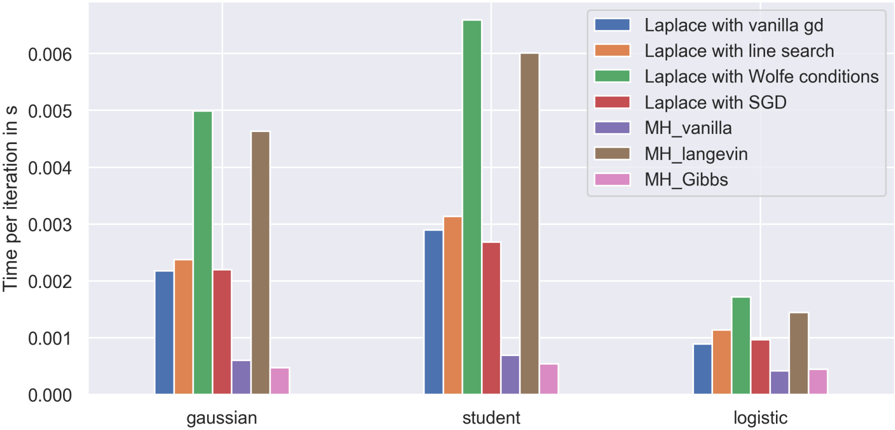
How to compare the methods

All the methods seem to perform equally when looking at the prediction, how do we choose between them ?

Comparison in term of total time



Comparison in term of time per iteration



Conclusion

Modelization

- Simpler methods and models performed the best
- Relationship highly non-linear

Conclusion

Modelization

- Simpler methods and models performed the best
- Relationship highly non-linear

Improvement

- Tuning of hyper-parameters
- Feature engineering

Conclusion

Modelization

- Simpler methods and models performed the best
- Relationship highly non-linear

Improvement

- Tuning of hyper-parameters
- Feature engineering
- Gamma model
- Classification in multiple ordered classes

Conclusion

Modelization

- Simpler methods and models performed the best
- Relationship highly non-linear

Improvement

- Tuning of hyper-parameters
- Feature engineering
- Gamma model
- Classification in multiple ordered classes
- More robust and faster module to use more advanced techniques

References

- E. R. Berndt
  The practice of econometrics : classic and contemporary. Addison-Wesley Pub. Co., 1991
- GitHub repository
  https://github.com/dufourc1/Bayesian computation