# Bayesian-Computation Framework, implementation and application

MATH-435 Project

William Cappelletti

EPFL - Instute of Mathematics

*william.cappelletti@epfl.ch*

June 18, 2019

# Overview

# General setting

**Content**

Hourly wages of 550 randomly selected employed workers from a 1978 population survey conducted by the USA government.

# General setting

**Content**

Hourly wages of 550 randomly selected employed workers from a 1978 population survey conducted by the USA government.

**The variables**

- The response:
  - Logarithm of hourly wage.
- The explanatory variables:
  - 3 continuos variables:
    - Age
    - Experience
    - Education
  - 13 categorical variables.

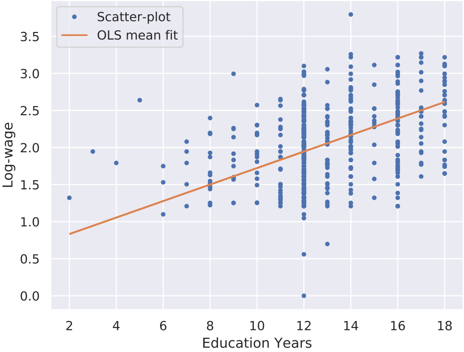
# General setting

**Content**

Hourly wages of 550 randomly selected employed workers from a 1978 population survey conducted by the USA government.

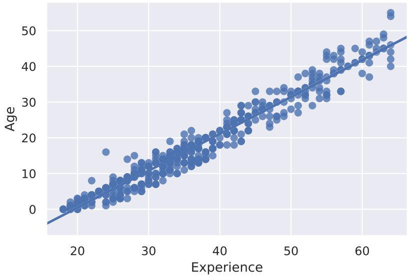
**The variables**

- The response:
  - Logarithm of hourly wage.
- The explanatory variables:
  - 3 continuos variables:
    - Age
    - Experience
    - Education
  - 13 categorical variables.



# Exploration

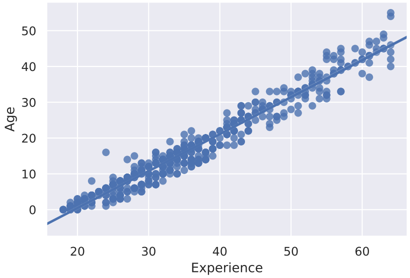
**Collinearity**




Some fatures are highly collinear. In particular 'age', 'education' and 'experience', the latter being computed as AGE - ED - 6.

# Exploration

## Collinearity




Some fatures are highly collinear. In particular 'age', 'education' and 'experience', the latter being computed as AGE - ED - 6.

## Outliers and leverage points

There are some 'peculiar' individuals, whose personal characteristics and wages strongly part from the others.

In particular a 48-years-old black man performing a clerical work and earning 0.625$ per hour, and a sales worker earning 3.195$ less than the mean on the same job category.

# Gaussian Models

## I implement three different Gaussian Linear models.

$$y = x^{\mathrm{T}}\beta + \varepsilon$$

Where $y$ is the response, $x$ the vector of covariates and the conditional densities of $\varepsilon$, $\beta$ are

$$\varepsilon \sim \mathcal{N}(0, \sigma^2), \quad \beta \sim \mathcal{N}(\mathbf{0}, \sigma_\beta^2 Id).$$

# Gaussian Models

## I implement three different Gaussian Linear models.

$$y = x^{\mathrm{T}}\beta + \varepsilon$$

Where $y$ is the response, $x$ the vector of covariates and the conditional densities of $\varepsilon$, $\beta$ are

$$\varepsilon \sim \mathcal{N}(0, \sigma^2), \quad \beta \sim \mathcal{N}(\mathbf{0}, \sigma_\beta^2 Id).$$

**Exponential prior on variance**

Two models with:

1. $\sigma \sim \exp(2)$, $\sigma_\beta = 3$;

# Gaussian Models

## I implement three different Gaussian Linear models.

$$y = x^{\mathrm{T}}\beta + \varepsilon$$

Where $y$ is the response, $x$ the vector of covariates and the conditional densities of $\varepsilon$, $\beta$ are

$$\varepsilon \sim \mathcal{N}(0, \sigma^2), \quad \beta \sim \mathcal{N}(\mathbf{0}, \sigma_\beta^2 Id).$$

**Exponential prior on variance**

Two models with:

1. $\sigma \sim \exp(2)$, $\sigma_\beta = 3$;
2. $\sigma \sim \exp(2)$, $\sigma_\beta \sim \exp(3)$.

# Gaussian Models

## I implement three different Gaussian Linear models.

$$y = x^{\mathrm{T}}\beta + \varepsilon$$

Where $y$ is the response, $x$ the vector of covariates and the conditional densities of $\varepsilon$, $\beta$ are

$$\varepsilon \sim \mathcal{N}(0, \sigma^2), \quad \beta \sim \mathcal{N}(\mathbf{0}, \sigma_\beta^2 Id).$$

**Exponential prior on variance**

Two models with:

1. $\sigma \sim \exp(2)$, $\sigma_\beta = 3$;
2. $\sigma \sim \exp(2)$, $\sigma_\beta \sim \exp(3)$.

**Gamma prior on variance**

One model with

- $\sigma \sim \Gamma(2), \ \sigma_\beta = 3$

# Gaussian Models

Gaussian models are simple, but the different priors let us give some freedom to the model.
Nonetheless, they are sensible to outliers.

# Student-t Models

## I implement two Student-t Linear models.

$$y = x^{\mathrm{T}}\beta + \sigma\varepsilon$$

Where $y$ is the response, $x$ the covariates vector and the conditional densities of $\varepsilon$, $\beta$ are

$$\varepsilon \sim t(df), \quad \beta \sim \mathcal{N}(\mathbf{0}, 9 \cdot Id).$$

# Student-t Models

## I implement two Student-t Linear models.

$$y = x^{\mathrm{T}}\beta + \sigma\varepsilon$$

Where $y$ is the response, $x$ the covariates vector and the conditional densities of $\varepsilon$, $\beta$ are

$$\varepsilon \sim t(df), \quad \beta \sim \mathcal{N}(\mathbf{0}, 9 \cdot Id).$$

**Gamma prior only on df**

- $df/2 \sim \Gamma(2)$, $\sigma = 3$;

# Student-t Models

## I implement two Student-t Linear models.

$$y = x^{\mathrm{T}}\beta + \sigma\varepsilon$$

Where $y$ is the response, $x$ the covariates vector and the conditional densities of $\varepsilon$, $\beta$ are

$$\varepsilon \sim t(df), \quad \beta \sim \mathcal{N}(\mathbf{0}, 9 \cdot Id).$$

**Gamma prior only on df**

- $df/2 \sim \Gamma(2)$, $\sigma = 3$;

**Gamma prior on both df and scale**

- $df/2 \sim \Gamma(2)$, $\sigma/2 \sim \Gamma(2)$;

# Student Models

Student models are more robust to outliers and the prior on the degrees of freedom let the model find by itself how much weight give to the tails. Adding on top of that a prior distribution for the scale lets the width of the tails adapts by themselves.

# Approximations

I use different methods to approximate the posterior distribution function

$$f(\theta|d),$$

where $\theta$ is the parameter vector and $d$ are the data.

# Approximations

I use different methods to approximate the posterior distribution function

$$f(\theta|d),$$

where $\theta$ is the parameter vector and $d$ are the data.

1. Laplace approximation
2. Importance and Rejection sampling
3. Metropolis Hastings
4. Gaussian Variational Approximation

# Laplace Approximation

Suppose $\log \tilde{f}(\theta|d)$ is the unnormalized log-posterior of our model.

## Definition

The Laplace approximation is given by

$$f(\theta|d) \approx C \exp\left\{-\frac{1}{2}(\theta - \theta_{\mathrm{MAP}})^{\mathrm{T}}\beta(\theta - \theta_{\mathrm{MAP}})\right\},$$

where $\theta_{\mathrm{MAP}}$ is the Maximum A Posteriori estimate of the parameters and $\beta = -H_\theta\left[\log \tilde{f}(\theta_{\mathrm{MAP}}|d)\right]''$.

# Laplace Approximation

Suppose $\log \tilde{f}(\theta|d)$ is the unnormalized log-posterior of our model.

## Definition

The Laplace approximation is given by

$$f(\theta|d) \approx C \exp\left\{-\frac{1}{2}(\theta - \theta_{\mathrm{MAP}})^{\mathrm{T}}\beta(\theta - \theta_{\mathrm{MAP}})\right\},$$

where $\theta_{\mathrm{MAP}}$ is the Maximum A Posteriori estimate of the parameters and $\beta = -H_\theta\left[\log \tilde{f}(\theta_{\mathrm{MAP}}|d)\right]''$.

We can obtain $\theta_{\mathrm{MAP}}$ and $\beta$ using Stochastic Gradient descent by using an automatic differentiator like Autograd.

**Definition**

We can sample from the posterior distribution $f(\theta|d)$ by using a proposal ditribution $h(\theta)$.

For each sample $\theta_i$ we store its *weight*

$$w_i = \frac{\tilde{f}(\theta_i|d)}{h(\theta_i)},$$

and we use it either to compute de desired value (IS), either to give us a treshold to reject the samples (RS).

# Metropolis Hastings

**Definition**

The *Metropolis-Hastings* algorithm let us create a Monte-Carlo-Markov-Chain which stationary distribution is the posterior we want to sample from.

# Metropolis Hastings

## Definition

The *Metropolis-Hastings* algorithm let us create a Monte-Carlo-Markov-Chain which stationary distribution is the posterior we want to sample from.

I implement a *random-walk-MH* algorithm, which generate a proposal using gaussian noise and then choose randomly, based on the acceptance ratio, wheter to keep the new sample.

# Gaussian Variational Approximation

Similarly to Laplace, it approximates the posterior distribution with a Gaussian.

**Definition**

The GVA approximation is given by

$$f(\theta|d) = exp\{-\phi(\theta)\},$$

by reparametrizing the random variable $\theta = \mu + \exp(L)\eta$, where $\eta$ is a standard gaussian.

# Gaussian Variational Approximation

Similarly to Laplace, it approximates the posterior distribution with a Gaussian.

## Definition

The GVA approximation is given by

$$f(\theta|d) = exp\{-\phi(\theta)\},$$

by reparametrizing the random variable $\theta = \mu + \exp(L)\eta$, where $\eta$ is a standard gaussian.

The parameters $\mu$ and $L$ are obtained by maximizing the ELBO.

$$\mathrm{ELBO} \approx \frac{1}{l}\sum_{i=1}^{l} \log\left(f(\mu + \exp(L)\eta_i|d)\right) + \frac{p}{2}\log(2\pi e) + \mathrm{Tr}(L),$$

where $\eta_i$, $i = 1, ..., l$ are random samples from a standard gaussian.
I implement it using Autograd and gradient ascent.

# My framework

I implement all models and methods in a OOP fashion.

# My framework

I implement all models and methods in a OOP fashion.

### Parameter

The class encoding a parameter we want to study. It contains the prior values and two main functions, `log_prior` and `proposal`.

# My framework

I implement all models and methods in a OOP fashion.

### Parameter

The class encoding a parameter we want to study. It contains the prior values and two main functions, `log_prior` and `proposal`.

### LinearModel

The class encoding the linear model. It contains the dataset, the parameters and its `log_likelihood`, which uses to compute its `log_unnorm_posterior`.

# My framework

I implement all models and methods in a OOP fashion.

**Parameter**

The class encoding a parameter we want to study. It contains the prior values and two main functions, `log_prior` and `proposal`.

**LinearModel**

The class encoding the linear model. It contains the dataset, the parameters and its `log_likelihood`, which uses to compute its `log_unnorm_posterior`.

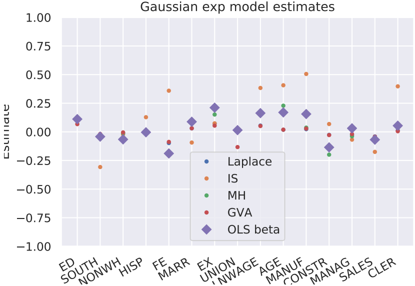
Thanks to this implementation, once the model is set, all methods can use it.

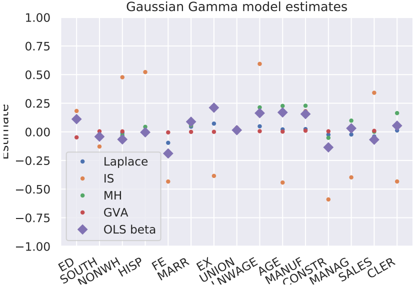
# Example

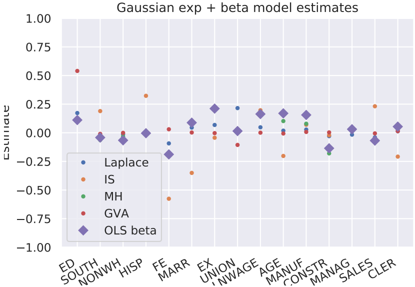
## Example

```
from bayesian.model import StudentModel_scaleVar
from bayesian.parameter import NormalPar, GammaPar
from bayesian.methods import importance_sampling

beta = NormalPar(mu = np.zeros(X.shape[1]), scale = 3)
df = GammaPar(2., 2.)
scale = GammaPar(2., 2.)

model = StudentModel_scaleVar(y, X, beta, df, scale)

beta.set_proposal(GaussianProposal, .4)
df.set_proposal(GammaProposal, 2., 2.)

n_samples = 10000

samples, log_weights = importance_sampling(model, n_samples)
```

# Comments on Gaussian Models

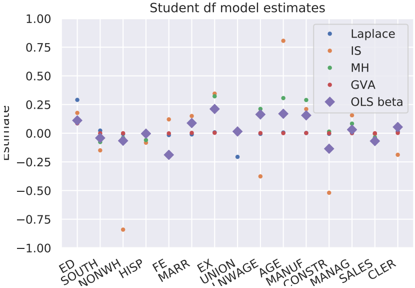


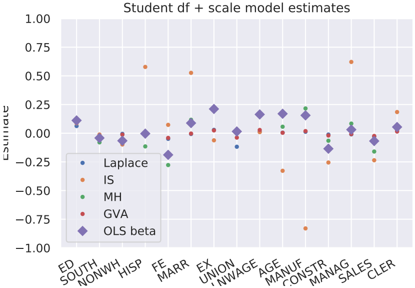




All models perform similarly. The one with an exp prior on $\beta$ variance is more numerically unstable.

# Comments on Student Models

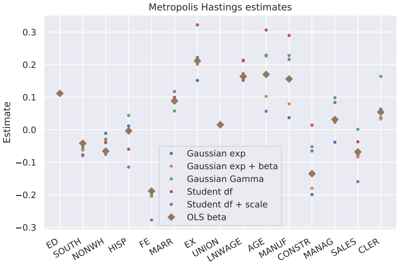


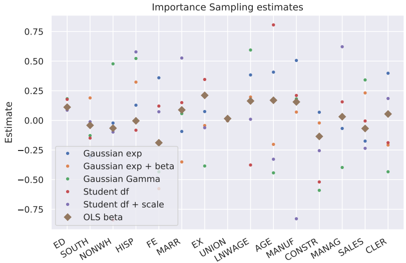



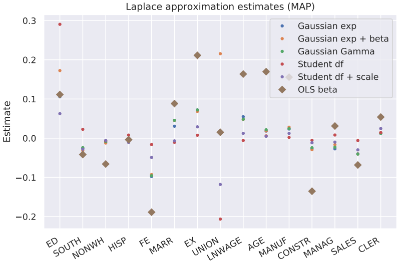

Student model are more stable, probably beacause of the presence of some outliers.

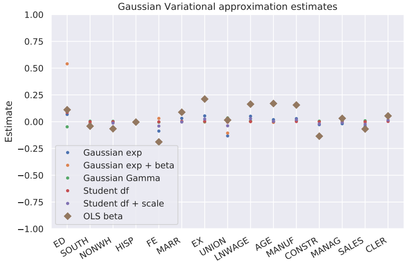
The model with gamma prior on both the degrees of fredom and the scale is even better.

# Stability of methods

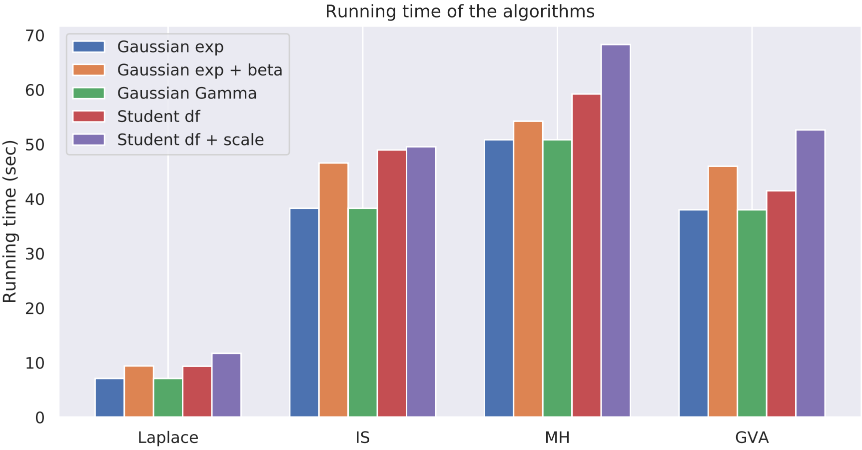


# Running times



# Conclusion

## My favourite Model

Amongst the one I tested, the model with the least computational issues, for this dataset, is the *student model with gamma prior on both the degrees of freedom and the scale*.

- It would be interesting to test for a sparcity inducing prior on $\beta$ on top of this model.

# Conclusion

## My favourite Model

Amongst the one I tested, the model with the least computational issues, for this dataset, is the *student model with gamma prior on both the degrees of freedom and the scale*.

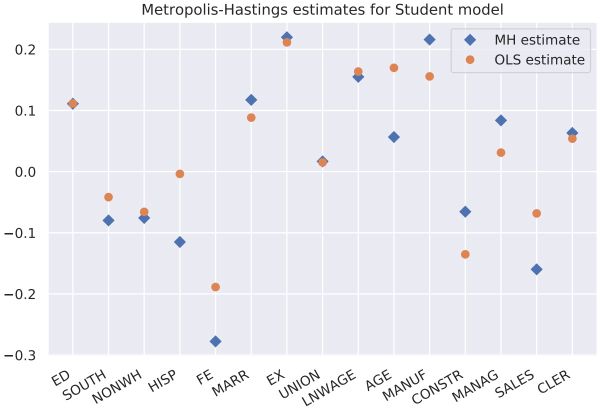
- It would be interesting to test for a sparcity inducing prior on $\beta$ on top of this model.

## Best Method

Between the four Approximation methods implemented, the best one is *Metropolis-Hastings*, both in terms of stability and “ease of use”.

- A further improvement could come by implementing other versions of MH, such as MH-within Gibbs.

# Conclusion




- Estimated degrees of freedom: 3.31
- Estimated scale : 0.82

## Conclusion

#ThreeQuestionsChallenge