HMM project: application to genetics

Léo Houairi, Yao Pacome Kouame, Théo Lorthios and Sonali Mohan Patekar

ENSAE Paris

January 6, 2022

ENSAE
IP PARIS

Overview

Introduction

Our study is based on the article of **"Parseq: reconstruction of microbial transcription landscape from RNA-Seq read counts using state-space models"**.

It develops a way to reconstruct the transcription level accross the genom (**hidden variable**) based on the read counts (**observed variable**) produced by modern RNA-sequencing techniques.

Project organization

- Implement the model and generate data
- Implement a FK and use the bootstrap filter
- Use the PMMH algorithm to perform bayesian inference

The variables

To model contains 3 main variables:

- $(y_t)_{t>0}$ is the sequence of observations, the read counts.
- $(u_t)_{t>0}$ is the underlying expression level.
- $(s_t)_{t>0}$ is a local scaling variable.

The hidden variable X_t is a vetor (u_t, s_t). It is deterministically initizalied at $(0, 0)$.

Markov chains for s_t

A piecewise constant Gamma:

$$k_s(s_{t+1}|s_t) = \alpha_s \times \delta_{s_t}(s_{t+1}) + (1-\alpha_s) \times \Gamma(s_{t+1}; shape = \kappa_s, scale = \kappa_s)$$

Markov chain for u_t

A complicated mixture with seven types of moves:

$$k_u(u_{t+1}|u_t) = \mathbb{1}_{\{u_t=0\}} \times \Big[(1-\eta) \times \delta_0(u_{t+1}) + \eta \times \mathcal{E}(u_{t+1};\zeta)\Big]$$
$$+ \mathbb{1}_{\{u_t>0\}} \times \Big[\alpha \times \delta_{u_t}(u_{t+1}) + \beta \times \mathcal{E}(u_{t+1};\zeta) + \beta_0 \times \delta_0(u_{t+1})$$
$$+ \gamma_u \mathbb{1}_{\{u_{t+1}>u_t\}} \left(u_t + \mathcal{E}(Z;\frac{\lambda_u}{u_t})\right) + \gamma_d \mathbb{1}_{\{u_{t+1}<u_t\}} \left(u_t - \mathcal{E}(Z;\frac{\lambda_d}{u_t})\right)\Big]$$

With $\alpha + \beta + \beta_0 + \gamma_u + \gamma_d = 1$.

Emission model - First version

Two intermediary variables:

- $a_t \sim \Gamma(\kappa, \theta)$ (the amplification coefficient)
- $x_t \sim \mathcal{P}(\frac{u_t s_t}{\kappa\theta})$ (the number of molecules)

Then the emission model has density:

$$e(y_t|x_t, a_t) = (1 - \epsilon_b - \epsilon_0) \times \mathcal{P}(y_t; x_t \times a_t) + \epsilon_b \times \mathcal{P}_{-\{0\}}(y_t; a_t) + \epsilon_0 \times \mathcal{U}(y_t; 0...b)$$

This version was useful for generating the data.

Emission model - Second version

Without those intermediary variable, it writes:

$$e(y_t|u_t, s_t) = (1 - \epsilon_b - \epsilon_0) \times \sum_{x_t=0}^{\infty} \mathcal{P}(x_t; \frac{u_t s_t}{\kappa\theta}) \times \mathcal{NB}(y_t; \kappa, \frac{x_t\theta}{x_t\theta + 1})$$
$$+ \epsilon_b \times \mathcal{NB}_{-\{0\}}(y_t; \kappa, \frac{\theta}{\theta + 1})$$
$$+ \epsilon_0 \times \mathcal{U}(y_t; 0...b)$$

This version was used to compute the log likelihood.

Hidden variable and observations

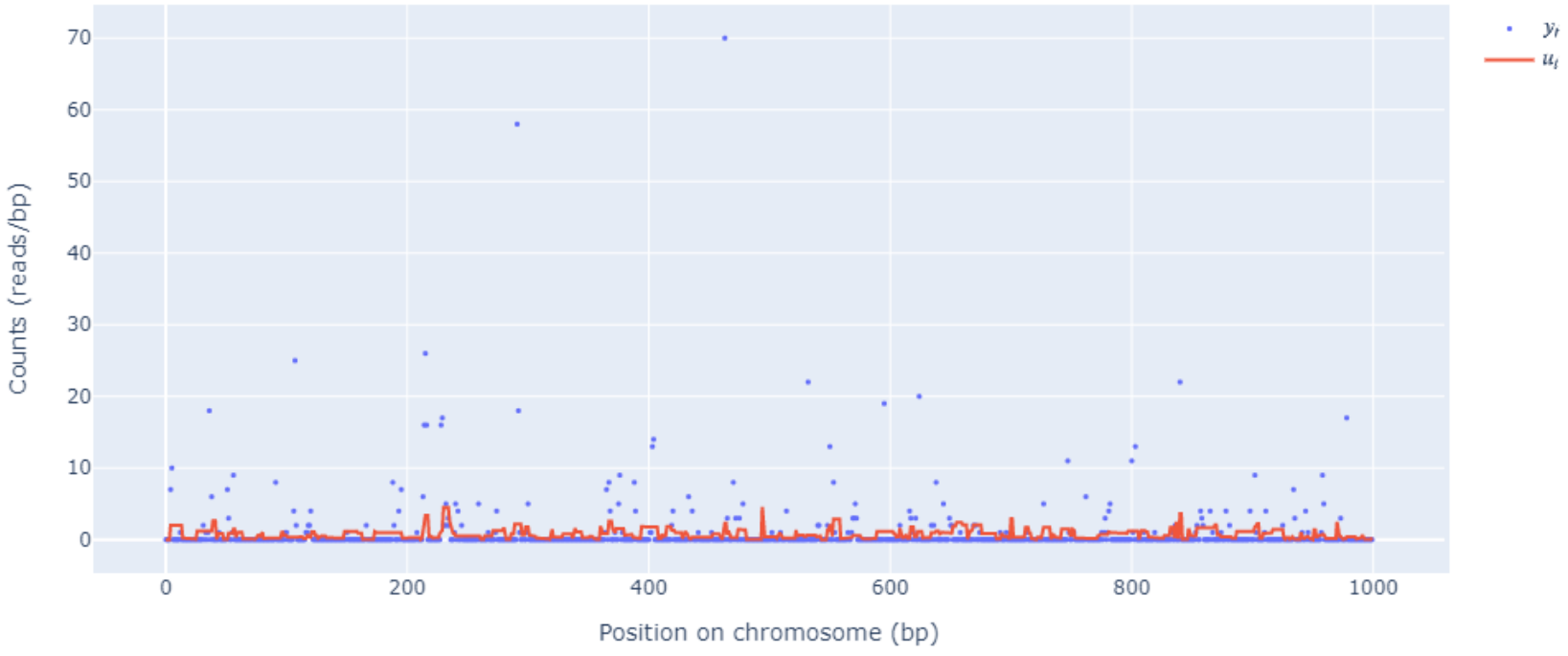

Figure: The expression level (u_t) and the observations (y_t) in the generated data

Zoom on the hidden variable

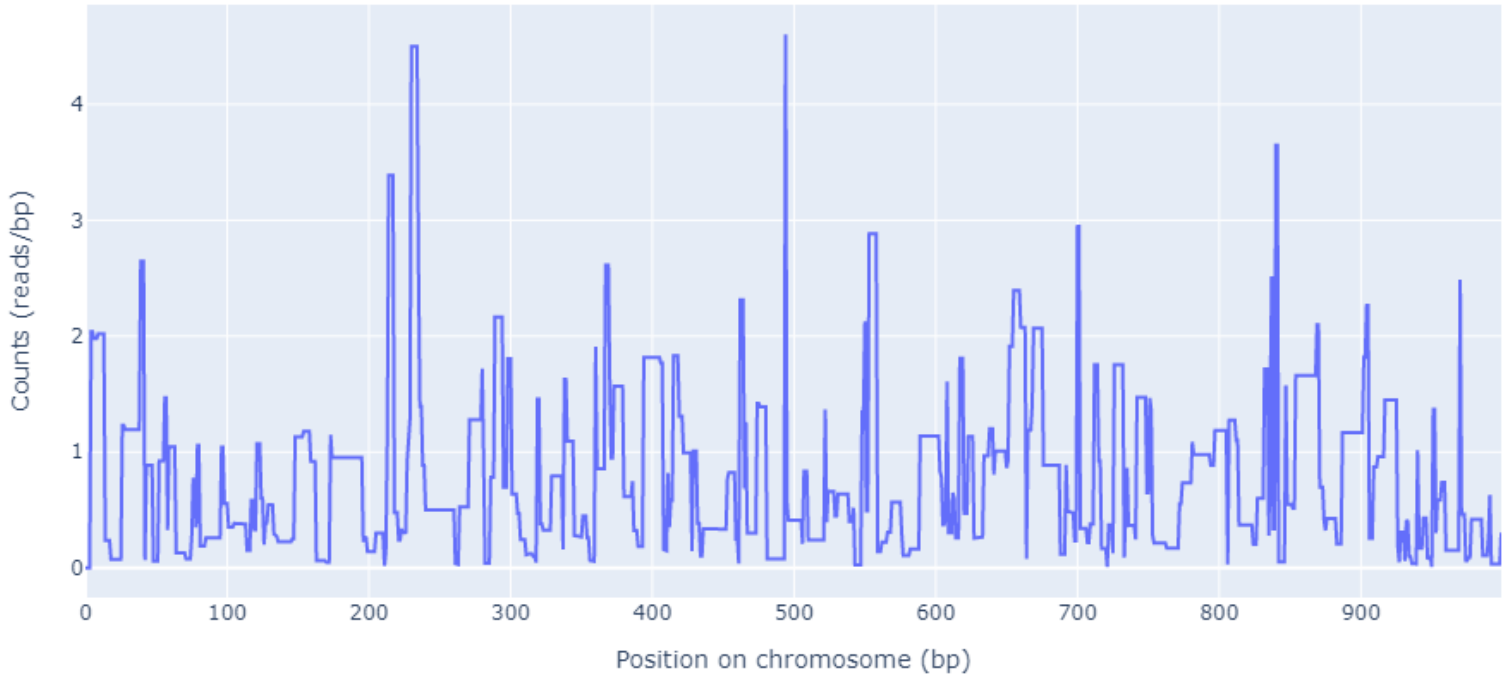

Figure: The expression level u_t in the generated data

Results obtained with 10 runs

100 particles were used in each bootstrap filter.

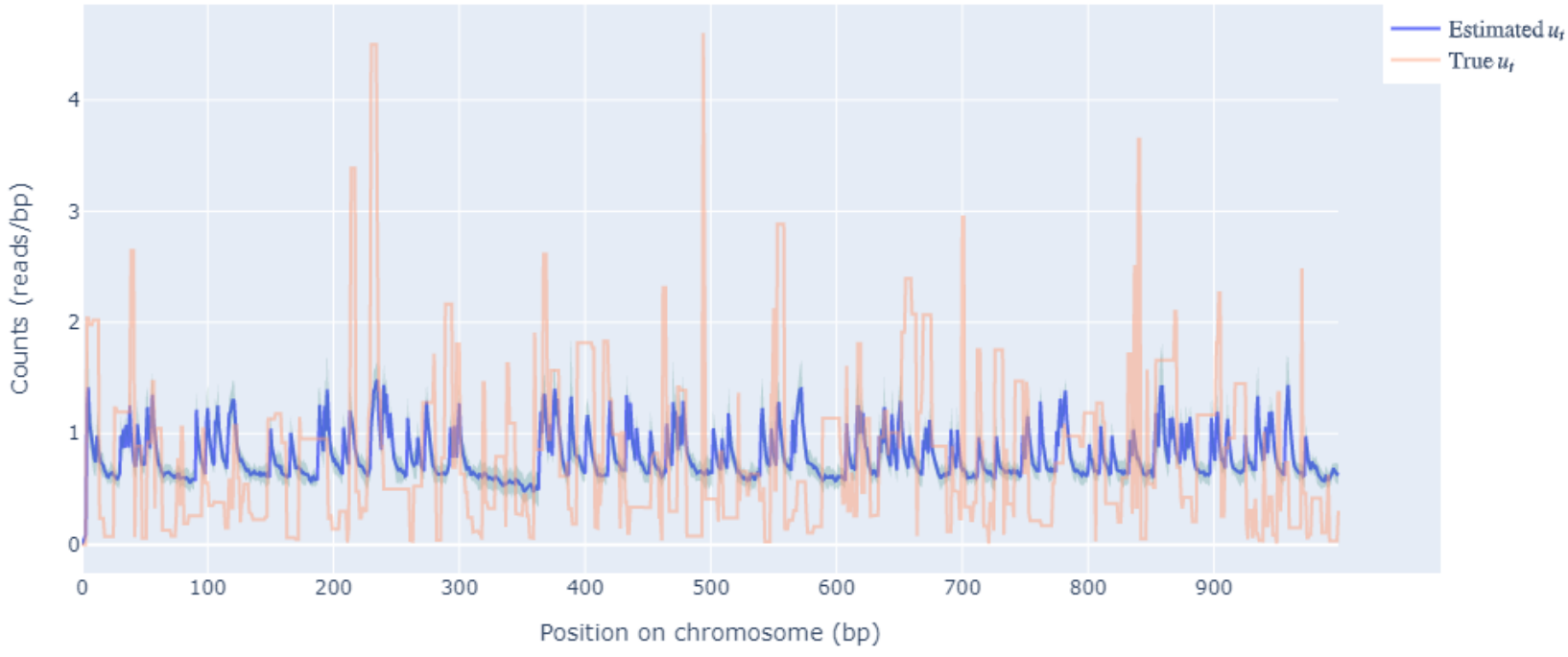

Figure: Comparison of the bootstrap filter with the real u_t

On a subset of parameters

The model includes **16 parameters** but we used PMMH on only two, because:

- Some of them can be estimated on the data.
- A lot of them are probabilities, and would require a Dirichlet distribution.

Priors:

$$\eta \sim Beta(1, 100)$$
$$\zeta \sim \mathcal{E}(1)$$

Results for η and ζ

1000 iterations and 20 particles (about 6 hours of running time).

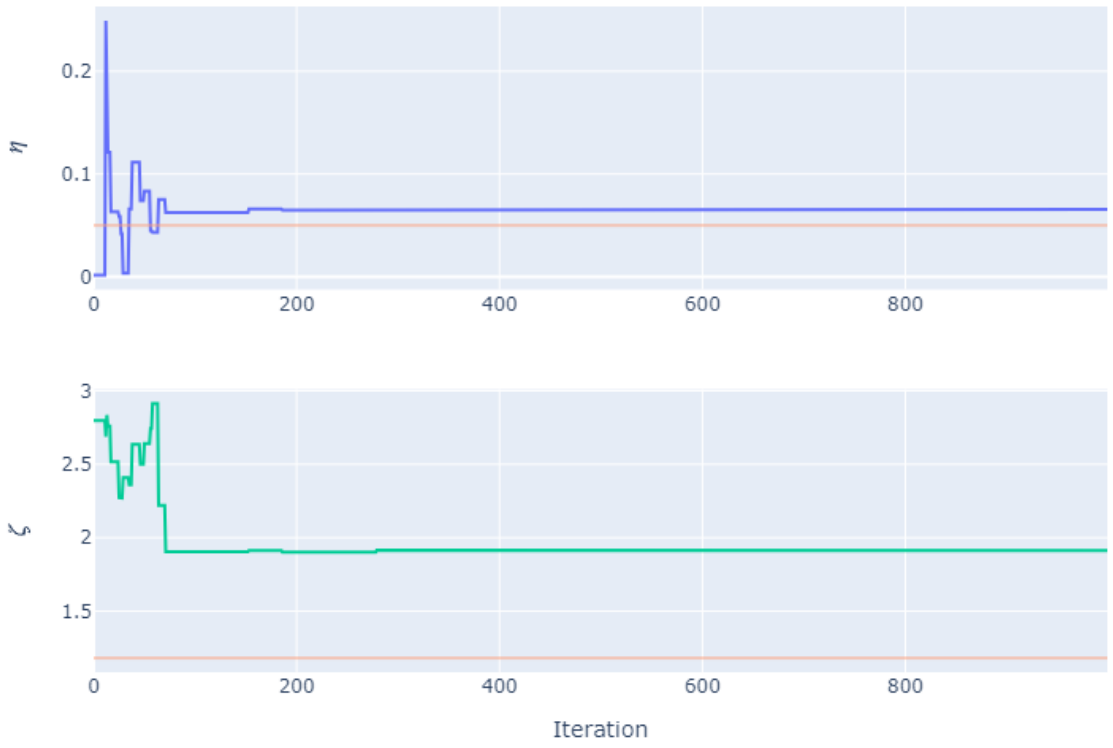

Figure: MCMC traces obtained using the PMMH algorithm

Conclusion

- Successful implementation of the bootstrap filter and the PMMH algorithm (using the **particles** package)
- But we did not have the time for the tuning of the algorithm
- It would be interesting to do the bayesian inference on the parameters representing probabilities

The end

Thank you for your attention !