Tempered Adaptive Multiple Importance Sampling for Galaxy parameter estimation

Grégoire Aufort

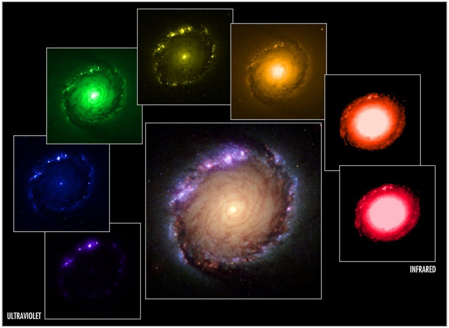
Joint work with Pierre Pudlo, and Denis Burgarella
Aix-Marseille Université

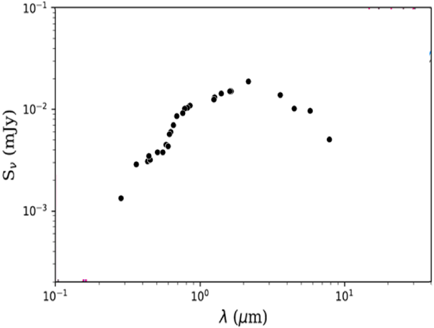
AppliBUGS
December 18 2020

INSTITUT
de MATHÉMATIQUES
de MARSEILLE

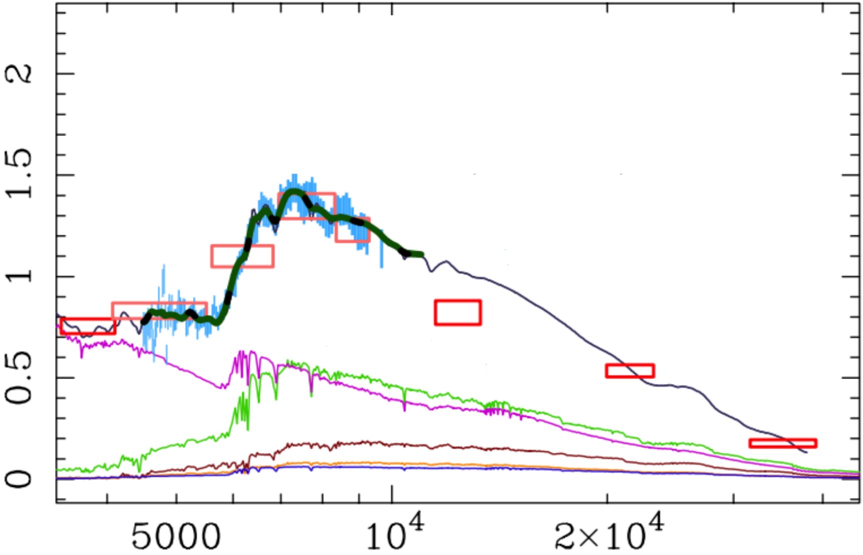
LAM
LABORATOIRE D'ASTROPHYSIQUE
DE MARSEILLE

Spectral Energy Distribution (SED)

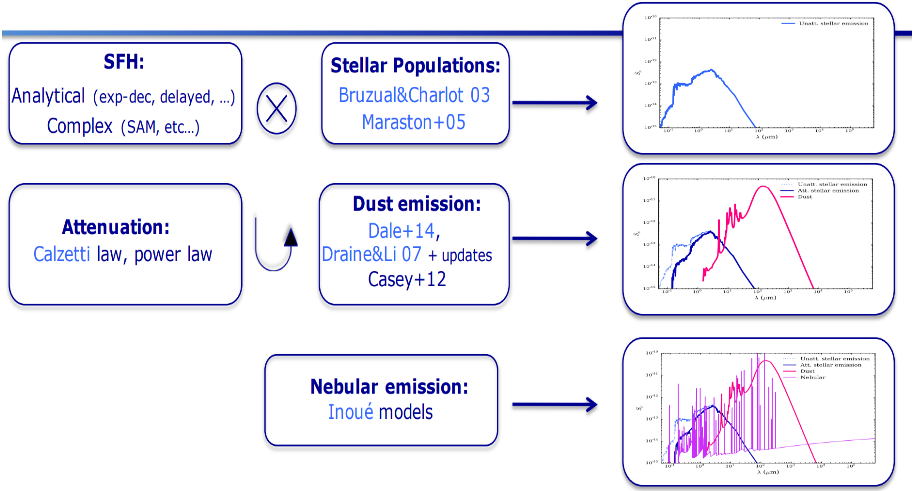




Spectroscopy



Modeling galaxies SED



SFH:
Analytical (exp-dec, delayed, ...)
Complex (SAM, etc...)

Stellar Populations:
Bruzual&Charlot 03
Maraston+05

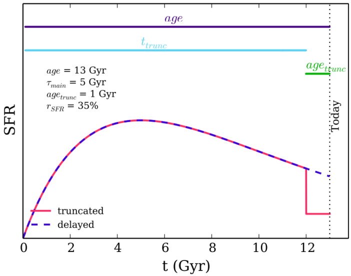
Attenuation:
Calzetti law, power law

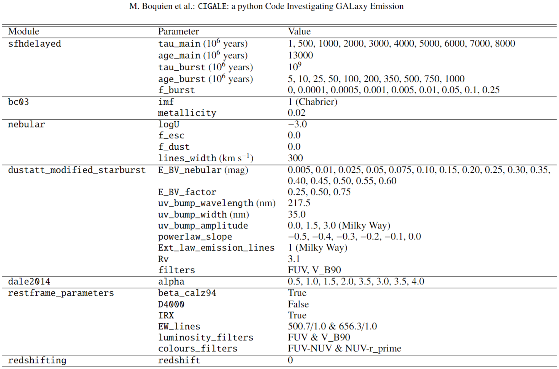
Dust emission:
Dale+14,
Draine&Li 07 + updates
Casey+12

Nebular emission:
Inoué models

Figure: The numerical code implementing the physical model

Example





M. Boquien et al.: CIGALE: a python Code Investigating GALaxy Emission

Module	Parameter	Value
sfhdelayed	tau_main (10^6 years)	1, 500, 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000
	age_main (10^6 years)	13000
	tau_burst (10^6 years)	10^9
	age_burst (10^6 years)	5, 10, 25, 50, 100, 200, 350, 500, 750, 1000
	f_burst	0, 0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1, 0.25
bc03	imf	1 (Chabrier)
	metallicity	0.02
nebular	logU	−3.0
	f_esc	0.0
	f_dust	0.0
	lines_width (km s^{-1})	300
dustatt_modified_starburst	E_BV_nebular (mag)	0.005, 0.01, 0.025, 0.05, 0.075, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50, 0.55, 0.60
	E_BV_factor	0.25, 0.50, 0.75
	uv_bump_wavelength (nm)	217.5
	uv_bump_width (nm)	35.0
	uv_bump_amplitude	0.0, 1.5, 3.0 (Milky Way)
	powerlaw_slope	−0.5, −0.4, −0.3, −0.2, −0.1, 0.0
	Ext_law_emission_lines	1 (Milky Way)
	Rv	3.1
	filters	FUV, V_B90
dale2014	alpha	0.5, 1.0, 1.5, 2.0, 3.5, 3.0, 3.5, 4.0
restframe_parameters	beta_calz94	True
	D4000	False
	IRX	True
	EW_lines	500.7/1.0 & 656.3/1.0
	luminosity_filters	FUV & V_B90
	colours_filters	FUV-NUV & NUV-r_prime
redshifting	redshift	0

Objective

The goal is to do Bayesian SED fitting for parameter inference:
We want to compute $p(\theta|x) \propto \pi(\theta) f(x|\theta)$ with

- x the observed data
- θ the galaxy parameters
- $f(\theta|x)$ the likelihood is multivariate Gaussian, centered around the theoretical SED(θ):

$$f(x|\theta) = \varphi\Big(x\Big|\text{mean}{=}SED(\theta), \text{var}{=}\Sigma_x\Big)$$

- $\pi(\theta)$ the prior distribution (usually uniform over hypercubes)

Weighting the measurements

Introduction of a spatially dependant covariance matrix

$$\Sigma_x = \begin{bmatrix} \sigma_1(x) & K_{1,2} & \dots & K_{1,N} \\ K_{2,1} & \sigma_2(x) & \dots & \\ \vdots & \vdots & \ddots & \vdots \\ K_{N,1} & \dots & & \sigma_N(x) \end{bmatrix}$$

where $\sigma_i(x)$ the estimated error on each flux and $K_{i,j} = k(\lambda_i, \lambda_j)$ an autocorrelation function

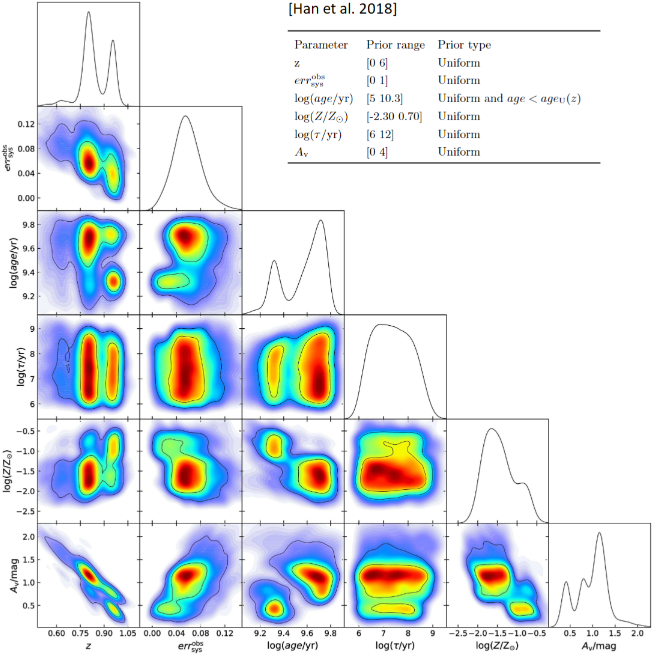
Takes care of :

- Instrument dependant noise correlations
- Systematic discrepancies between the model and the data
- Balancing weights between densely probed areas (spectroscopy) and scarce areas (photometry)

The problem

The MOONS spectrograph should be installed on the Very Large Telescope in the coming years (2021).
It will provide medium and high resolution spectrospic measurements for hundreds of thousands of galaxies
A few difficulties appear :

- SED modeling doesn't allow for an analytic solution
- The likelihood computation requires is a relatively expensive black box → the CIGALE software
- Non-linearities and dependencies between parameters due to the physical model → complex posterior
- A large number of observed datasets → No informative prior in general



[Han et al. 2018]

Parameter	Prior range	Prior type
z	[0 6]	Uniform
$err_{\rm sys}^{\rm obs}$	[0 1]	Uniform
$\log(age/{\rm yr})$	[5 10.3]	Uniform and $age < age_{\rm U}(z)$
$\log(Z/Z_\odot)$	[-2.30 0.70]	Uniform
$\log(\tau/{\rm yr})$	[6 12]	Uniform
$A_{\rm v}$	[0 4]	Uniform

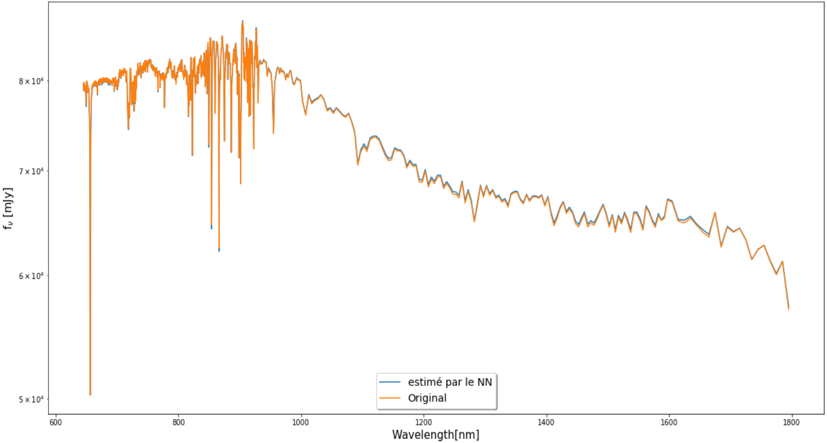
Approximating the likelihood

Splitting the problem in "$f(x|\theta) \rightarrow f(x|SED(\theta))$" we can move the approximation from the statistical model to the physical model.
Since $f(x|\theta)$ depends on both $SED(\theta)$ and x, we chose to only approximate $SED(\theta)$
Simple strategies (PCA + small Neural Network) work !

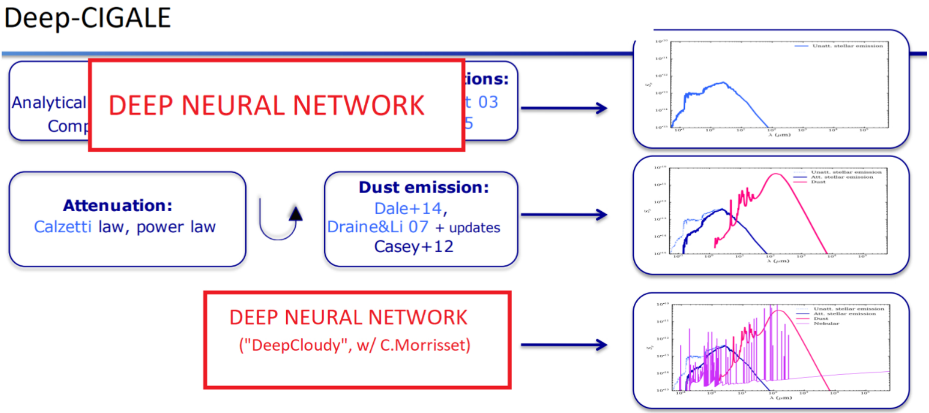
- Make a training set
- Use a PCA on the spectra
- Learn a Dense Neural Network to predict the PCA coordinates from the model parameters
- compute the real likelihood with the approximated spectra

Example for MOONS-like data : DeepCIGALE

This approximation is 1000 times faster !



Resulting : DeepCIGALE



Deep-CIGALE

Analytical
Comp

DEEP NEURAL NETWORK

ions:
t 03
5

Attenuation:
Calzetti law, power law

Dust emission:
Dale+14,
Draine&Li 07 + updates
Casey+12

DEEP NEURAL NETWORK
("DeepCloudy", w/ C.Morrisset)

Monte Carlo

No analytic solution → Numerical integration
Several usual schemes :

- HMC/ NUTS :
  No gradient because of the blackbox
- Gibbs
  No conditionals readily available
- Metropolis-Hastings :
  Too sequential and too expensive
- Variational inference :
  Complex posterior
- Importance Sampling :
  (kind of) the current method, poor prior but great potential for heavy parallelization

Importance Sampling

Assume we want to compute the integral $E_p[x] = \int xp(x)dx$. We can introduce $q(x)$ such that $q(x) > 0$ wherever $p(x) > 0$,and rewrite

$$
\begin{aligned}
\mathbb{E}_p[x] &= \int xp(x)\,dx \\
&= \int x\frac{p(x)}{q(x)}q(x)\,dx \\
&= E_q\big[X\,\frac{p(X)}{q(X)}\big]
\end{aligned}
$$

denoting $\omega_i = \frac{p(x_i)}{q(x_i)}$ this yields the unbiased estimator

$$\widehat{\mathbb{E}[X]}_N^{IS} = \frac{1}{N}\sum_{i=1}^{N} x_i\omega_i$$

and for a function h

$$\widehat{\mathbb{E}[h(x)]}_N^{IS} = \frac{1}{N}\sum_{i=1}^{N} h(x_i)\omega_i$$

Self-Normalised Importance Sampling

In most cases we know the target p up to a normalising constant Z. We can estimate

$$\hat{Z} = \frac{1}{N} \sum_{i=1}^{N} \omega_i$$

and

$$\widehat{\mathbb{E}[h(x)]}_N^{SNIS} = \sum_{i=1}^{N} h(x_i)\tilde{\omega}_i$$

with $\tilde{\omega}_i = \frac{\omega_i}{\sum_{i=1}^{N} \omega_i}$.

Algorithm :

- Sampling step : Draw N samples $x_i, i = 1, \dots, N$ from q
- weighting step : compute the weights $\omega_i = \frac{p(x_i)}{q(x_i)}$
- return the weighted sample to compute $\widehat{\mathbb{E}[h(x)]}_N^{SNIS}$ and/or $\hat{Z}$

The usual diagnostic is $\widehat{ESS} = \frac{1}{\sum_{i=1}^{N} \tilde{\omega}_i^2}$

Role of the proposal

The big problem : How to choose q ?

- Must be easy to sample from
- Optimally is proportional to the target
- should have heavier tails than the target, otherwise $\frac{p(x)}{q(x)}$ might explode

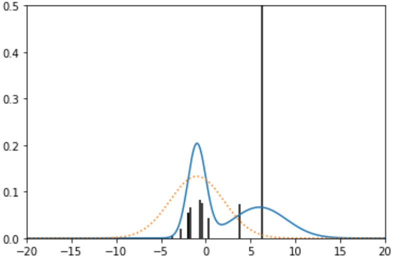


Figure: Exploding weights

Adaptive Importance Sampling

General form of AIS :

- Choose a first proposal q_0,T and a sequence of sample sizes N_t
- for $t = 0, \dots, T - 1$
  - Draw N_t samples $x_i, i = 1, \dots, N_t$ from q_t
  - Compute the weights $\omega_{t,i} = \frac{p(x_i)}{q_t(x_i)} = 1, \dots, N_t$
  - update the proposal to get q_{t+1} according to w_i^t

Adative Multiple Importance Sampling

- **Initialisation :** Choose T, N_t and take q_0 to be a Gaussian Mixture

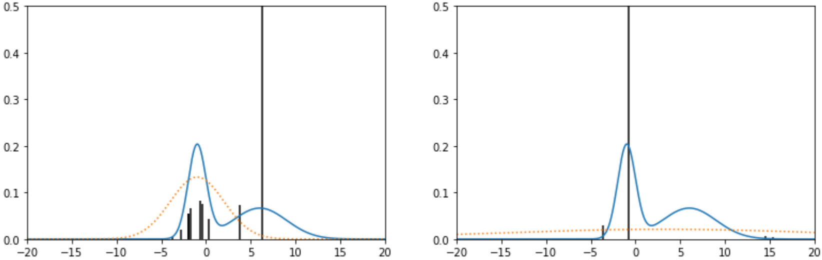
$$q_0(x) = \sum_{i=1}^{K} \pi_i \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu_i}, \boldsymbol{\Sigma_i})$$

- **Iterations :** for $t = 0, \dots, T-1$
    - Draw N_t samples $x_i, i = 1, \dots, N_t$from q_t
    - Compute the weights $\omega_{t,i} = \frac{p(x_i)}{q_t(x_i)} = 1, \dots, N_t$
    - update $\pi_i, \mu_i \text{and } \Sigma_i, i = 1, \dots, K$ using EM on the weighted sample (θ_i^t, w_i^t)
- **Recycling**
    - set $\Omega_T = N_0 + \cdots + N_t$
    - update $\omega_{t,i} = \frac{p(x_i^t)}{\Omega_T^{-1} \sum_{k=0}^{T} q_k(x_i^t)}$

Significant improvement over simple IS

Drawbacks

Still relying heavily on the initial proposal : ”what you see is what you get”, ”these adaptive methods are typically unstable.”



Solutions

Proposed solutions :

- Optimizing a scale parameter in every dimension with the Nelder-Mead algorithm[Cornuet 2012] :
  - extremely expensive in high dimension (”as long as all the iterations”)
  - doesn’t really adress a bad center intialisation
- Metropolis-Hastings to initialise PMC [Beaujean 2013]
- Nonlinear weight transformation [Koblents 2013]
- covariance updates modification for stability [El-Laham 2018]

Tempering

Mostly seen in the MCMC / SMC litterature
Choose a sequence $0 = \beta_0 < \beta_1 < \cdots < \beta_T = 1$ and target sequentially

$$p_t(x) = p(x)^{\beta_t} \qquad \text{[Kirkpatrick 83]}$$

or

$$p_t(x) = p(x)^{\beta_t} q_0(x)^{(1-\beta_t)} \qquad \text{[Neal 98]}$$

Can we do the same for AMIS ?

$$p_t(x) = p(x)^{\beta_t} q_{t-1}(x)^{(1-\beta_t)}$$

Yes ! But it doesn't really work

TAMIS

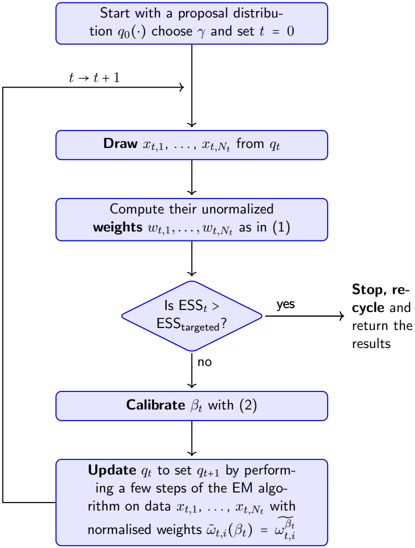
The right auxiliary target :

$$p_t(x) = p(x)^{\beta_t} q_t(x)^{(1-\beta_t)}$$

The corresponding weights :

$$\begin{aligned}\omega_{t,i}(\beta_t) &= \frac{p_t(x)}{q_t(x)} \\ &= \frac{p(x)^{\beta_t} q_t(x)^{(1-\beta_t)}}{q_t(x_i)} \\ &= \left(\frac{p(x)}{q_t(x)}\right)^{\beta_t}\end{aligned}$$

(An other useful one is $p_t(x) = p(x) + \epsilon_t q_t(x)$)



Start with a proposal distribution $q_0(\cdot)$ choose γ and set $t = 0$

$t \to t+1$

Draw $x_{t,1}, \dots, x_{t,N_t}$ from q_t

Compute their unnormalized **weights** $w_{t,1}, \dots, w_{t,N_t}$ as in (1)

Is ESS$_t$ > ESS$_{\text{targeted}}$?

yes

Stop, recycle and return the results

no

Calibrate β_t with (2)

Update q_t to set q_{t+1} by performing a few steps of the EM algorithm on data $x_{t,1}, \dots, x_{t,N_t}$ with normalised weights $\tilde{\omega}_{t,i}(\beta_t) = \widetilde{\omega_{t,i}^{\beta_t}}$

Figure: Our adaptive algorithm.

$$\omega_{t,i} = \frac{p(x_i)}{q_t(x_i)} \qquad (1)$$

$$\widehat{\text{ESS}}(\beta) = \frac{\left(\sum_{i=1}^{N_t} \omega_{t,i}(\beta)\right)^2}{\sum_{i=1}^{N_t} \omega_{t,i}^2(\beta)}$$

$$\beta_t = \sup\left\{\beta \ : \ \widehat{\text{ESS}}(\beta) > \gamma\right\} (2)$$

Why those targets ?

- The auxiliary target is always closer than the proposal to the true target

$$\begin{aligned}\mathrm{D}_{\mathrm{KL}}(p \| p^{\beta} q_t^{1-\beta}) &= \int p(x) \log\left(\frac{p(x)}{p(x)\beta q_t(x)^{1-\beta}}\right) dx \\ &= (1-\beta)\mathrm{D}_{\mathrm{KL}}(p \| q_t)\end{aligned}$$

- The function $\beta \rightarrow \widehat{\mathrm{ESS}}(\beta)$ is continuously decreasing from N to the untempered $\widehat{\mathrm{ESS}}$

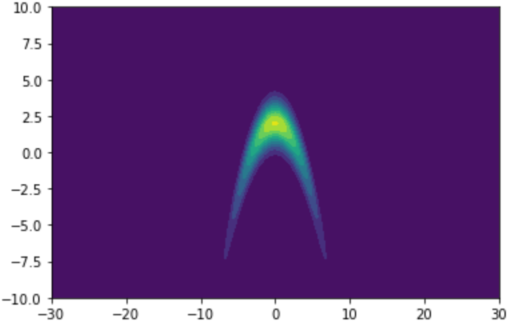
  → We can always find a β to get the desired $\widehat{\mathrm{ESS}}$
- mixing with the current proposal ensures non-zero weights

Toy examples

3 Examples :

- Multivariate gaussian : $\mathcal{N}([10, \dots]^T, 5 \times \mathbf{I_D})$
- banana likelihood :
  let $\sigma^2\Sigma = \mathrm{diag}(\sigma^2, 1, \dots, 1)$ and $b = 0.2$

$$p(x) = f_{\mathcal{N}(0_D,\Sigma)}(x_1, x_2 + b(x_1^2 - \sigma^2), x_3, \dots, x_D)$$



Gaussian-LogGamma model :

$$p(x) = \prod_{i=1}^{D} p(x_i)$$

$$\begin{aligned} p(x_1) &= 0.5\mathrm{LogGamma}(x_1|10,1,1) + 0.5\mathrm{LogGamma}(x_1|-10,1,1) \\ p(x_2) &= 0.5\mathcal{N}(x_2|10,1) + 0.5\mathcal{N}(x_2|-10,1) \end{aligned}$$

and $p(x_i) = \begin{cases} \mathrm{LogGamma}(x_i|-10,1,1) & \text{if } 3 \le i \le (D+2)/2 \\ \mathcal{N}(x_i|10,1) & \text{else} \end{cases}$

TAMIS : Numerical results Banana

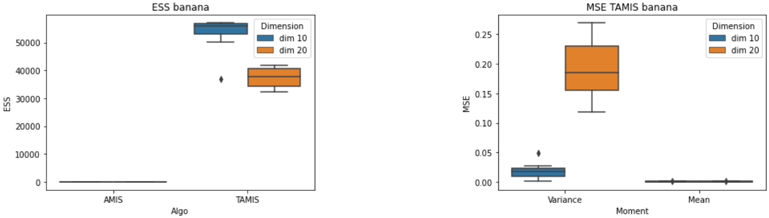


Figure: 200.000 draws over 20 iterations. The real mean is 0^d, the variance diag$(10, 9, 1 \dots, 1)$

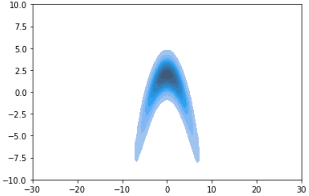


Figure: Dimension 10

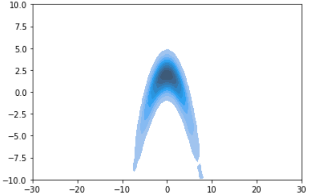


Figure: Dimension 20

TAMIS : Numerical results Gaussian

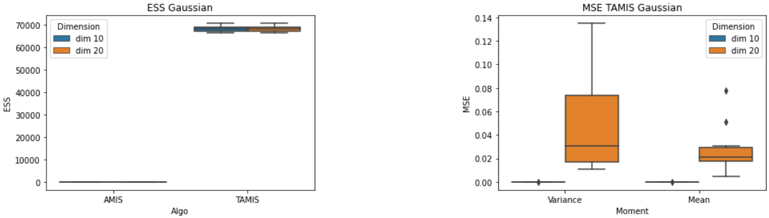
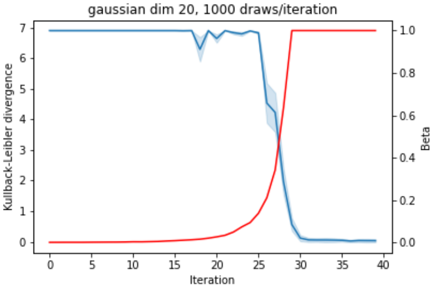


Figure: 200,000 draws over 20 iterations. The real mean is $[50, \ldots, 50]^T$, the covariance $\mathbf{I_d}$



TAMIS : Numerical results LogGamma

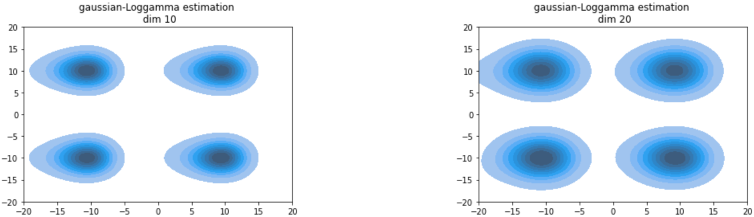


Figure: 200.000 draws over 20 iterations on the Gaussian Log gamma problem

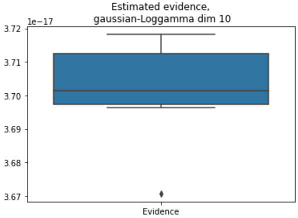


Figure: Dimension 10

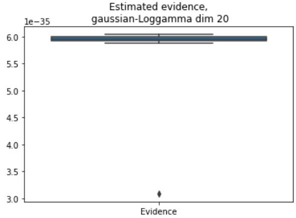
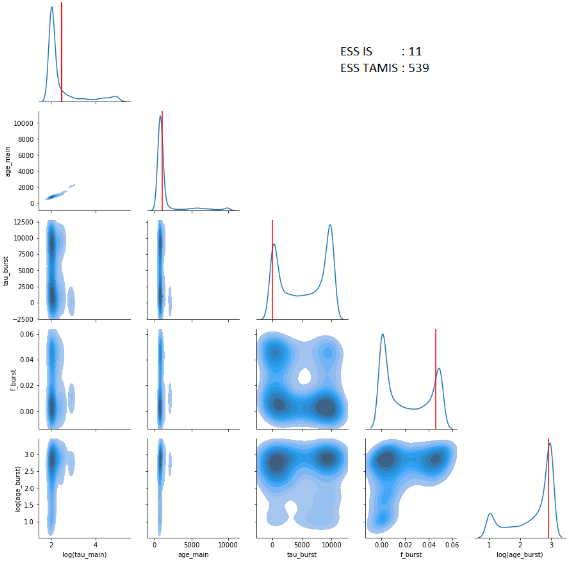


Figure: Dimension 20

TAMIS : Numerical result synthetic Galaxy



ESS IS : 11
ESS TAMIS : 539

TAMIS : Bonus

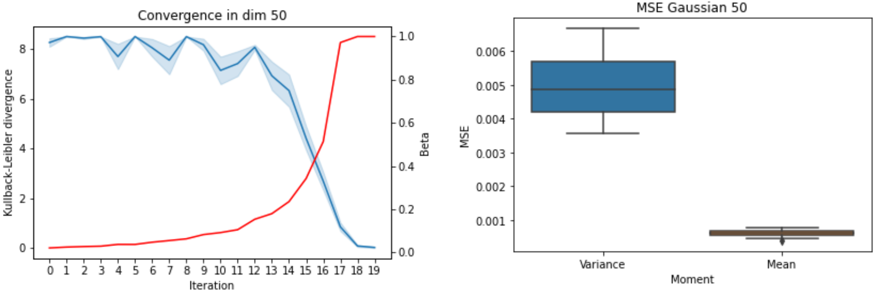


Figure: target $\mathcal{N}([10,\dots]^T, 5 \times \mathbf{I_{50}})$ initialization with $\mu_0 = 0$ and $\Sigma_0 = 200 \times \mathbf{I_{50}}$

Conclusion

We propose an easy modification of the AMIS algorithm to mitigate the initialization difficulty with almost no additional cost.
It is easy to tune, heavily parallelisable and suffers less from the curse of dimensionality. It also drastically decreases the required number of likelihood evaluation.
It is effective on the galaxy parameter inference problem and provides a clear speed up.
Next :

- Studying TAMIS
- Transfert learning to initialize the proposal

References

Jean-Marie Cornuet and Jean-Michel Marin and Antonietta Mira and Christian P. Robert "Adaptive Multiple Importance Sampling." Scandinavian Journal of Statistics, vol. 39, no. 4, 2012, pp. 798–812.,

Jean-Michel Marin, Pierre Pudlo, Mohammed Sedki. Consistency of the Adaptive Multiple Importance Sampling. Bernoulli, Bernoulli Society for Mathematical Statistics and Probability, 2019

Neal, R.M. Annealed importance sampling. Statistics and Computing 11, 125–139 (2001).

Bugallo, Mónica Elvira, Victor Martino, Luca Luengo, David Miguez, Joaquin Djuric, Petar. (2017). Adaptive Importance Sampling: The past, the present, and the future. IEEE Signal Processing Magazine. 34. 60-79. 10.1109/MSP.2017.2699226.

Jun Liu

Y. El-Laham, V. Elvira and M. F. Bugallo, "Robust Covariance Adaptation in Adaptive Importance Sampling," in IEEE Signal Processing Letters, vol. 25, no. 7, pp. 1049-1053, July 2018

Koblents, Eugenia Miguez, Joaquin. (2012). A population Monte Carlo scheme with transformed weights and its application to stochastic kinetic models. Statistics and Computing.

Cameron, E. Pettitt, Anthony. (2012). Approximate Bayesian Computation for Astronomical Model Analysis: A Case Study in Galaxy Demographics and Morphological Transformation at High Redshift. Monthly Notices of the Royal Astronomical Society. 425. 10.1111/j.1365-2966.2012.21371.x.