Current MAF results on reference problem

Sebastian Pina Otey

Supervisors: Vicens Gaitán, Federico Sánchez, Anna Espinal

Grupo AIA/IFAE/SEA

January 10, 2018

GRUPO AIA

Institut de Física d'Altes Energies

Servei d'Estadística
Universitat Autònoma de Barcelona

Problem statement

- Given observed data $\mathbf{X}_{\text{obs}} = \{\mathbf{x}_{\text{obs}}^{(1)}, \ldots, \mathbf{x}_{\text{obs}}^{(N)}\}$ coming from some true distribution $f_{\text{real}}(\mathbf{x})$ and a scientific model $f(\mathbf{x}|\theta)$ with parameters θ, find θ^* such that $d\left(f_{\text{real}}(\mathbf{x}), f(\mathbf{x}|\theta)\right)$ for some distance/divergence d is minimized.

Problem statement

- Given observed data $\mathbf{X}_{\text{obs}} = \{\mathbf{x}_{\text{obs}}^{(1)}, \ldots, \mathbf{x}_{\text{obs}}^{(N)}\}$ coming from some true distribution $f_{\text{real}}(\mathbf{x})$ and a scientific model $f(\mathbf{x}|\theta)$ with parameters θ, find θ^* such that $d\left(f_{\text{real}}(\mathbf{x}), f(\mathbf{x}|\theta)\right)$ for some distance/divergence d is minimized.
- Bayesian approach, apply Bayes theorem on priors $f(\theta)$ to get posterior $f(\theta|\mathbf{X}_{\text{obs}})$ after observing the data:

$$f(\theta|\mathbf{X}_{\text{obs}}) \propto L(\theta) f(\theta) = \prod_{\mathbf{x}_{\text{obs}} \in \mathbf{X}_{\text{obs}}} f(\mathbf{x}_{\text{obs}}|\theta) f(\theta)$$

Problem statement

- Given observed data $\mathbf{X}_{\text{obs}} = \{\mathbf{x}_{\text{obs}}^{(1)}, \dots, \mathbf{x}_{\text{obs}}^{(N)}\}$ coming from some true distribution $f_{\text{real}}(\mathbf{x})$ and a scientific model $f(\mathbf{x}|\theta)$ with parameters θ, find θ^* such that $d\left(f_{\text{real}}(\mathbf{x}), f(\mathbf{x}|\theta)\right)$ for some distance/divergence d is minimized.
- Bayesian approach, apply Bayes theorem on priors $f(\theta)$ to get posterior $f(\theta|\mathbf{X}_{\text{obs}})$ after observing the data:

$$f(\theta|\mathbf{X}_{\text{obs}}) \propto L(\theta) f(\theta) = \prod_{\mathbf{x}_{\text{obs}} \in \mathbf{X}_{\text{obs}}} f(\mathbf{x}_{\text{obs}}|\theta)\, f(\theta)$$

- Problem: sometimes $f(\mathbf{x}|\theta)$ is not available, we can only sample from it given a parameter θ.

MAF: approximating densities

- Masked Autoregressive Flow is a neural network (NN) with parameters ψ that takes as an input $(\theta, \mathbf{x})$ and outputs a probability $q_\psi(\mathbf{x}|\theta)$.

MAF: approximating densities

- Masked Autoregressive Flow is a neural network (NN) with parameters ψ that takes as an input $(\theta, \mathbf{x})$ and outputs a probability $q_\psi(\mathbf{x}|\theta)$.
- The NN is trained by minimizing the Kullback-Leibler divergence between the distributions f and q_ψ, $KL(f, q_\psi)$. This is equivalent to maximizing with respect to ψ the log-likelihood $\sum_{(\theta,\mathbf{x})} \log q_\psi(\mathbf{x}|\theta)$ if we sample $\mathbf{x} \sim f(\mathbf{x}|\theta)$.

MAF: approximating densities

- Masked Autoregressive Flow is a neural network (NN) with parameters ψ that takes as an input $(\theta, \mathbf{x})$ and outputs a probability $q_\psi(\mathbf{x}|\theta)$.
- The NN is trained by minimizing the Kullback-Leibler divergence between the distributions f and q_ψ, $KL(f, q_\psi)$. This is equivalent to maximizing with respect to ψ the log-likelihood $\sum_{(\theta,\mathbf{x})} \log q_\psi(\mathbf{x}|\theta)$ if we sample $\mathbf{x} \sim f(\mathbf{x}|\theta)$.
- If the samples are weighted, the objective function has to be simply reweighted: $\sum_{(\theta,\mathbf{x},w)} w \cdot \log q_\psi(\mathbf{x}|\theta)$

MAF: approximating densities

- Masked Autoregressive Flow is a neural network (NN) with parameters ψ that takes as an input $(\theta, \mathbf{x})$ and outputs a probability $q_\psi(\mathbf{x}|\theta)$.
- The NN is trained by minimizing the Kullback-Leibler divergence between the distributions f and q_ψ, $KL(f, q_\psi)$. This is equivalent to maximizing with respect to ψ the log-likelihood $\sum_{(\theta,\mathbf{x})} \log q_\psi(\mathbf{x}|\theta)$ if we sample $\mathbf{x} \sim f(\mathbf{x}|\theta)$.
- If the samples are weighted, the objective function has to be simply reweighted: $\sum_{(\theta,\mathbf{x},w)} w \cdot \log q_\psi(\mathbf{x}|\theta)$
- If q_ψ approximates f well enough, it can be used in the Bayesian approach to obtain the posterior.

Reference Problem

Magnitudes $\mathbf{x} = \{x_1, x_2\}$ and a single parameter θ with density

$$f(\mathbf{x}|\theta) = f(x_1, x_2|\theta) = f(x_2|x_1, \theta)\, f(x_1|\theta).$$

Reference Problem

Magnitudes $\mathbf{x} = \{x_1, x_2\}$ and a single parameter θ with density

$$f(\mathbf{x}|\theta) = f(x_1, x_2|\theta) = f(x_2|x_1, \theta)\, f(x_1|\theta).$$

The random variables and the parameter are related by:

$$X_1 = \log(Y),\ \ Y \sim \mathrm{Exp}(0.2); \quad X_2/X_1 \sim \mathrm{Gamma}(6 + \theta \cdot X_1^2),$$

Reference Problem

Magnitudes $\mathbf{x} = \{x_1, x_2\}$ and a single parameter θ with density

$$f(\mathbf{x}|\theta) = f(x_1, x_2|\theta) = f(x_2|x_1, \theta)\, f(x_1|\theta).$$

The random variables and the parameter are related by:

$$X_1 = \log(Y), \;\; Y \sim \mathrm{Exp}(0.2); \quad X_2/X_1 \sim \mathrm{Gamma}(6 + \theta \cdot X_1^2),$$

The total density is

$$f(x_1, x_2|\theta) = \left(\frac{1}{0.2} \cdot e^{-\exp(x_1)/0.2 \; + \; x_1}\right) \cdot \frac{x_2^{5+\theta \cdot x_1^2}\, e^{-x_2}}{\Gamma(6 + \theta \cdot x_1^2)}$$

Reference Problem

Magnitudes $\mathbf{x} = \{x_1, x_2\}$ and a single parameter θ with density

$$f(\mathbf{x}|\theta) = f(x_1, x_2|\theta) = f(x_2|x_1, \theta)\, f(x_1|\theta)\,.$$

The random variables and the parameter are related by:

$$X_1 = \log(Y),\;\; Y \sim \mathrm{Exp}(0.2); \quad X_2/X_1 \sim \mathrm{Gamma}(6 + \theta \cdot X_1^2),$$

The total density is

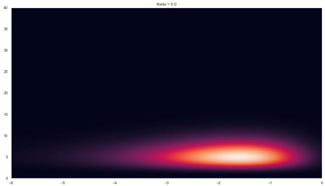
$$f(x_1, x_2|\theta) = \left(\frac{1}{0.2} \cdot e^{-\exp(x_1)/0.2 \;+\; x_1}\right) \cdot \frac{x_2^{5+\theta \cdot x_1^2}\, e^{-x_2}}{\Gamma(6 + \theta \cdot x_1^2)}$$

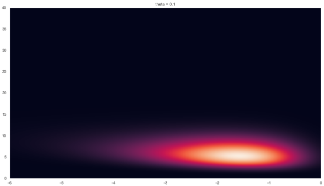
Our prior is $\theta \in \mathrm{Unif}(0, 2)$, and we sample the observations from $\theta = 0.6$.

Problems and solutions we had in MAF in reference problem

1. Densities over all the parameter possibilities, $\theta \in [0.0, 0.1]$, were too similar:

Left: $\theta = 0.0$. Right: $\theta = 0.1$.

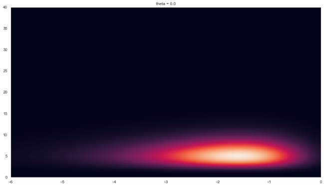


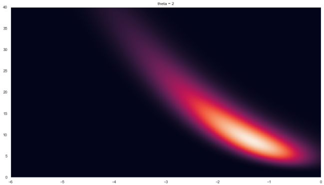


Problems we had in MAF in reference problem

New problem was defined with $\theta \in [0.0, 2.0]$

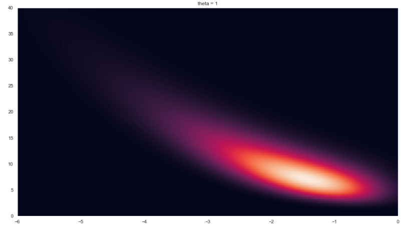
Left: $\theta = 0.0$. Right: $\theta = 2.0$.

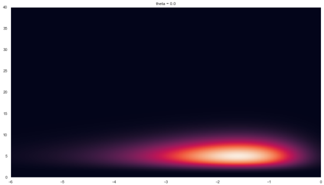


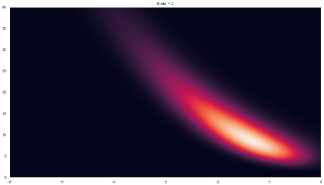


Problems and solutions we had in MAF in reference problem

2. Nominal density to be reweighted did not cover all the space ($\theta = 1.0$):



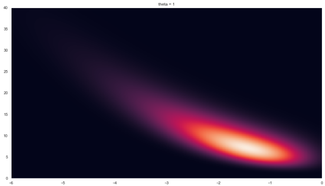


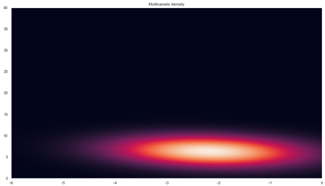


Problems and solutions we had in MAF in reference problem

Sample from all the space $(\theta, \mathbf{x})$, then define nominal distribution as a multivariate with mean and covariance of the samples:

Left: $\theta = 1.0$. Right: Multivariate distribution.

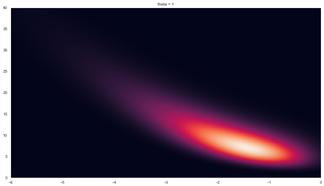


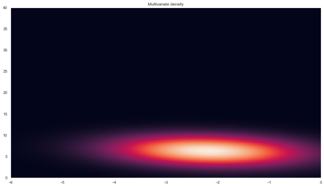


Problems and solutions we had in MAF in reference problem

Sample from all the space $(\theta, \mathbf{x})$, then define nominal distribution as a multivariate with mean and covariance of the samples:

Left: $\theta = 1.0$. Right: Multivariate distribution.





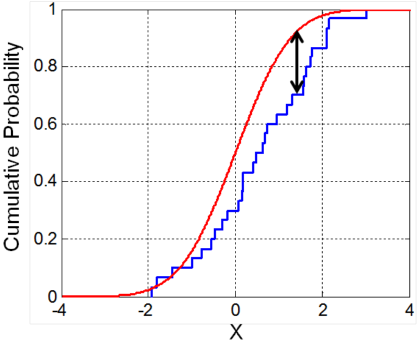
Why did this work?

Results

Results: Compute posterior of θ of 100 sets of 100 observations.

Results

Results: Compute posterior of θ of 100 sets of 100 observations. Compare them to exact posterior using the Kolmogorov-Smirnov (KS) statistics (maxium between cumulative density functions, CDF).



Results

Results: Compute posterior of θ of 100 sets of 100 observations. Compare them to exact posterior using the Kolmogorov-Smirnov (KS) statistics (maxium between cumulative density functions, CDF).
Three likelihood-free methodologies were applied under the same conditions:

- MAF.
- Binned MarkovChain Montecarlo.
- Approximate Bayesian Computation (discarded, the results were bad).

Results. 1. MAF

We trained the folllowing combinations of MAF structures:

- ns mades=(5 10)
- architectures=("[5]*2" "[20]*2" "[20]*10")
- batch sizes=(100 500 -1)
- early stoppings=(100 1000 10000)

54 combinations in total.

Results. 1. MAF

We trained the folllowing combinations of MAF structures:

- ns mades=(5 10)
- architectures=("[5]*2" "[20]*2" "[20]*10")
- batch sizes=(100 500 -1)
- early stoppings=(100 1000 10000)

54 combinations in total.
We compute 100 KS for each of them and sort them by mean KS score.

Results. 1. MAF

We trained the folllowing combinations of MAF structures:

- ns mades=(5 10)
- architectures=("[5]*2" "[20]*2" "[20]*10")
- batch sizes=(100 500 -1)
- early stoppings=(100 1000 10000)

54 combinations in total.

Results. 1. MAF

We trained the folllowing combinations of MAF structures:

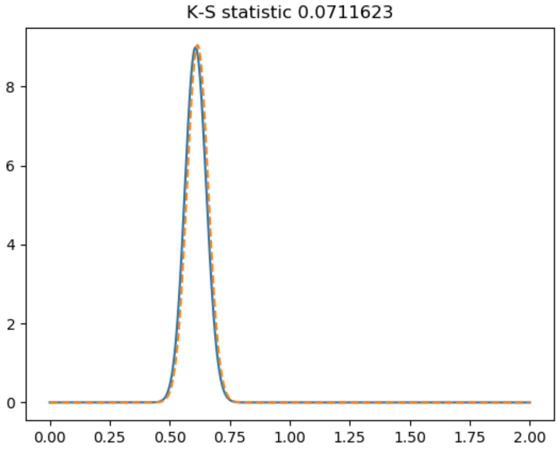
- ns mades=(5 10)
- architectures=("[5]*2" "[20]*2" "[20]*10")
- batch sizes=(100 500 -1)
- early stoppings=(100 1000 10000)

54 combinations in total.
We compute 100 KS for each of them and sort them by mean KS score.

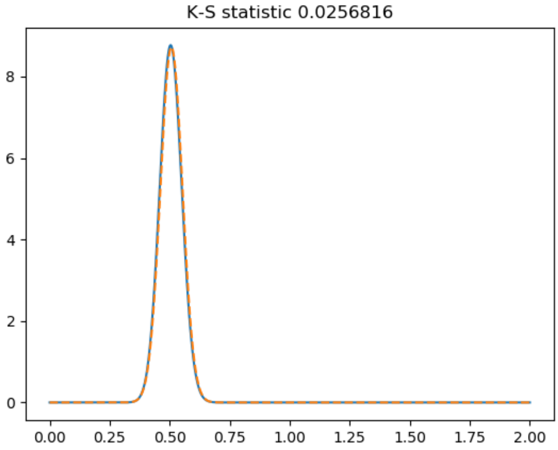
Results. 1. MAF

	n_mades	architecture	batch_size	early_stopping	rmse	ks	ks_mean	ks_var
n_test								
34	10	[5, 5]	-1	100	0.001157	[0.07116227209832351, 0.03356896686956544, 0.1...	0.046895	0.001295
41	10	[20, 20]	500	1000	0.001510	[0.042223365869773856, 0.08163860652569518, 0....	0.055031	0.001978
33	10	[5, 5]	500	10000	0.001214	[0.025629406157212802, 0.055911237822021495, 0...	0.055109	0.004221
20	5	[20, 20, 20, 20, 20, 20, 20, 20, 20, 20]	100	1000	0.001864	[0.017524367724661743, 0.02845554345480889, 0....	0.059193	0.004627
21	5	[20, 20, 20, 20, 20, 20, 20, 20, 20, 20]	100	10000	0.001265	[0.05189295669797125, 0.01702239092069454, 0.0...	0.060888	0.002801
15	5	[20, 20]	500	10000	0.001161	[0.05502323619636504, 0.06668260839000646, 0.0...	0.061524	0.002702
7	5	[5, 5]	-1	100	0.001217	[0.006113853726749614, 0.054812745606840664, 0...	0.063003	0.004289

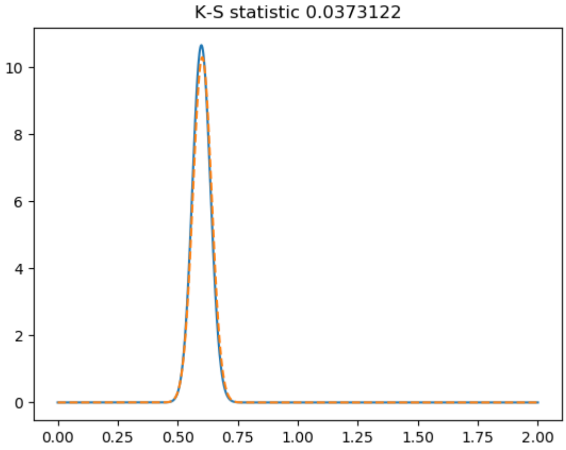
Results. 1. MAF



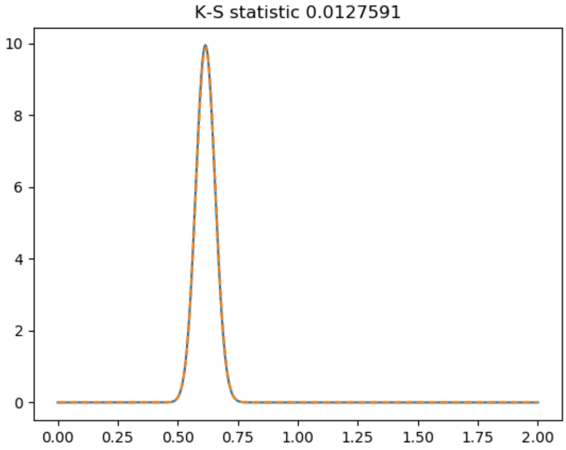
Results. 1. MAF



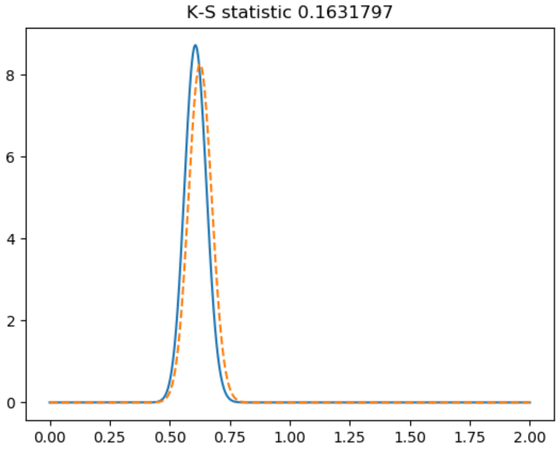
Results. 1. MAF



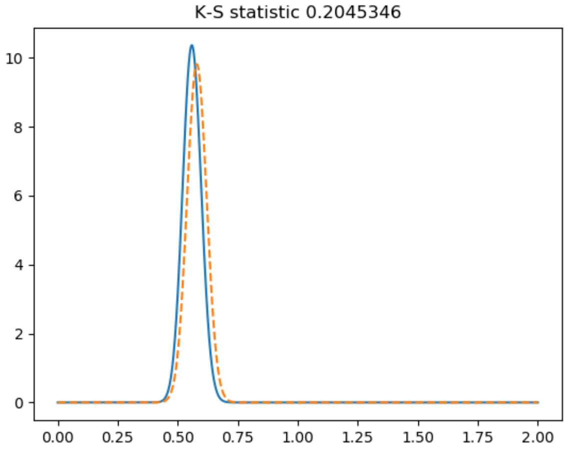
Results. 1. MAF



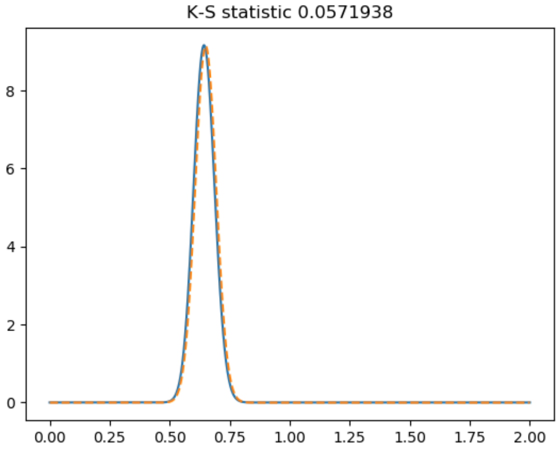
Results. 1. MAF



Results. 1. MAF



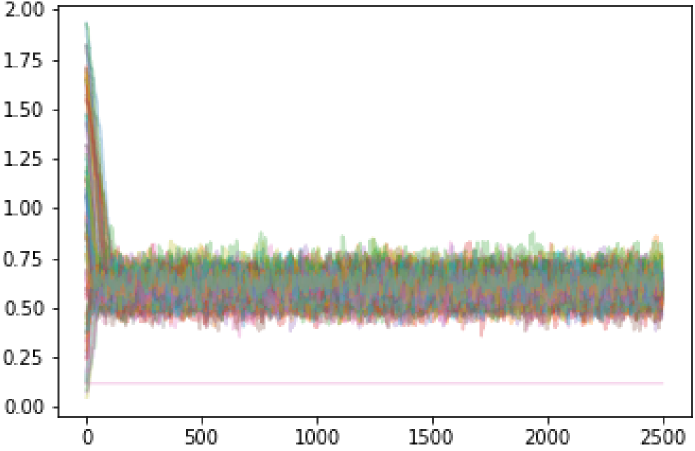
Results. 1. MAF



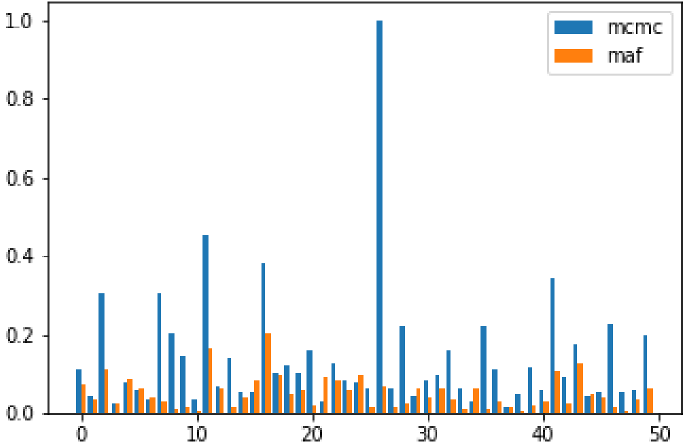
Results. 2. MCMC

30x30 bins all over the same space from min to max of nominal values to be reweighted. Poisson distribution is assumed in each bin.
2500 samples obtained from posterior, considering 200 first as burn in.

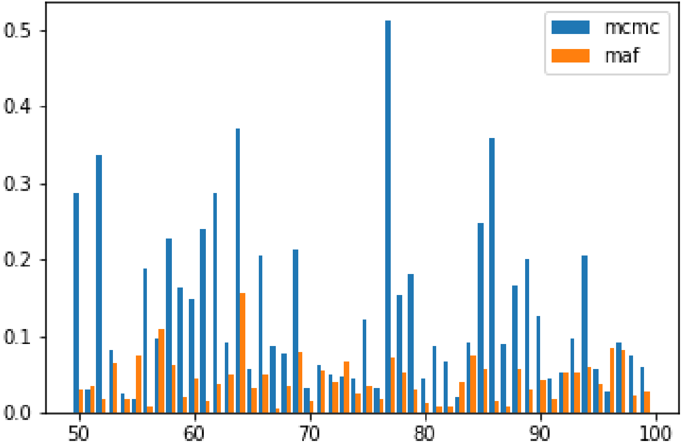
Results. 2. MCMC



Results. 2. MCMC



Results. 2. MCMC



Results. MAF vs MCMC

MCMC achieved mean KS of 0.1358 with variance of 0.0181. In 17 out of 100, MCMC binned performed better than MAF. MAF results were still really good in these cases:

`0.0257, 0.0865, 0.0635, 0.0373, 0.0835, 0.0917, 0.0961, 0.0637, 0.0618, 0.0496, 0.0342, 0.0729, 0.1093, 0.0664, 0.0401, 0.0515, 0.083`

Pearson coefficient between both KS vectors is 0.3152.

ks_mean	ks_var
0.046895	0.001295
0.055031	0.001978
0.055109	0.004221
0.059193	0.004627
0.060888	0.002801
0.061524	0.002702