Data Analysis Project in Bayesian Statistics: Techniques and models

Alexey Popov

27 august 2017

Abstract

Using most actual and most interested data provided by GAIA space mission less then one year ago, parameters of the Ogorodnikov-Miln kinematic model was obtained for two samples of stars, B-stars and K-stars of the Main Sequense. Two methods were used: MCMC regression technik and classic Least-Squares method (or Ordinary Least Squares method, OLS). The difference in kinematic of the stars from different spectal types is confirmed, good estimations of the Oort's parameters has been derived. MCMC methods gives more precise estimation on the worse data them OLS methods.

Introduction

It's hard to determine the structure of our Galaxy and find out how it rotates. First reason is beacause of the place of the observer, us. We are inside the Galaxy. The second reason: gas and dust in Galaxy makes many stars fainter or even covers it completely. The third reason: Galaxy is really huge so we can't observe distant stars because they becomes too faint for our optics. As result we can observe and measure a small part of Galaxy only. What we know already is that Galaxy rotates not as a solid disk, angular velocity does not the same on the different distance from Galaxy's center. Its not enough to measure angular velocity at Solar distance. We need much more information.

This problem was solved (on some level of precision) in last centures by introducing a different kinematic models. These models explain how observed proper motion of the stars on the sky depends on there real star motion in the space. Most popular and well-known are Oort-Lindbood and Ogorodnikov-Miln models.

The equation of the Ogorodnikov-Miln models looks like:

$$\begin{aligned}\mathcal{K}\mu_l \cos b = \quad & U/r\ \sin l - V/r\ \cos l - \omega_1 \sin b \cos l - \omega_2 \sin b \sin l + B \cos b - \\ - \quad & M_{13}^{+} \sin b \sin l + M_{23}^{+} \sin b \cos l + A \cos b \cos 2l - \\ - \quad & C \cos b \sin 2l,\end{aligned}$$

$$\begin{aligned}\mathcal{K}\mu_b = \quad & U/r\ \cos l \sin b + V/r\ \sin l \sin b - W/r\ \cos b + \omega_1 \sin l - \omega_2 \cos l - \\ - \quad & \frac{1}{2} A \sin 2b \sin 2l + M_{13}^{+} \cos 2b \cos l + M_{23}^{+} \cos 2b \sin l - \\ - \quad & \frac{1}{2} C \sin 2b \cos 2l - \frac{1}{2} K \sin 2b\end{aligned}$$

where, $K = 4.74 km/s/kpc$ is a constant coefficient; μ_l, μ_b is a stars proper motions, observed data; l, b is a stars coordinates in Galactic coordinate system, observed data; r is a stars Solar distance, observed variable;

$V, U, W, A, B, C, K, M_i j^+$ - explanatory variables, parameters of the kinematic model we need to find out.

Usually, OLS gives good results if our star samples satisfies some conditions: stars uniformly enough distributed over the celestial sphere, stars observed parameters are independent and have the same variance. Unfortunately, the last condition about equal variance for all variables does not satisfied often. The proper motions, distance and positions come with a different variance. So, I hope that MCMC let us get better results then we can get with OLS. The biggest problem we have is a error in distance. It is not possible to observe it directly. From the observation we have parallax, the angle between different positions of the star during the year. The distance r can be calculated form parallax px as $r = 1/px$. But when px is small, uncertancy in px makes a troubles in r. This means that first 3 terms of the model V, U, W are seriously affected by the errors in r. This is a big problem considered in some reseaches (e.g. T.L. Astraatmadja, C.A.L. Bailer-Jones, Astrophys. J. 832,Issue 2, article id. 137 26 pp. (2016))

We also know from previous reseaches that kinematic of the stars sufficiently depends of the Luminocity class and Spectral class of the stars in sample and on the some other parametes (like mean age in the sample, population, distance from galaxy disk plane, etc). Galaxy disk is not a solid disk and it has really complecated kinematic.

Data

The data is used in current reseach was provided by ESA space telescope mission GAIA ((Lindegren, L., Lammers, U., Bastian, U., et al.), Gaia Data Release Astrometry one billion positions, two million proper motions and parallaxes, Astronomy and Astrophysics, Vol. 595, id.A4, 323 pp. (2016)).

The GAIA DR1 TGAS catalogue can be downloaded here. Dataframe contains some missing data, some variables are provided in inconvinient units. So, on the first all data was cleaned and transformed.

On the second step, based on the photometry of the stars the L-class and Sp-class for each star was estimated.

After that we can make more or less homogeneous samples of stars with the same L-class and Sp-class. Let's considers sample of B-stars and sample of K-stars, both of the Main Sequence (5-th Luminocity class). The prepared data in R looks like:

```
head(stars)
              l          b         r      mu_l        mu_b
 [1,] 2.904064 -0.7895090 0.3211907 -3.935030 -18.9630966
 [2,] 2.869837 -0.5454267 0.5769224  2.378347   0.2660637
 [3,] 2.958054 -0.4187951 0.2965226  5.863128  -4.0032136
```

Based on the observed data we have to make matrix with coefficients according with the equations of the kinematic model listed above. The final data set of the equations coefficients looks like:

```
head(a)
              U        V W         Wx         Wy         B        M13       M23         A         C K
 [1,] 0.7325900 3.025998 0 -0.6900724 0.16706558 0.7041940 0.16706558 0.6900724 0.6262163 0.3220905 0
 [2,] 0.4652679 1.669724 0 -0.4997441 0.13925351 0.8549060 0.13925351 0.4997441 0.7317121 0.4421106 0
 [3,] 0.6155011 3.315781 0 -0.3998297 0.07421952 0.9135796 0.07421952 0.3998297 0.8527171 0.3278738 0
```

(Zero K is OK, because it appears in the second equation only)

Model

The model we used looks like:

"'r mod_string = " model { for (i in 1:length(y)){ y[i] ~ dnorm(mu[i], prec)

```
  mu[i] = x[1]*U[i] + x[2]*V[i] + x[3]*W[i]  + x[4]*Wx[i] + x[5]*Wy[i]+ x[6]*B[i] + x[7]*M13[i] + x[8]*M
}
```

```
for (i in 1:3){
  x[i] ~ dnorm(10.0, 1.0/1.0e3)
}
for (i in c(4, 5, 7, 8)){
  x[i] ~ dnorm(0.0, 1.0/1.0e2)
}
x[6] ~ dnorm(-13.0, 1.0/1.0e1)
x[9] ~ dnorm(14.0, 1.0/1.0e1)
x[10] ~ dnorm(-2.0, 1.0/1.0e2)
x[11] ~ dnorm(-2.0, 1.0/1.0e2)

prec ~ dgamma(1.0/2.0, 1.0*1000.0/2.0)
sig2 = 1.0 / prec
sig = sqrt(sig2)
```

} " ‘‘‘

Results

The first sample, B-stars of Main Sequence, includes 7019 rows (we have 7019 B-stars of Main Sequence with all parameters available in GAIA TGAS catalogue).

I use 3 chains of 5000 iterations witn 1000 burn-in terms.

The convergence diagnostics by Gelman test and autocorrelation test give very good results. The effective size looks good also:

```
> effectiveSize(mod4_sim)
   sig      x[1]    x[2]   x[3]    x[4]     x[5]   x[6]    x[7]    x[8]   x[9]     x[10]  x[11]
 15152.7  9972.4  9297.1 10749.3  1406.3  2319.8  9004.7  2330.8  1430.9 10741.5  9640.1 12439.6
```

And the result itself looks very similar to estimations I got with OLS.

```
 Iterations = 1:15000
 Thinning interval = 1
 Number of chains = 1
 Sample size per chain = 15000

 1. Empirical mean and standard deviation for each variable,
    plus standard error of the mean:

          Mean     SD Naive SE Time-series SE
 sig    21.418 0.1265 0.001033       0.001159
 x[1]   12.006 0.2690 0.002197       0.002686
 x[2]   13.488 0.3081 0.002516       0.003249
 x[3]    7.538 0.2073 0.001693       0.001992
 x[4]    4.425 0.8054 0.006576       0.021803
 x[5]   -3.326 0.7276 0.005941       0.015118
 x[6]  -13.969 0.2967 0.002423       0.003225
 x[7]   -3.637 0.7801 0.006370       0.016396
 x[8]   -3.654 0.8548 0.006979       0.022625
 x[9]   12.394 0.3850 0.003143       0.003874
 x[10]  -3.723 0.4017 0.003280       0.004092
 x[11]  -7.160 1.4562 0.011889       0.013173
```

OLS estimation gives us:

```
 > res_tgas$X
    U         V       W       Wx       Wy       B        M13      M23     M12(A)    C        K
 12.0114  13.4943   7.5343   4.4939  -3.3595 -13.9838  -3.6703  -3.7270  12.3608  -3.7275  -7.2481

 > res_tgas$s_X
   eU      eV     eW      eWx     eWy      B      eM13     eM23   A        eC     eK
0.26914 0.30874 0.20783 0.79209 0.72904 0.29768 0.77847 0.84224 0.39041 0.40369 1.47116
```

What's about K-stars? In a sample we have 29115 stars, 4 times more then in the sample with B-stars.

```
Iterations = 1:15000
Thinning interval = 1
Number of chains = 1
Sample size per chain = 15000

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:
            Mean     SD Naive SE Time-series SE
sig   242.011535 0.7105 0.005801       0.005652
x[1]    8.091786 0.1732 0.001414       0.001414
x[2]   18.439752 0.1784 0.001457       0.001494
x[3]    6.825042 0.1714 0.001400       0.001400
x[4]    2.335973 1.7678 0.014434       0.015152
x[5]   -2.403117 1.7727 0.014474       0.014474
x[6]  -11.790597 1.5021 0.012264       0.012404
x[7]   -1.340394 2.2361 0.018258       0.018511
x[8]   -0.004865 2.1961 0.017931       0.018433
x[9]   18.770465 1.8286 0.014931       0.015222
x[10]   0.228906 2.1654 0.017680       0.018398
x[11]  -1.625198 3.7886 0.030933       0.032052
```

What's about OLS?

```
> res_tgas5$X
    U       V        W        Wx        Wy       Wz(B)     M13       M23        A     C       K
8.09559  18.43699  6.82405  2.51330  -2.65509 -11.39131  -1.48056  -0.10083  21.16000 0.30110 -1.39316

> res_tgas5$s_X
   eU      eV        eW       eWx       eWy     eWz(B)    eM13      eM23       eA     eC      eK
0.17360  0.17781   0.17083  1.79910  1.80688   1.71379    2.28717   2.24067   2.236    2.220  4.122
```

That's interesting! MCMC makes more pricise estimations for meaningful parameters e.g. A and B.

Results

1. MCMC gives good estimations of the Ogorodnikov-Miln model's parameters similar to teh values, obtained by OLS, with a good precision;
2. MCMC works much slower then OLS, but requires less memory;
3. Sufficienty different means and suffeciently higher posterior variance of the explanatory variables tells us that the motion of K-stars differs from the motion of B-stars. K-stars fit Ogorodnikov-Miln model worse then B-star. This can be explained if we take into account the age of the K-stars. K-stars much older and during its lifetime it has been expirienced gravity infuence of other stars longer.
4. For K-stars MCMC gives us estimations with higher precision then OLS.

There are a lot consequenses and conclusions but it looks like it's out of this capstone.