Supplement to

Close encounters of the stellar kind (A&A 537, A75, 2015)

Coryn A.L. Bailer-Jones
Max Planck Institute for Astronomy, Heidelberg (calj@mpia.de)
22 February 2015

Use of the median as opposed to the mean

In the published paper I sample over the uncertainties in the astrometry and radial velocities to construct the posterior probability distribution function (PDF) over the perihelion time, distance, and speed for each encountering object. From an inference point of view this PDF is the final result. But in practice we want to summarize this distribution. A good summary (for unimodal distributions, which these all are) requires at least two numbers in order to characterize the range over which there is significant probability. One approach is to report the mean, μ, and standard deviation, σ. However, the standard deviation makes little sense for the perihelion distance, because that quantity must be strictly positive, whereas $\mu - \sigma$ could be negative. I therefore use the 5% and 95% quantiles of the posterior PDF, which together give a 90% confidence interval. For the subsequent analyses and comparisons I also need a single characteristic value, and for this I adopt the mean (expectation value), denoted $t_{\rm ph}^{\rm av}$, $d_{\rm ph}^{\rm av}$, and $\mathrm{v}_{\rm ph}^{\rm av}$.

Instead of the mean one could use the mode or median, or indeed any other reasonable estimator. Because the perihelion distance must be positive, the distribution over this quantity is generally skewed towards positive values, as you can see in Figure 4 in the published paper. The mean will, therefore, in general be larger than the median or mode. These is of course no "correct" estimator, although one may be preferable depending on what you want to show or test for. But how sensitive are my results to my choice?

Using the same set of samples as in the paper, I have calculated the median for the perihelion time, distance, and speed distributions. I denote these $t_{\rm ph}^{\rm med}$, $d_{\rm ph}^{\rm med}$, and $\mathrm{v}_{\rm ph}^{\rm med}$ respectively. Figure 1 shows the difference between the median and mean as a function of $d_{\rm ph}^{\rm av}$ for objects with $d_{\rm ph}^{\rm av} < 10\,$pc. The difference are generally very small (remember that there are 1548 objects plotted in each of these panels). Figure 2 shows the zoom of this for $d_{\rm ph}^{\rm av} < 2\,$pc, from which details of individual objects can be read off by comparison with Table 3 in the published paper. The absolute differences will of course grow with the estimates. The middle panel of Figure 3 shows the cumulative distribution of the ratio $d_{\rm ph}^{\rm med}/d_{\rm ph}^{\rm av}$ for all objects with $d_{\rm ph}^{\rm av} < 10\,$pc. This shows, for example, that only 7.5% of the objects have $d_{\rm ph}^{\rm med}/d_{\rm ph}^{\rm av} < 0.95$, and all objects have $d_{\rm ph}^{\rm med}/d_{\rm ph}^{\rm av} < 1.02$. The other two panels plot the ratio for the perihelion time and speed ratios. These plots shows that for the vast bulk of the encounters, the fractional difference between the mean and median is very small.

The number of objects with $d_{\rm ph}^{\rm med}$ below (0.5,1,2,3,5,10) pc is (5,17,75,188,465,1628) respectively (cf. Table 2 in the published paper).

The top panel of Figure 4 shows the median perihelion distances vs. time for all encounters with $d_{\rm ph}^{\rm med} < 10\,$pc. This looks very similar to the corresponding plot in the published paper (the top panel of Figure 2; the plotting ranges are identical and happen to include the full time range in $t_{\rm ph}^{\rm med}$). The lower two panels in Figure 4 are reproductions of the lower two panels of Figure 2 in the published paper, but here with the median estimates for the same objects overplotted in red. This shows that the median differs very little from the mean compared to the 90% confidence interval for the close encounters.

In conclusion, my overall results on close encounters are little effected by the choice of the mean or median as an estimator. The median estimates are available online in the file `peristats_median`.

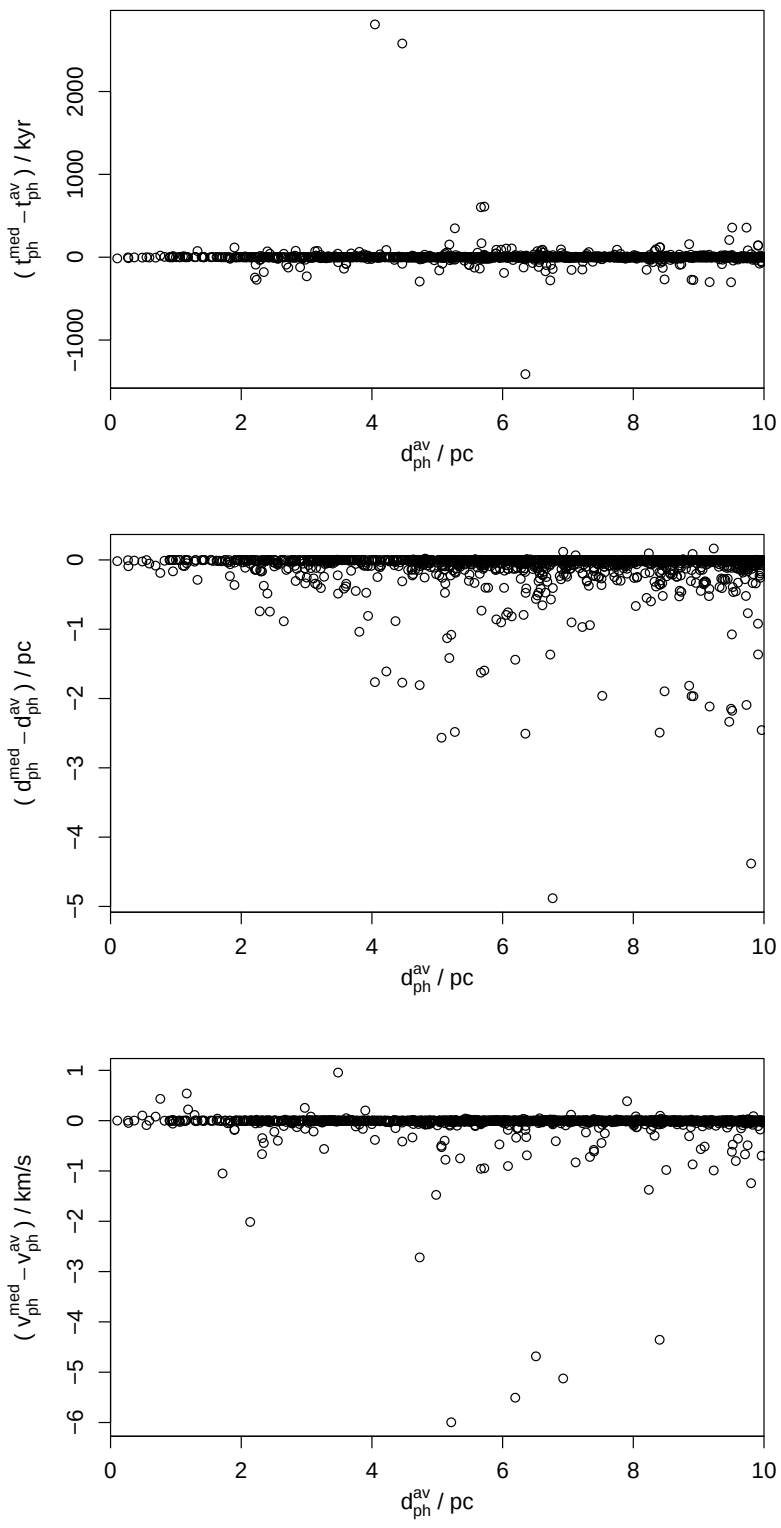

Figure 1: Comparison of median with mean. Each panel shows the difference between the median and mean estimate of the perihelion time, distance, and speed (from top to bottom) as a function of the mean perihelion distance, for all objects with $d_{\rm ph}^{\rm av} < 10\,{\rm pc}$. The vertical axes are selected to include all 1548 objects.

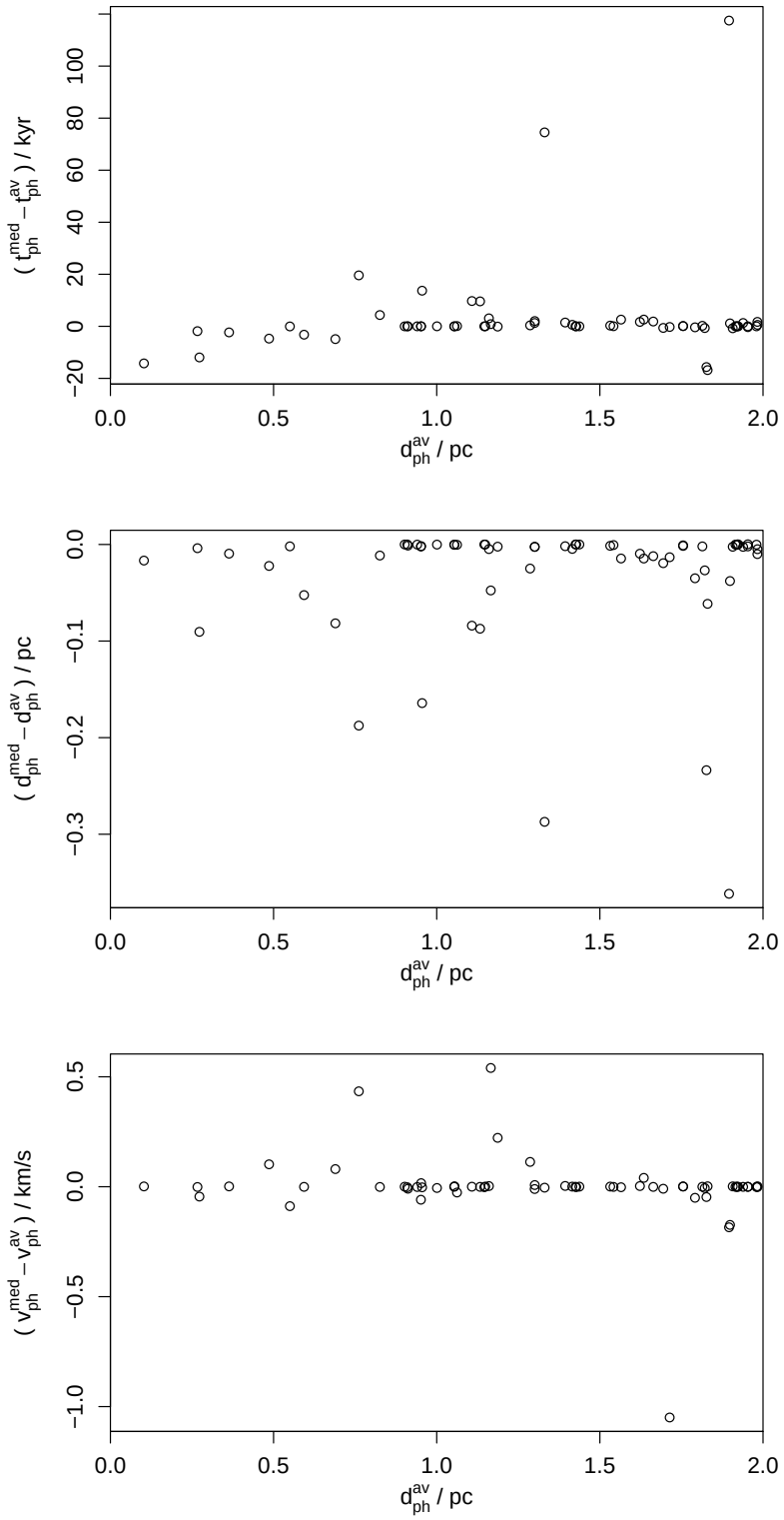

Figure 2: As Figure 1 but for all 65 objects with $d_{\rm ph}^{\rm av} < 2\,{\rm pc}$.

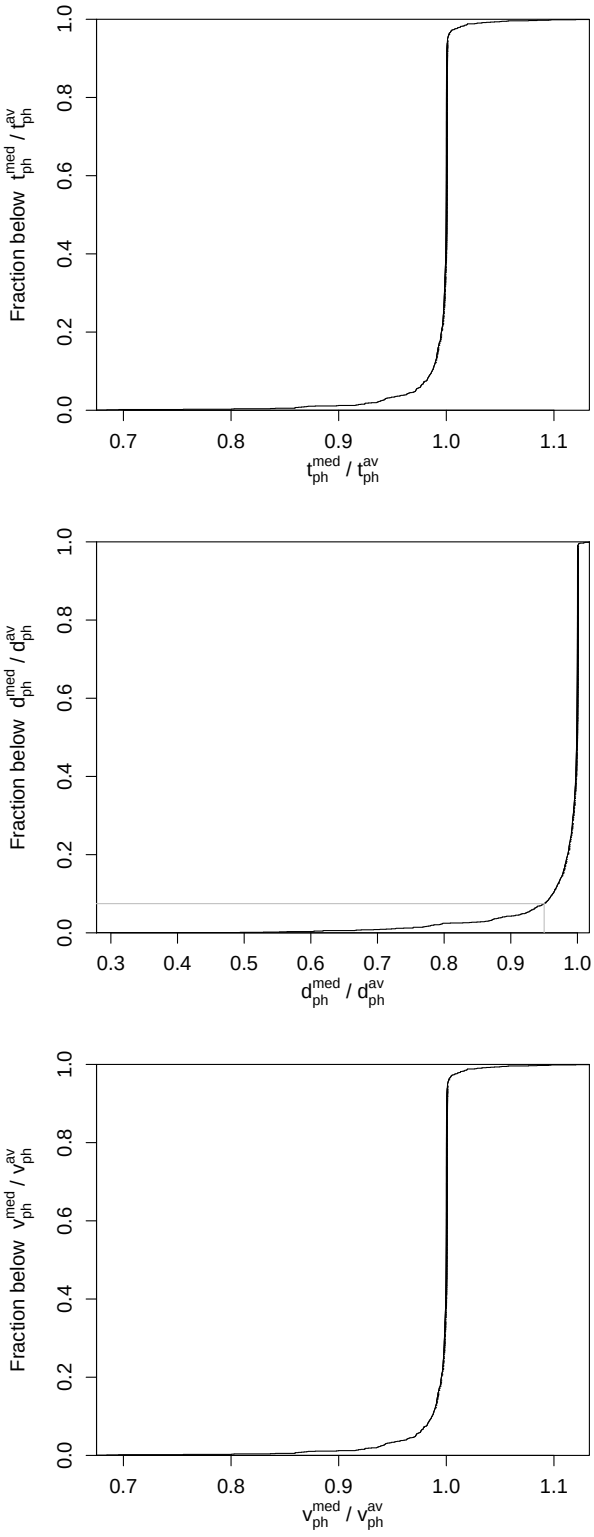

Figure 3: Cumulative distribution of the ratio of median to mean of the perihelion estimates, for all objects with $d_{\rm ph}^{\rm av} < 10\,$pc. The middle panel shows the fraction of objects (with $d_{\rm ph}^{\rm av} < 10\,$pc) which have $d_{\rm ph}^{\rm med}/d_{\rm ph}^{\rm av}$ less than the amount shown on the horizontal axis. The grey line shows that 7.5% of the objects have $d_{\rm ph}^{\rm med}/d_{\rm ph}^{\rm av} < 0.95$. The top and bottom panels show the same for $t_{\rm ph}^{\rm med}/t_{\rm ph}^{\rm av}$ and $\mathrm{v}_{\rm ph}^{\rm med}/\mathrm{v}_{\rm ph}^{\rm av}$ respectively. In all cases the full range of the ratio is plotted.

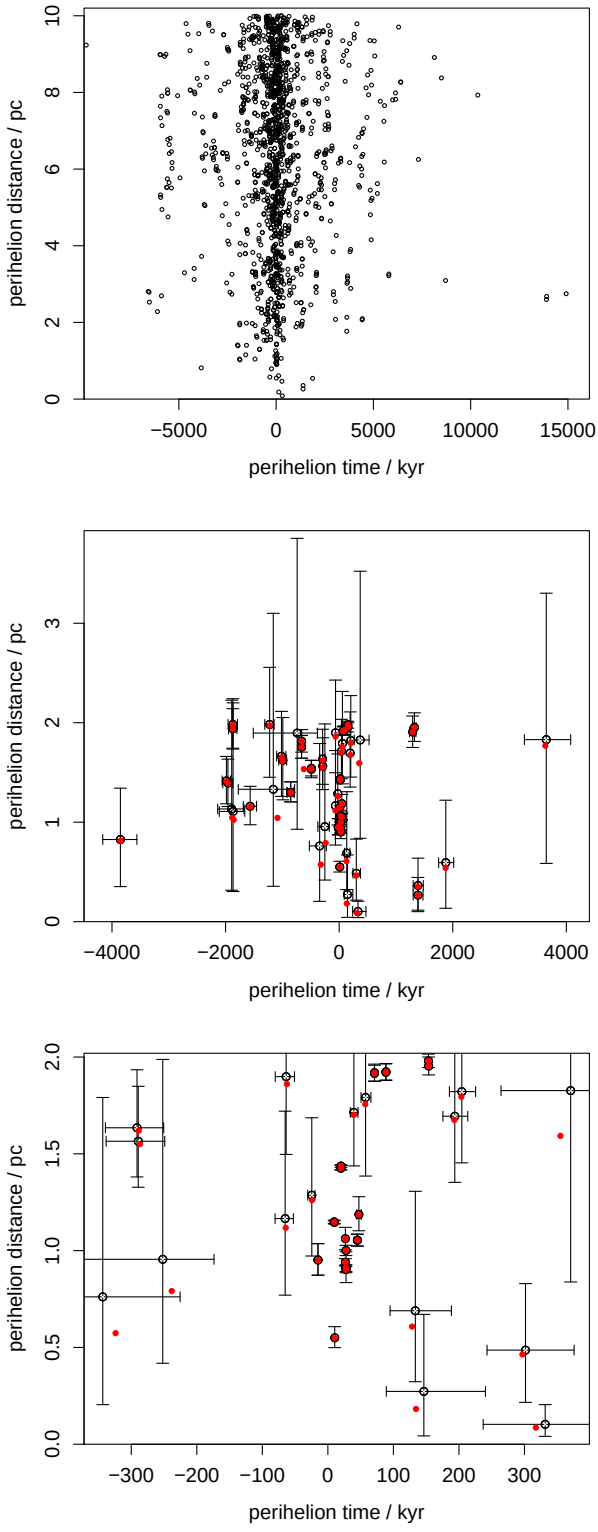

Figure 4: The top panel shows the median perihelion distance, $d_{\rm ph}^{\rm med}$, vs. median perihelion time, $t_{\rm ph}^{\rm med}$, for all objects with $d_{\rm ph}^{\rm med} < 10\,$pc. The middle panel shows all objects with $d_{\rm ph}^{\rm av} < 2\,$pc and the bottom panel is a zoom of this. The black points in these two panels are $(t_{\rm ph}^{\rm av}, d_{\rm ph}^{\rm av})$ and the nearest red points are $(t_{\rm ph}^{\rm med}, d_{\rm ph}^{\rm med})$ for the same objects. The ends of the error bars in the lower two panels denote the 5% and 95% quantiles of the distributions for each object, which together form a 90% confidence interval. These could equally well be applied to the median estimates (the red points) instead. Cf. Figure 2 in the published paper.