Working with astrometric data
- warnings and caveats -

X. Luri (U. Bastian, Stefan Jordan)

Scientist's dream

- **Error-free data**
  - No random errors
  - No biases
  - No correlations
- **Complete sample**
  - No censorships
- **Direct measurements**
  - No transformations
  - No assumptions

Never ever available

Degeneracy for less than 1 year

Errors 1: biases

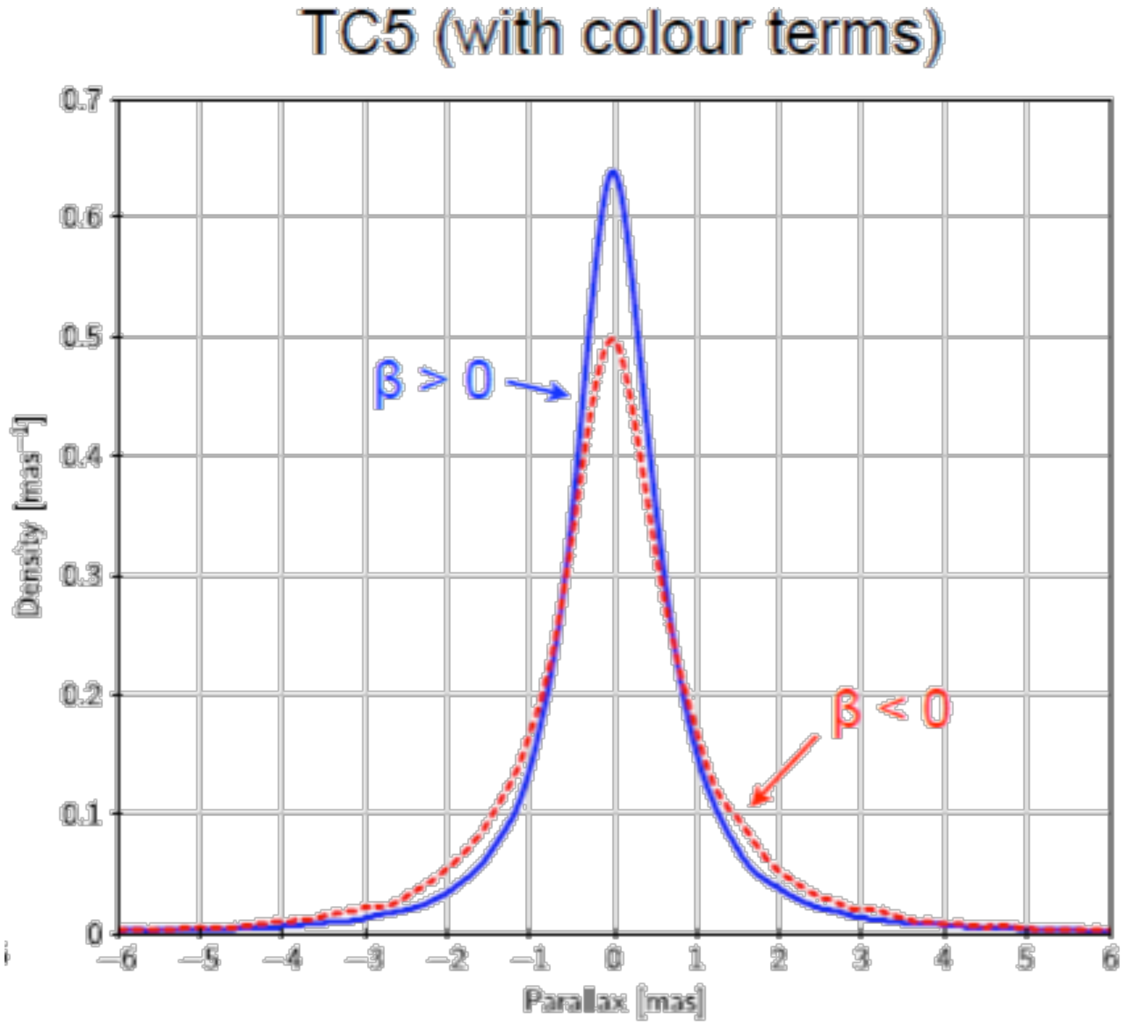
Bias:

your measurement is systematically too large or too small

For DR1 parallaxes:

- Probable global zero-point offset present; -0.04 mas found during validation
- Colour dependent and spatially correlated systematic errors at the level of 0.2 mas
- Over large spatial scales, the parallax zero-point variations reach an amplitude of 0.3 mas
- Over a few smaller areas (2 degree radius), much larger parallax biases may occur of up to 1 mas
- There may be specific problems in a few individual cases

Global zero point from QSO parallaxes

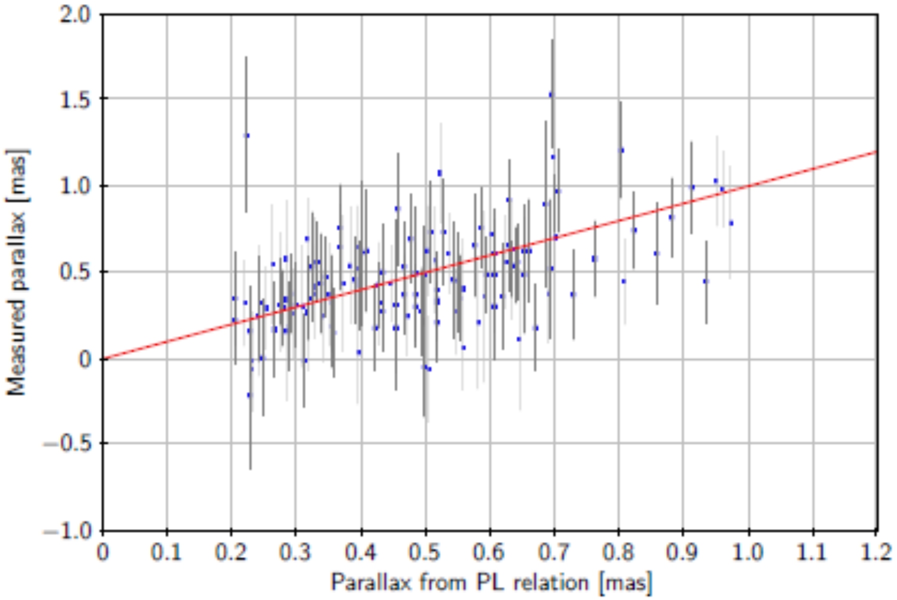


$\beta > 0$: $\mathrm{med}(\varpi) = +0.002$ mas

$\beta < 0$: $\mathrm{med}(\varpi) = -0.020$ mas

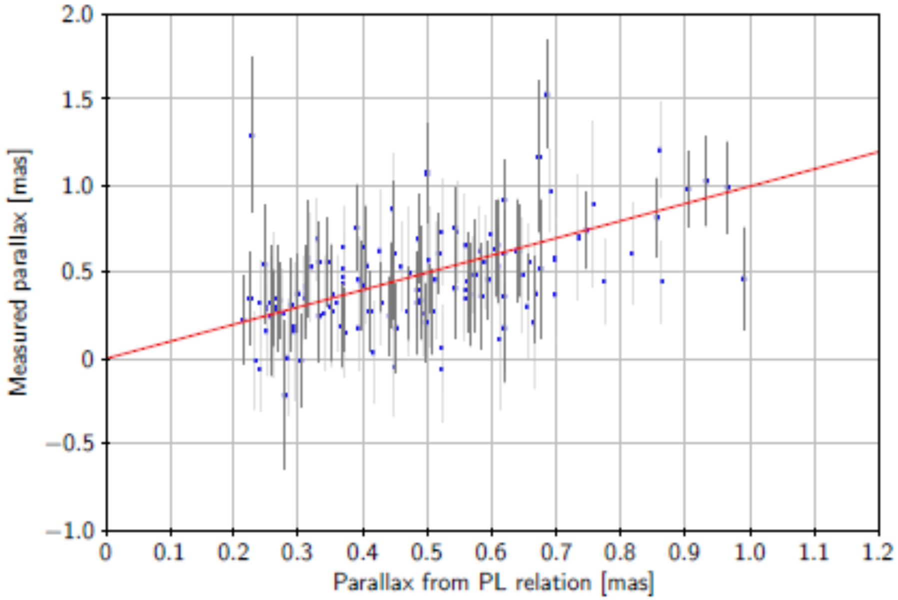
Global zero point from Cepheids

P-L relation from Tammann et al. (2003):
$M_V = -3.141 \log P - 0.820$



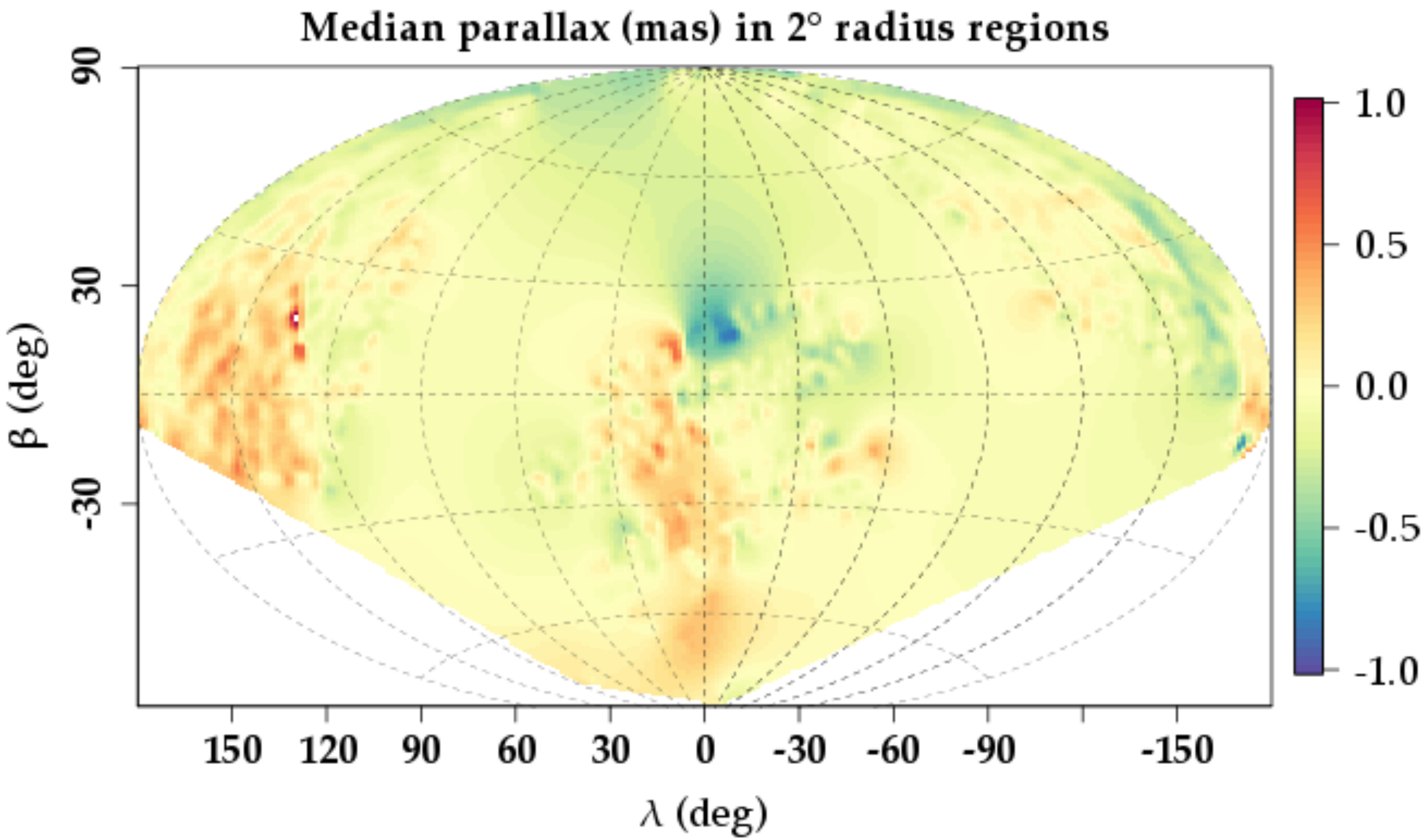
med($\Delta\varpi$) = −0.015 mas

P-L relation from Fouqué et al. (2007)
$M_V = -2.678 \log P - 1.275$

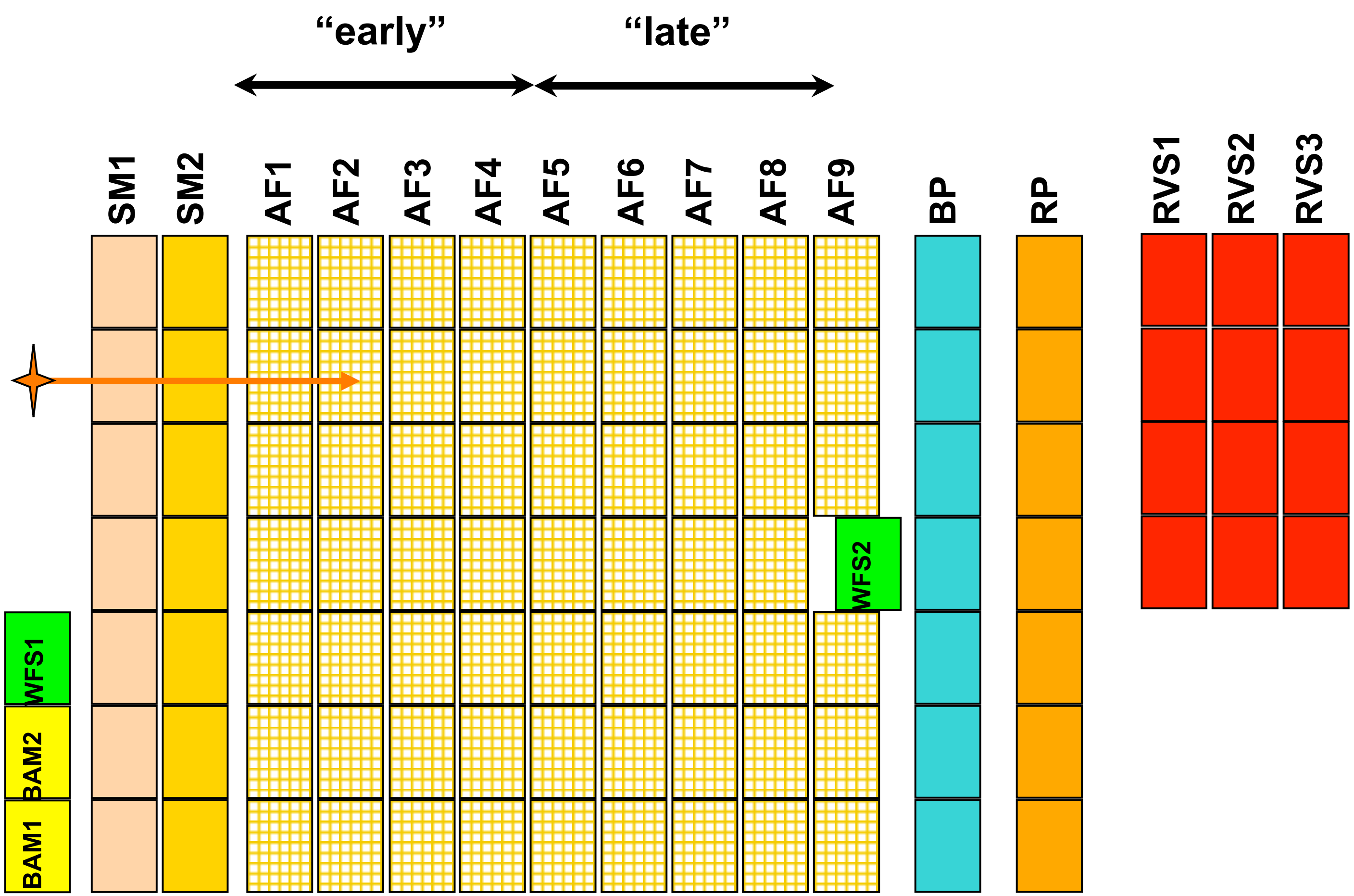


med($\Delta\varpi$) = −0.017 mas

Regional effects from QSOs
(ecliptic coordinates)



Split FoV

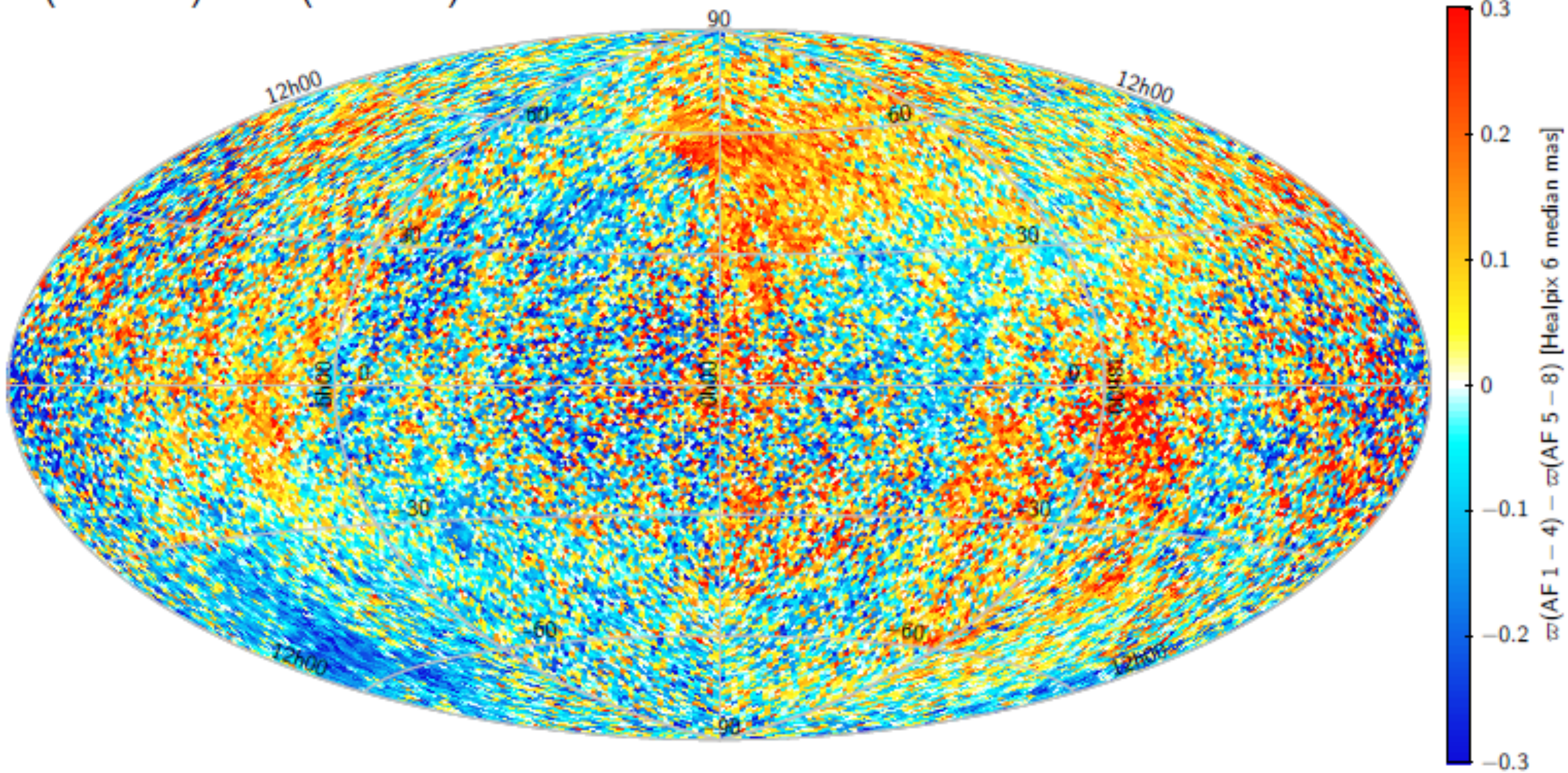


"early"
"late"
SM1
SM2
AF1
AF2
AF3
AF4
AF5
AF6
AF7
AF8
AF9
BP
RP
RVS1
RVS2
RVS3
WFS1
BAM2
BAM1
WFS2

Regional effects from split FOV solutions

(equatorial coordinates)

ϖ(AF1-4) − ϖ(AF5-8)



Median value per pixel (~1 deg^2)

A. Bombrun

How to take this into account

- You can introduce a global zero-point offset to use the parallaxes (suggested -0.04 mas)

- **You cannot correct the regional features:** if we could, we would already have corrected them. We have indications that these zero points may be present, but no more.

- For most of the sky assume an additional systematic error of 0.3 mas; your derived standard errors for anything cannot go below this value

  $$\boldsymbol{\varpi \pm \sigma_{\varpi}} \textbf{ (random) } \boldsymbol{\pm} \textbf{ 0.3 mas (syst.)}$$

- For a few smaller regions be aware that the systematics might reach 1 mas

This is possibly the sole aspect in which Gaia DR1
is not better than Hipparcos
(apart from the incompleteness for the brightest stars)

More specifically: treat separately random error and bias, but if you must combine them, a **worst case** formula can be as follows

- **For individual parallaxes:** to be on the safe side add 0.3 mas to the standard uncertainty

$$\sigma_{Total} \approx \mathrm{sqrt}(\sigma^2_{Std}+0.3^2)$$

- **When averaging parallaxes for groups of stars:** the random error will decrease as sqrt(N) but the systematic error (0.3 mas) will <u>not</u> decrease

$$\sigma_{final} \approx \mathrm{sqrt}(\sigma^2_{averageStd}+0.3^2)$$

where $\sigma_{averageStd}$ decrease is the formal standard deviation of the average, computed in the usual way from the sigmas of the individual values in the average (giving essentially the sqrt(N) reduction).

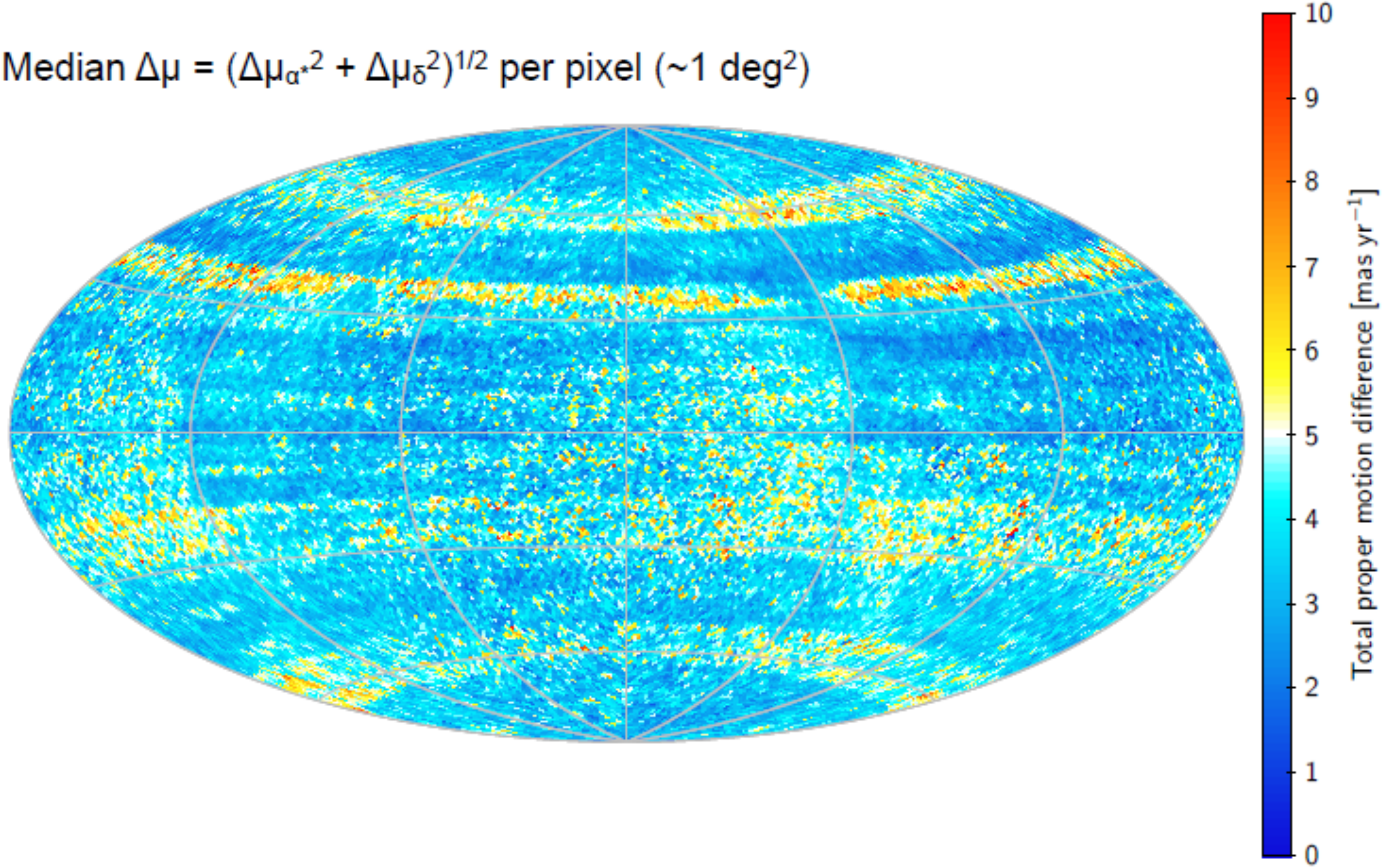
- **Don't try to get a "zonal correction" from previous figures**, it's too risky

For DR1 proper motions and positions:

- In this case Gaia data is the best available, by far.
- We do not have means to do a check as precise as the one done for parallaxes, but there are no indications of any significant bias
- For positions remember that for comparison purposes you will likely have to convert them to another epoch. You should propagate the errors accordingly.

Comparison with Tycho-2 shows that catalogue's systematics (not Gaia's)

Median $\Delta\mu = (\Delta\mu_{\alpha*}^2 + \Delta\mu_\delta^2)^{1/2}$ per pixel (~1 deg^2)



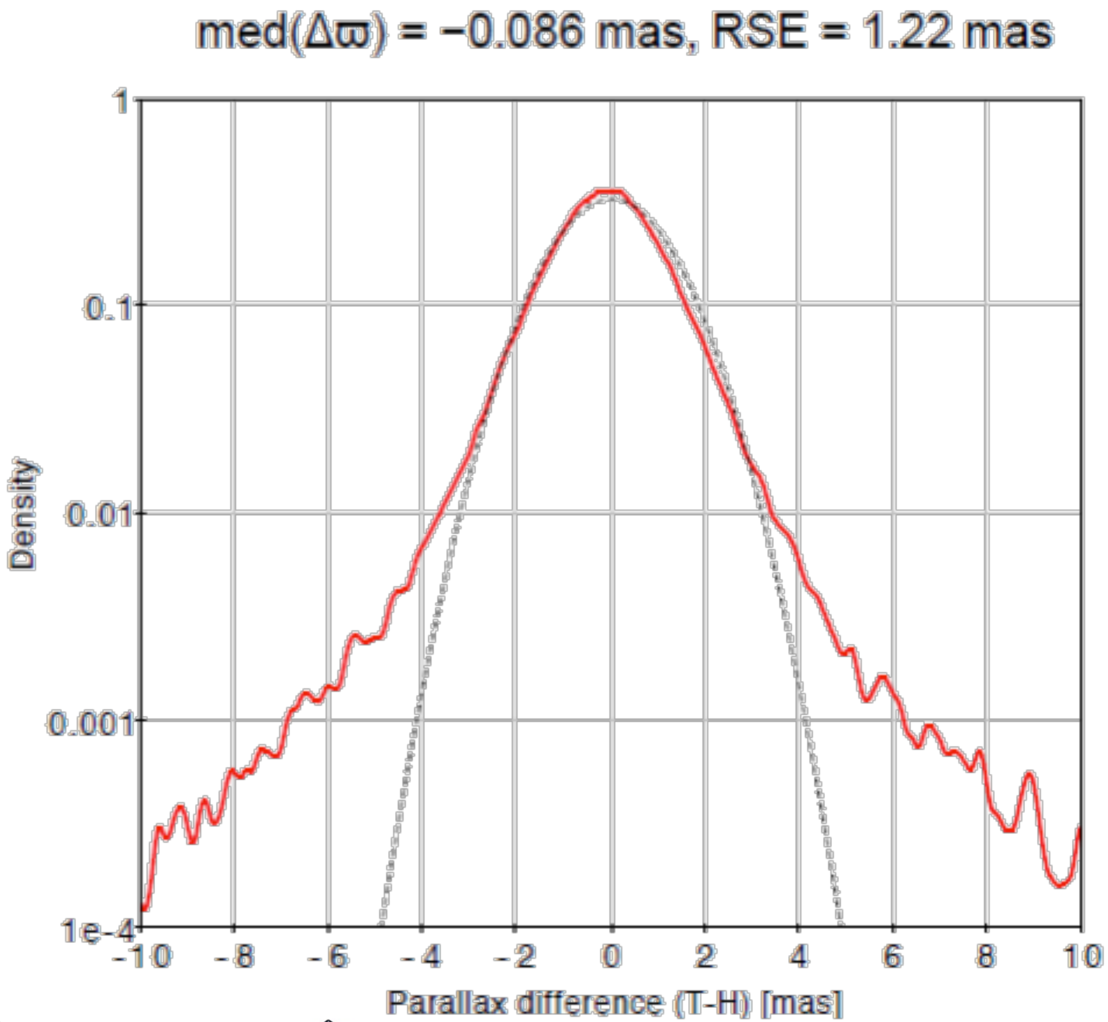
Errors 2: random errors

Random error:

your measurements are randomly distributed around the true value

- Each measurement in the catalogue comes with a formal error
- Random errors in Gaia are quasi-normal. The formal error can be assimilated to the variance of a normal distribution around the true value.
- Published formal errors for Gaia DR1 may be slightly overestimated

Warning: comparison with Hipparcos shows deviation from normality beyond ~2σ

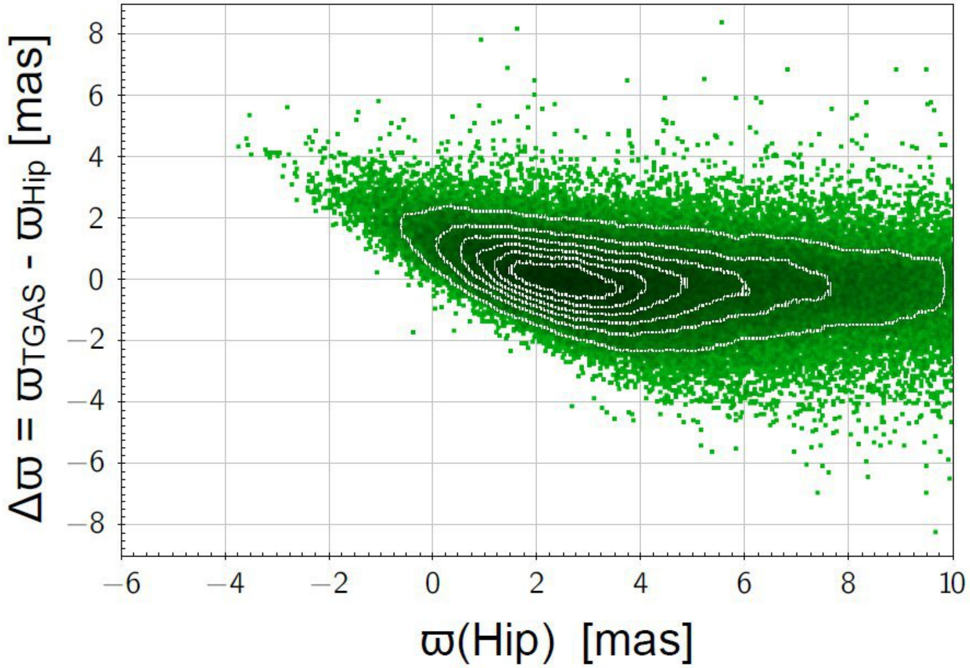


To take into account for outlier analysis

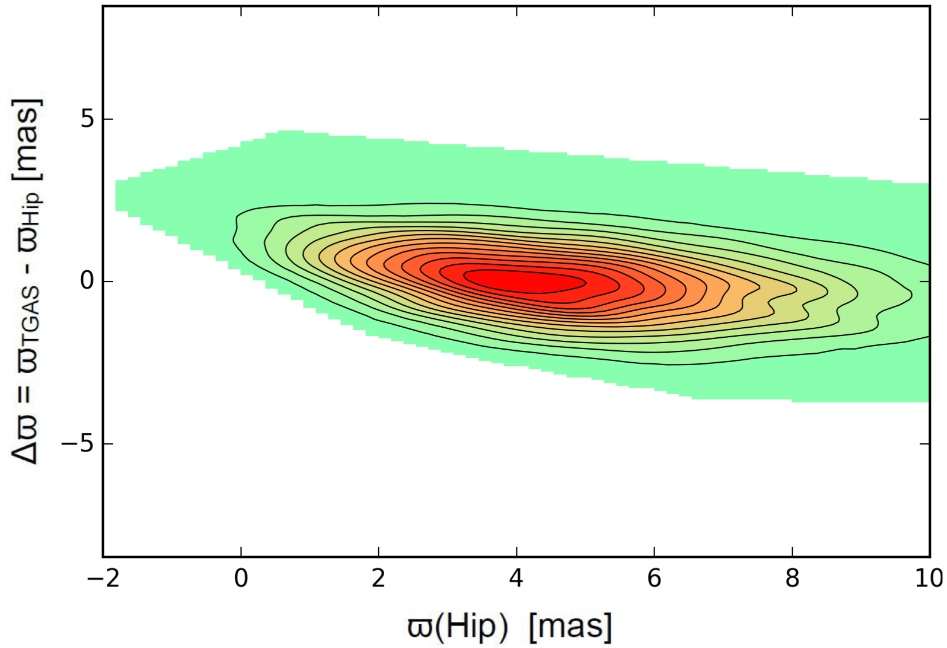
Warning: when comparing with other sources of trigonometric parallaxes take into account the properties of the error distributions

TGAS vs Hipparcos

Observations



Simulations

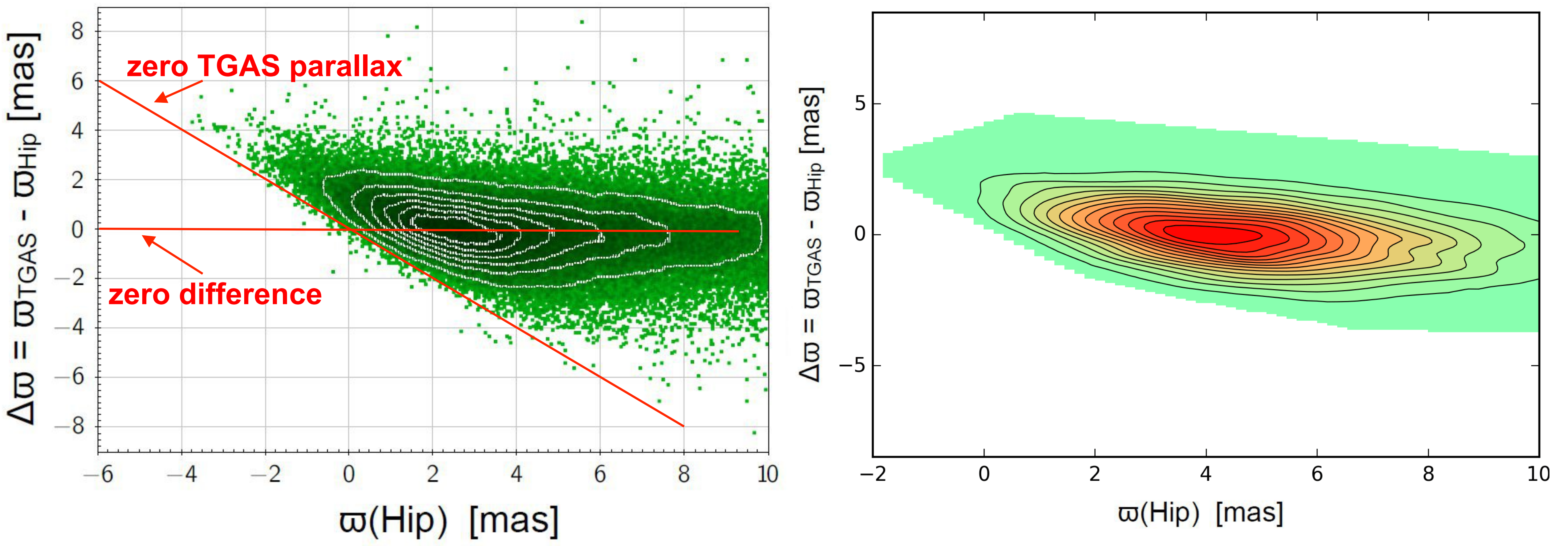


The "slope" at small parallaxes is not a bias in either TGAS or HIP, simply due to the different size of the errors in the two catalogues!

Warning: when comparing with other sources of trigonometric parallaxes take into account the properties of the error distributions

TGAS vs Hipparcos

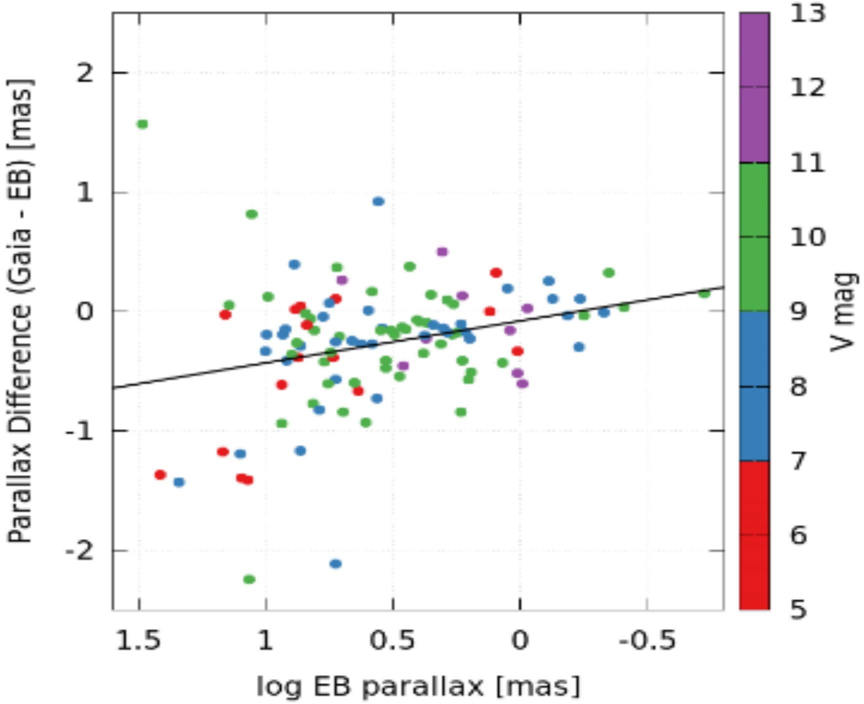
Observations | Simulations



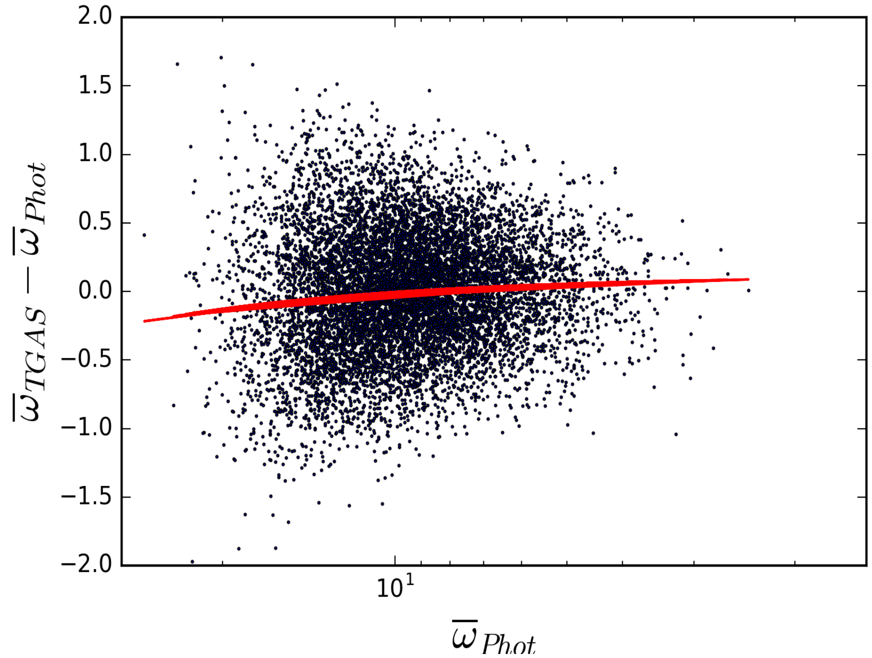
The "slope" at small parallaxes is not a bias in either TGAS or HIP, simply due to the different size of the errors in the two catalogues!

Eclipsing binaries parallaxes vs TGAS

arXiv:1609.05390v3



Simulation



The overall “slope” is due to the different error distributions in parallax (lognormal for photometric, normal for trigonometric)

Errors 3: correlations

Correlation:

the measurements of several quantities are not independent from each other

- Whenever you take linear combinations of such quantities, the correlations have to be taken into account in the error calculus (and even more so for non-linear functions)

- The errors in the five astrometric parameters provided are not independent
- The ten correlations between these parameters are provided in the Gaia DR1 archives (correlation matrix)

Errors 3: correlations

Correlation:

the measurements of several quantities are not independent from each other.

- Whenever you take linear combinations of such quantities, the correlations have to be taken into account in the error calculus (and even more so for non-linear functions !)

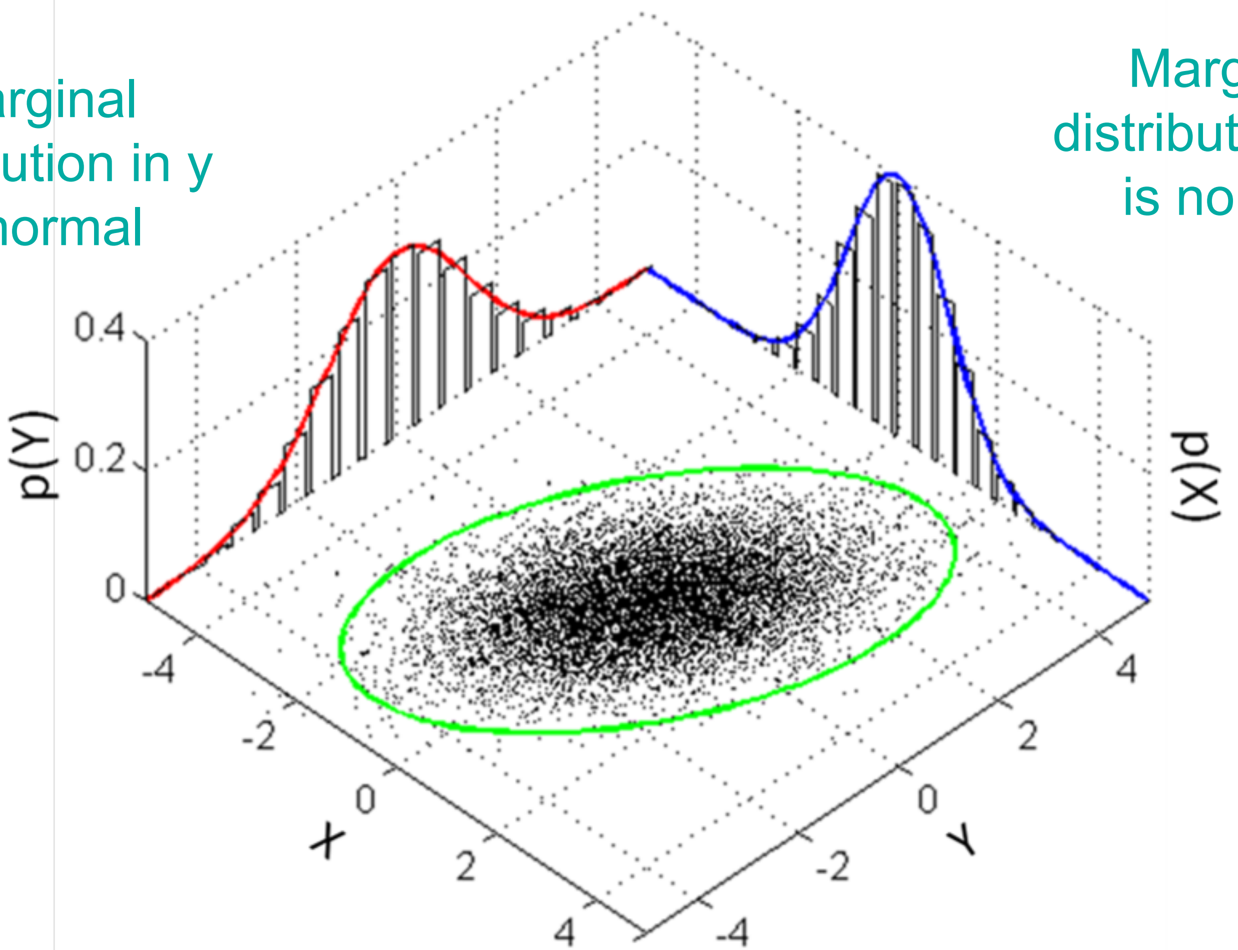
Variance of a sum: (x1+x2)
sigma^2 (x1+x2) = sigma^2(x1) + sigma^2 (x2) + 2 cov(x1,x2)
= sigma^2(x1) + sigma^2 (x2) + 2 sigma(x1) sigma (x2) corr(x1,x2)

Variance of any linear combination of two measured quantities, x1 and x2 : (ax1 + bx2)
sigma^2 = a^2 sigma^2(x1) + b^2 sigma^2 (x2) + 2ab cov(x1,x2)
= a^2 sigma^2(x1) + b^2 sigma^2 (x2) + 2ab sigma(x1) sigma (x2) corr(x1,x2)

Generally, for a whole set of linear combinations $\boldsymbol{y}$ of several correlated random variables $\boldsymbol{x}$:
If $\boldsymbol{y}$ = A'$\boldsymbol{x}$, then: Cov($\boldsymbol{y}$) = A' Cov($\boldsymbol{x}$) A = A' Sigma($\boldsymbol{x}$) Corr($\boldsymbol{x}$) Sigma'($\boldsymbol{x}$) A
where Cov and Corr indicate covariance and correlation matrices, Sigma($\boldsymbol{x}$) is a diagonal matrix having the sigmas of the components of x as elements, and A' is the relation matrix.
In the example above, for just two x and one y, the matrix A' is simply the row vector (a,b).

Example of two correlated parameters

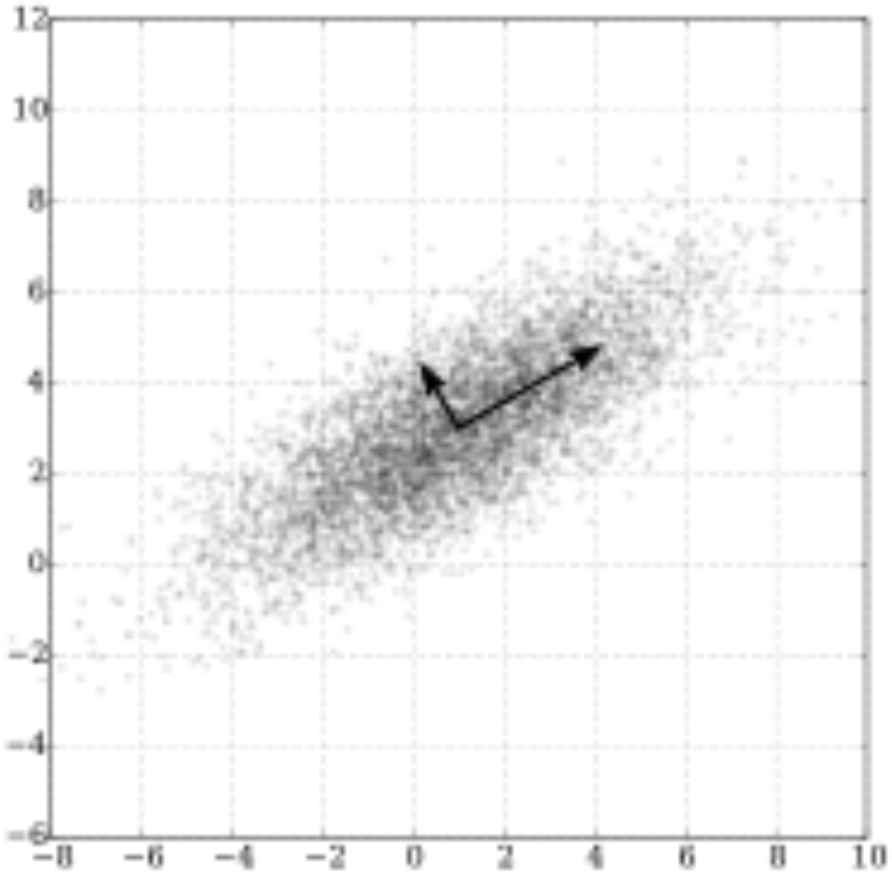
Marginal distribution in y is normal

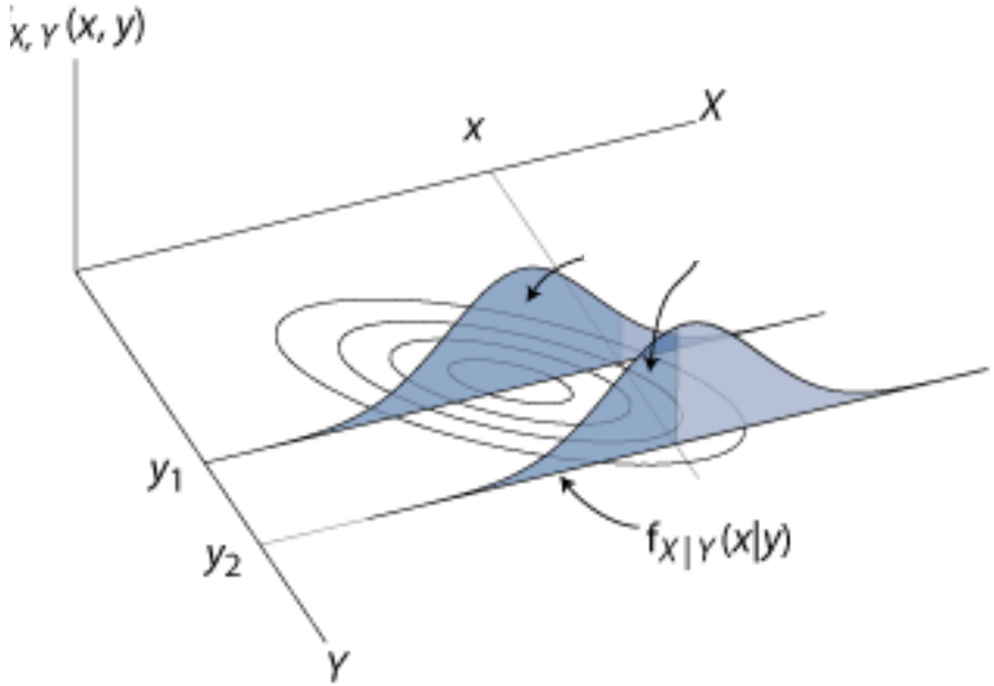


Marginal distribution in x is normal

By Bscan - Own work, CC0, https://commons.wikimedia.org/w/index.php?curid=25235145

Beware when using these quantities together



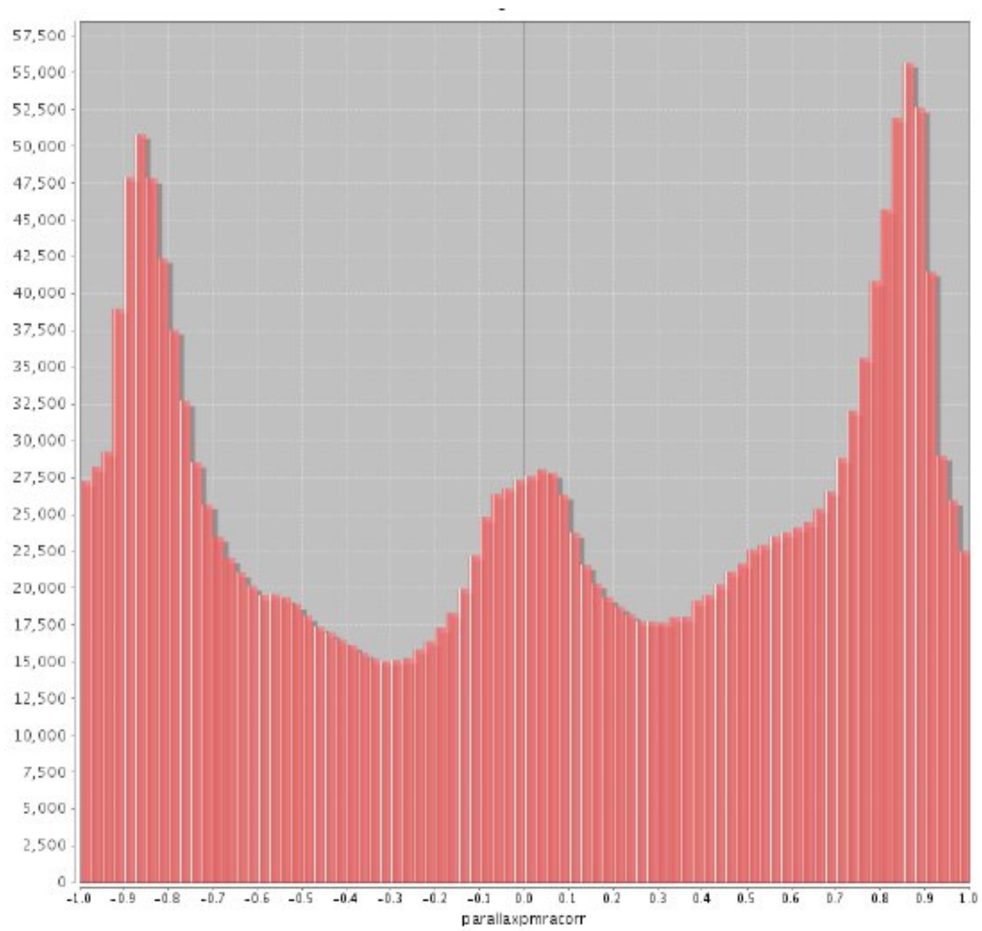


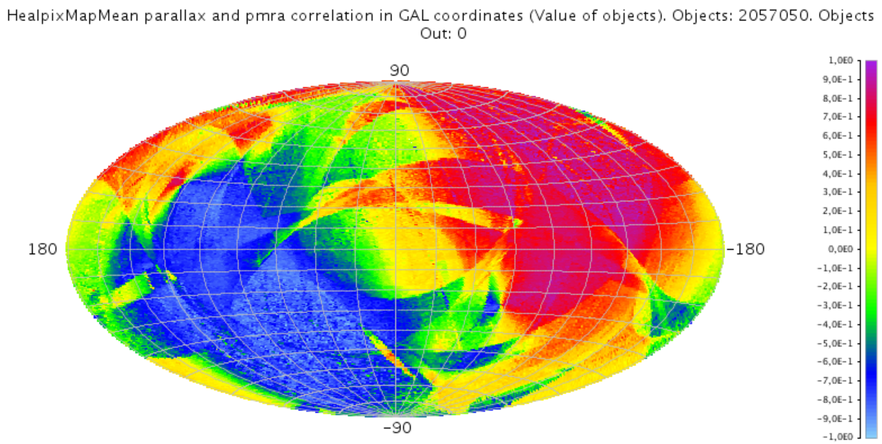
Examples of problematic use:

- Simple epoch propagation (!) pos&pm
- Calculation of proper directions pos&pm&parallax
- Proper motion in a given direction on the sky (other than north-south or east-west) proper-motion components
- Proper motion components in galactic or ecliptic coordinates proper-motion components

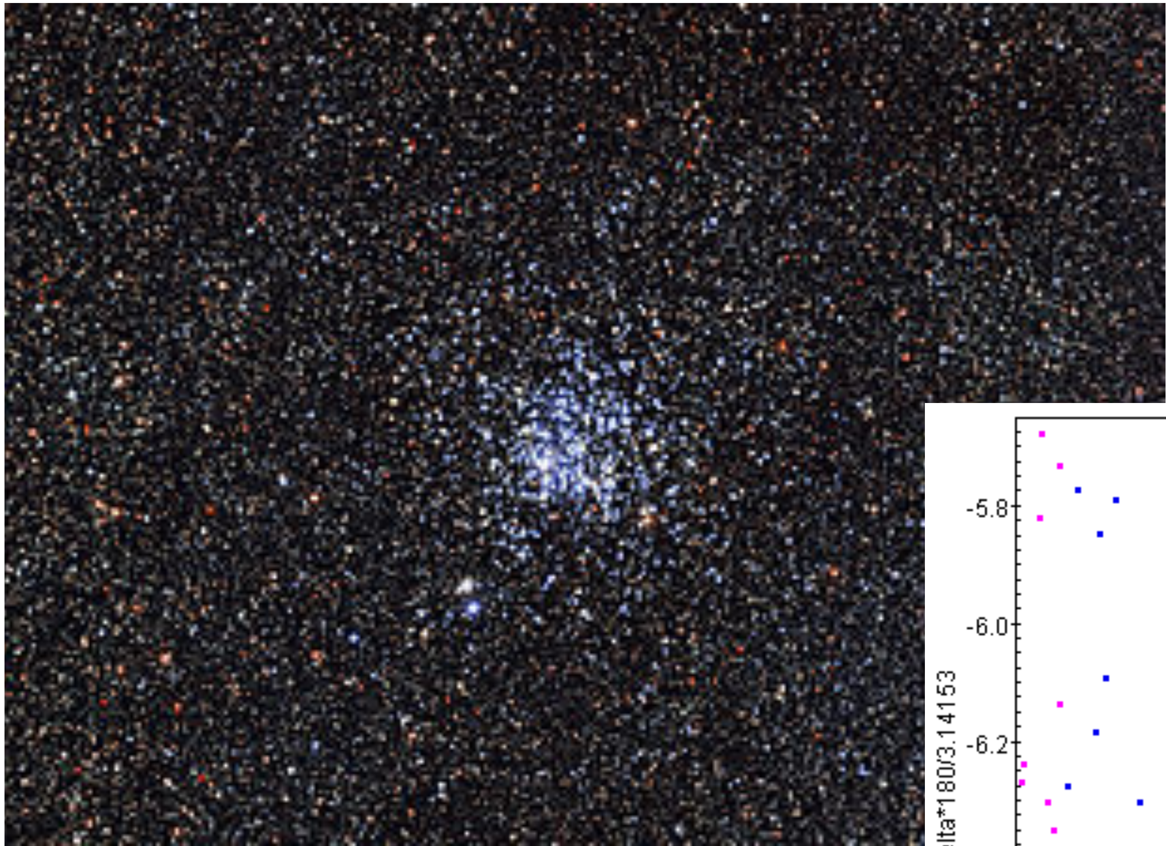
- More complex, non-linear example:
  Calculating the transversal velocities of a set of stars
  - The resulting dispersion of velocities is influenced by the errors in parallax and in proper motion; thus 3-dimensional case.
  - Its determination can not be done using the parallax and proper motion errors separately; the correlations have to be taken into account
  - But this time it's non-linear! The error distribution will no longer be Gaussian.
  - The A matrix of the previous page will become the Jacobian matrix of the local derivatives of the transversal velocity wrt parallax and pm components

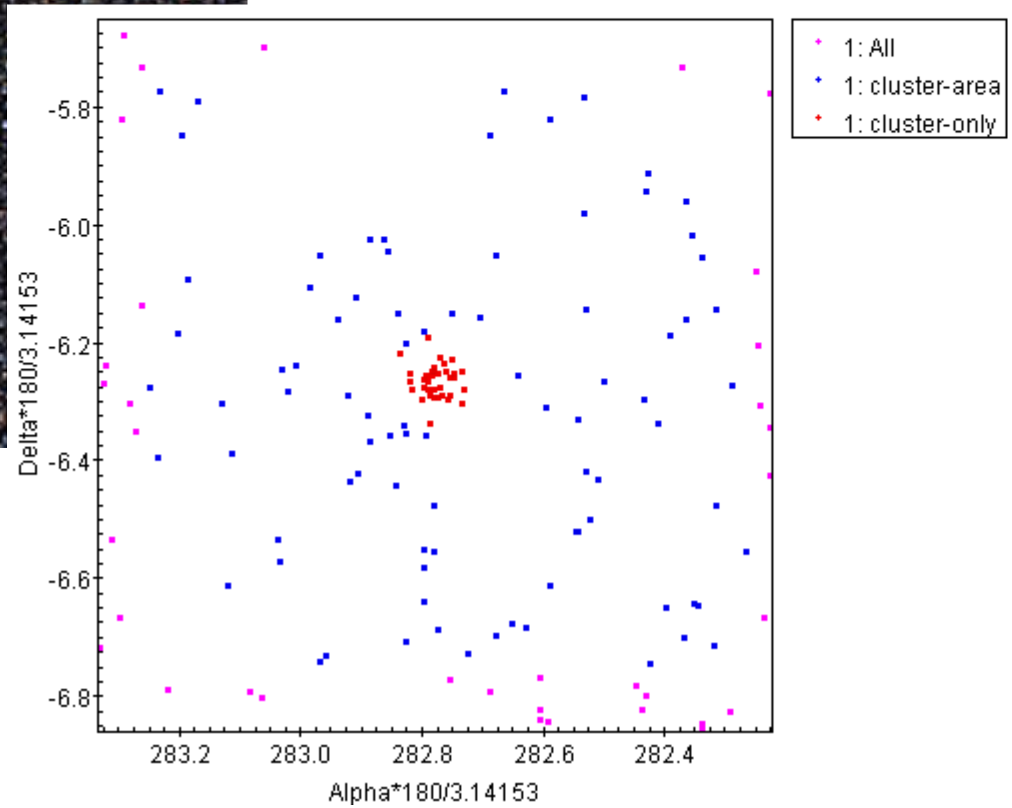
Beware: large and unevenly distributed correlations in DR1; example: PmRA-vs.-Parallax correlation



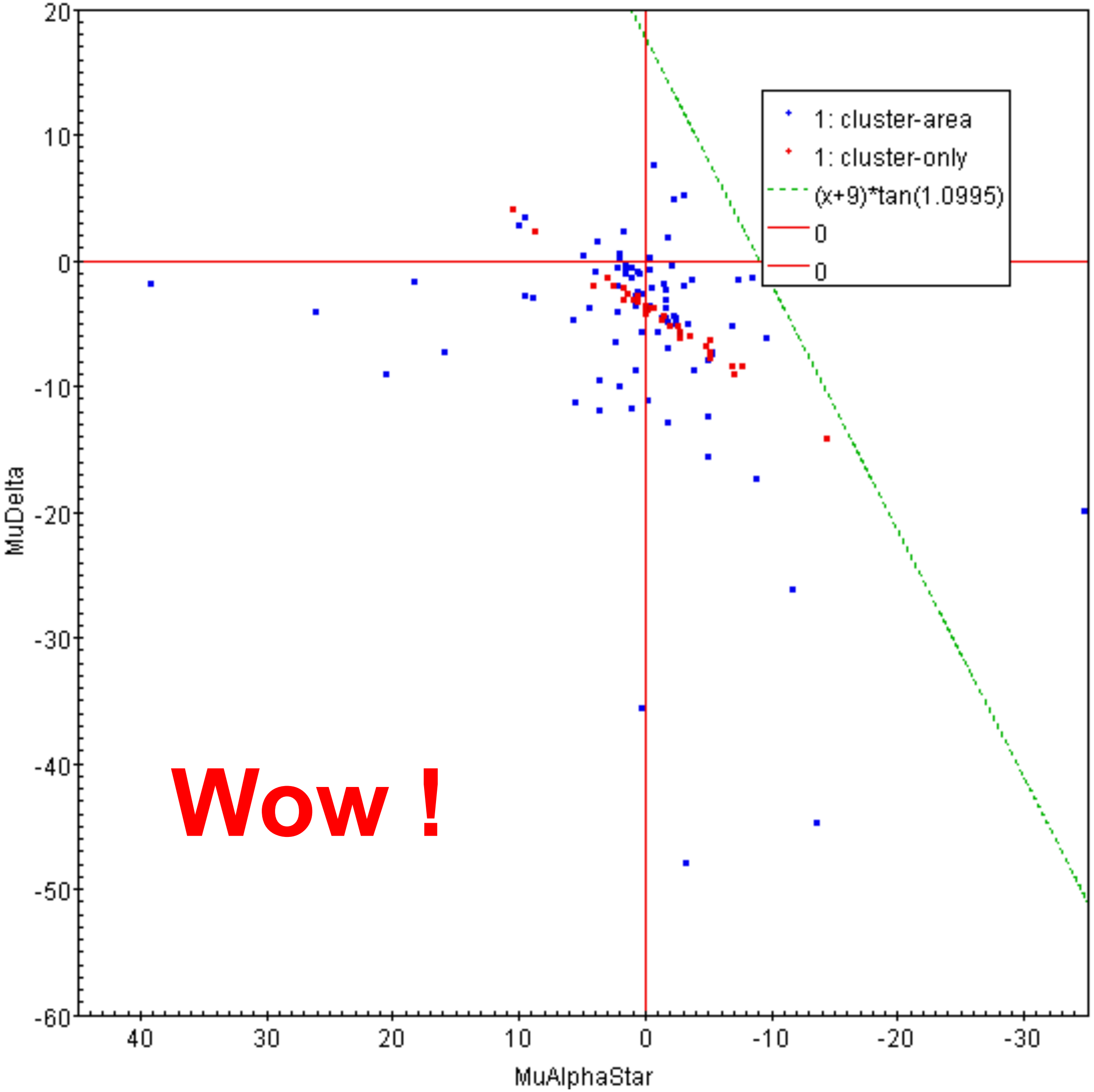


A really pretty example on correlations: M11

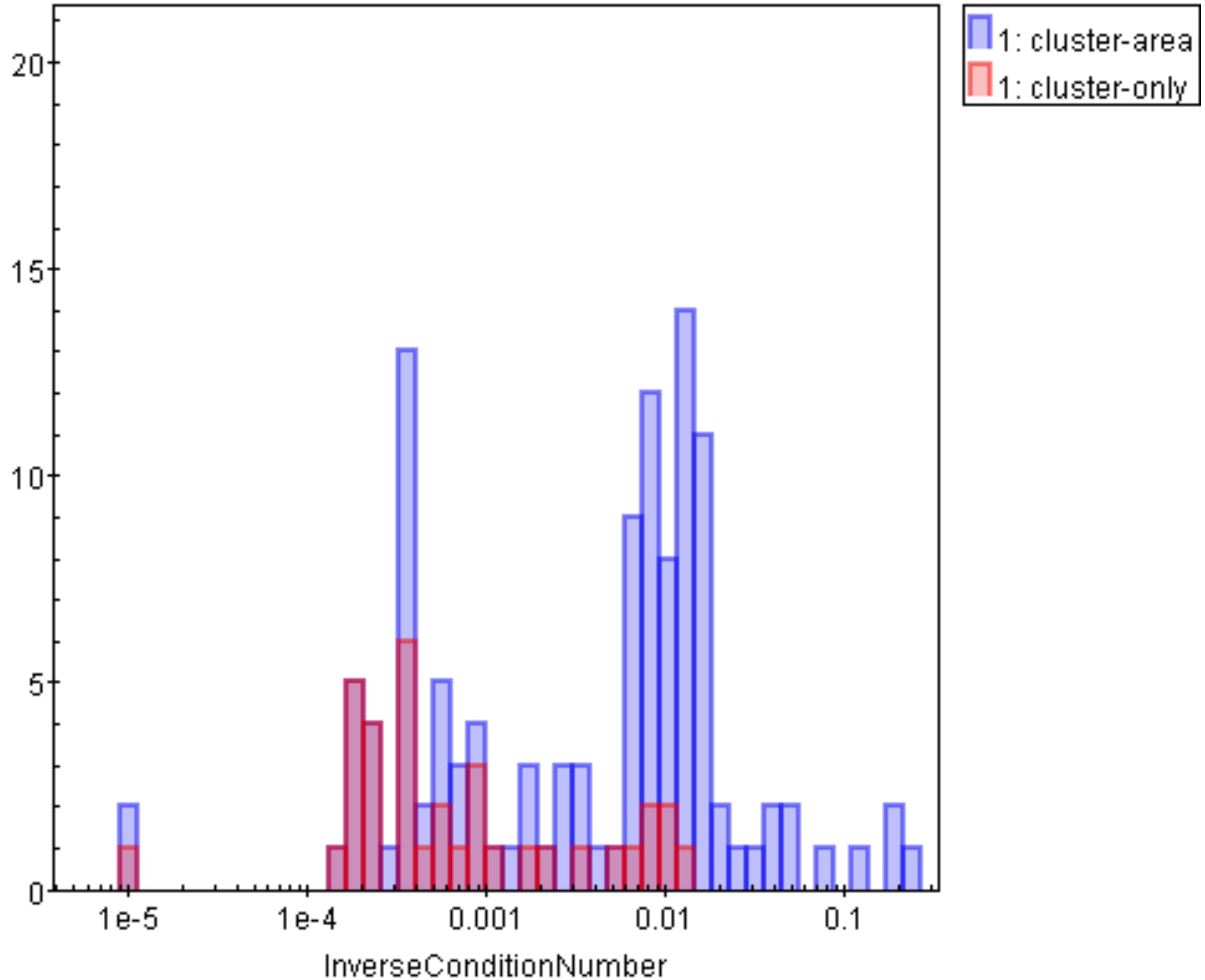




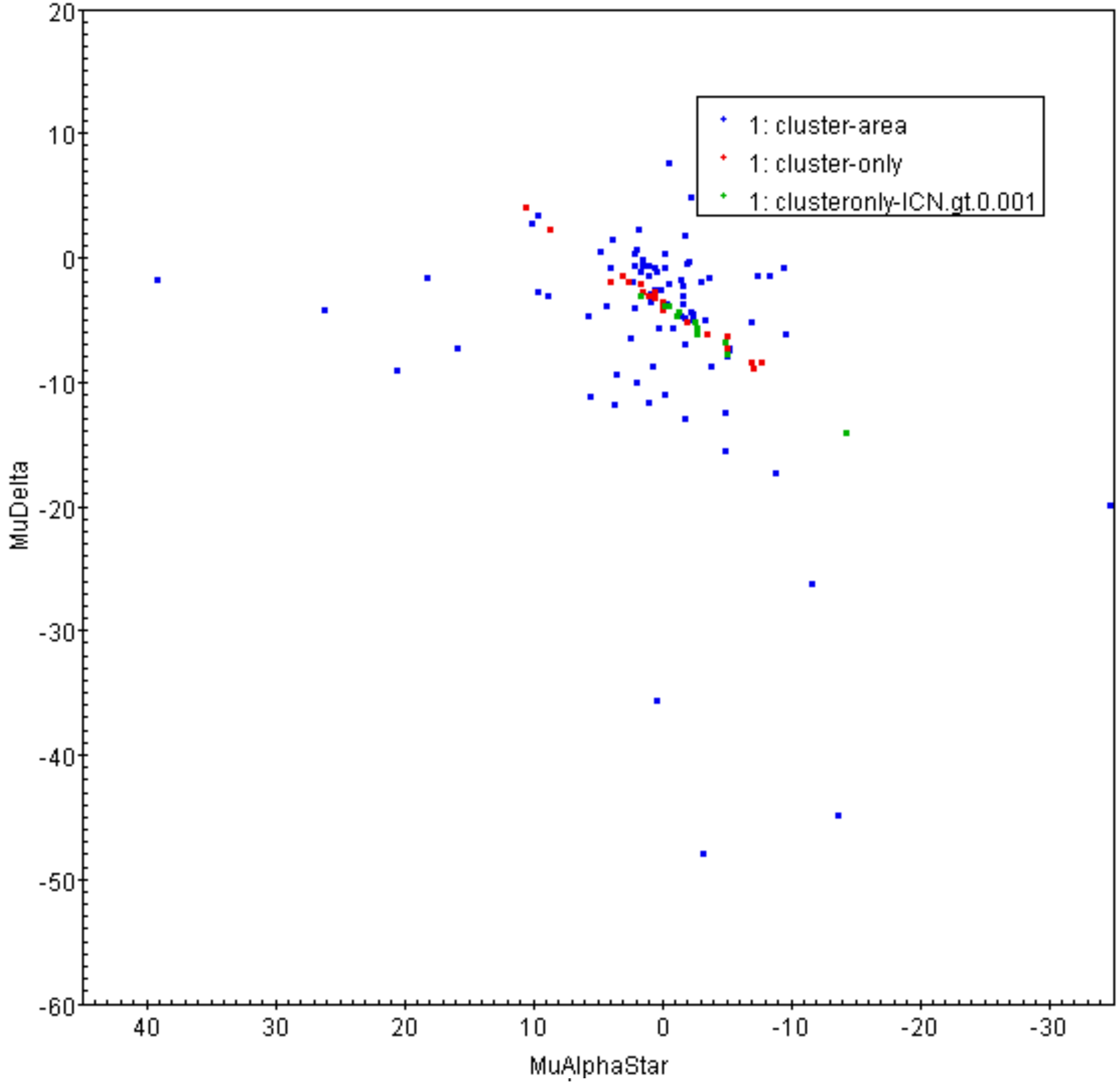
M11; proper motions in the AGIS-01 solution



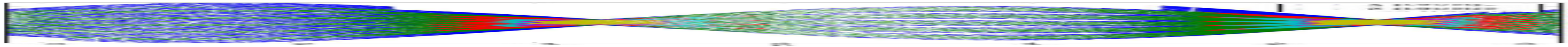
M11; scan coverage statistics

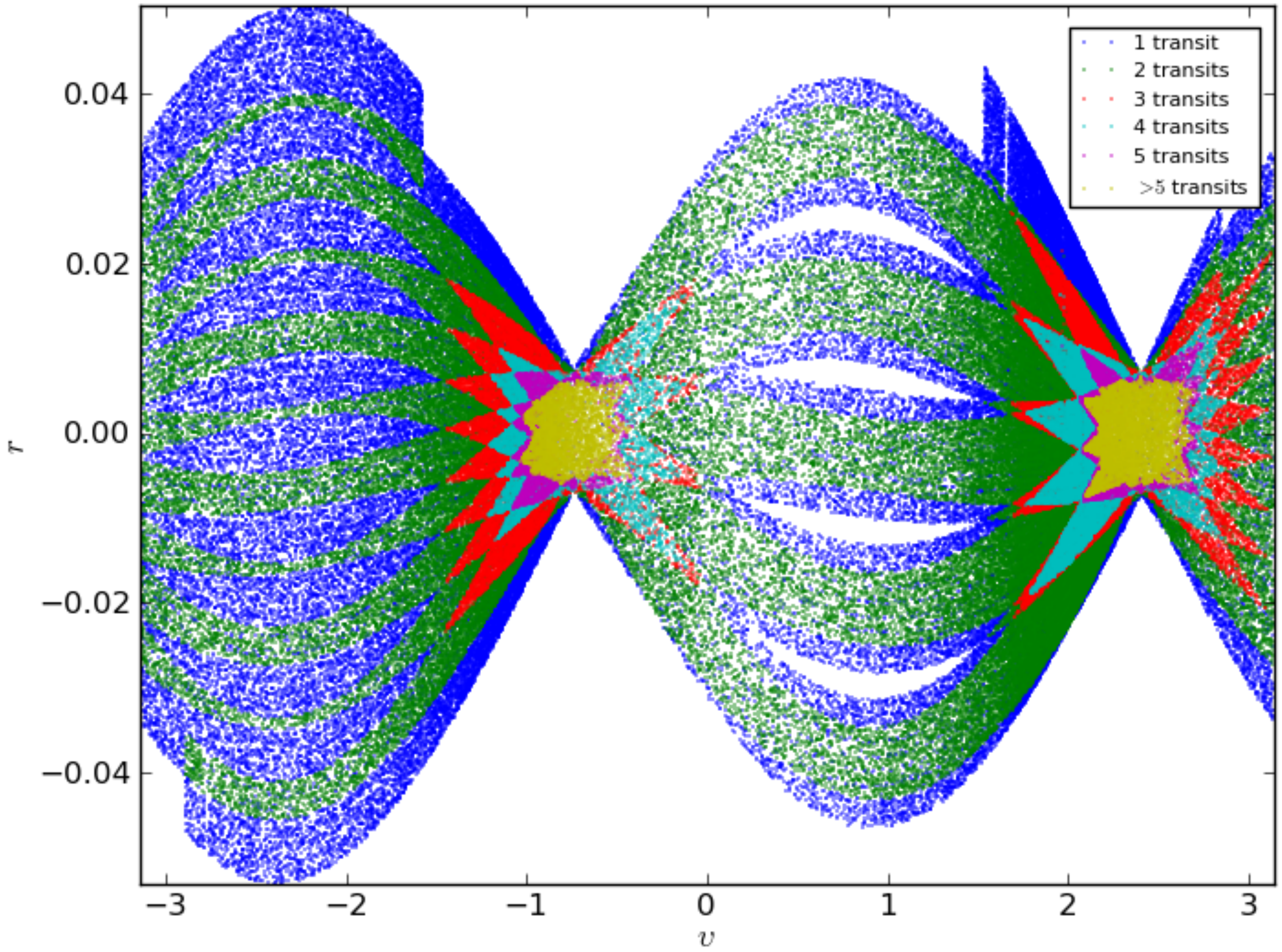


M11; selection of „better-observed“ stars



Just bad luck for poor M11:





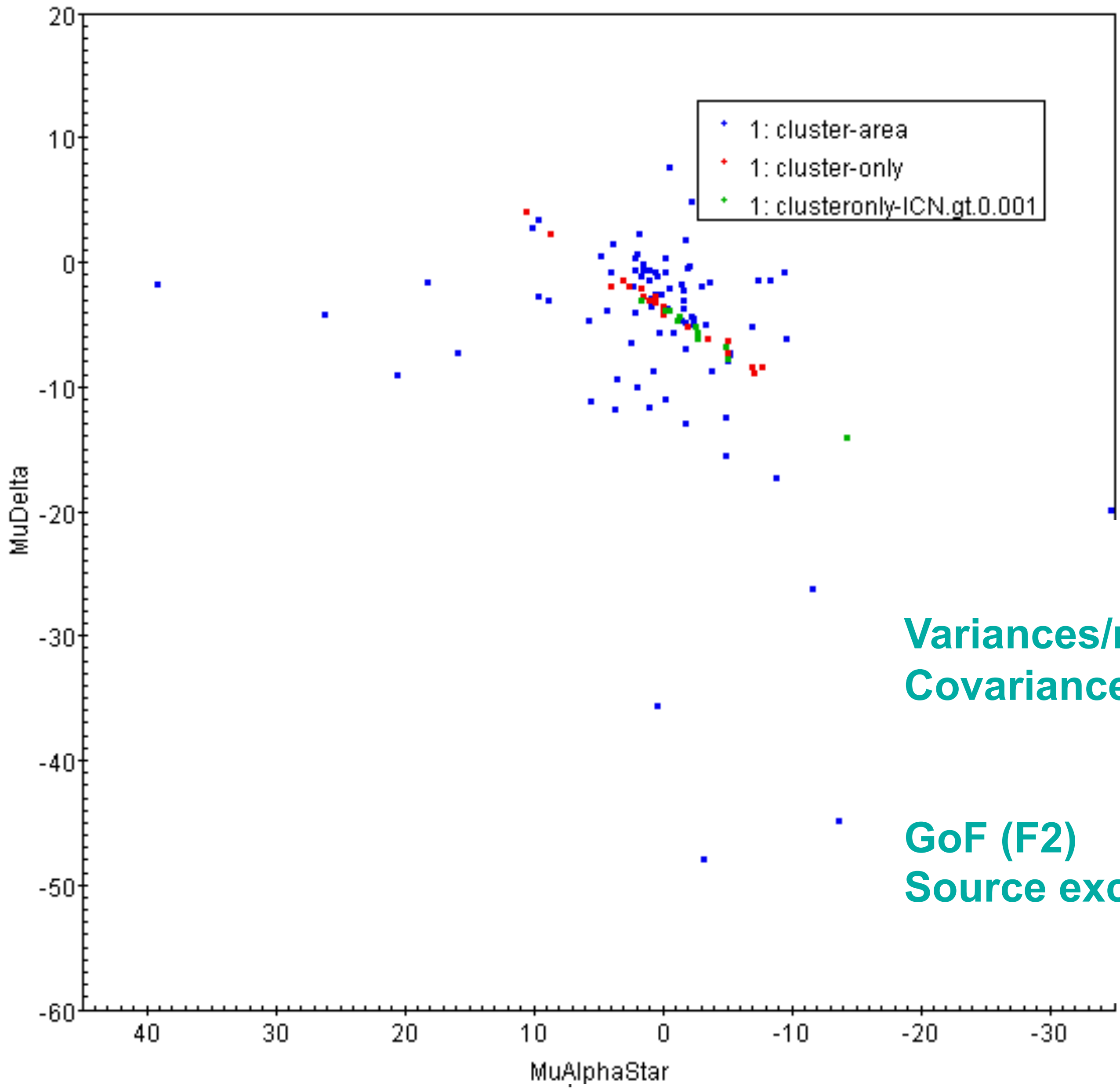
6 transits

all but one ...

slits

hickups

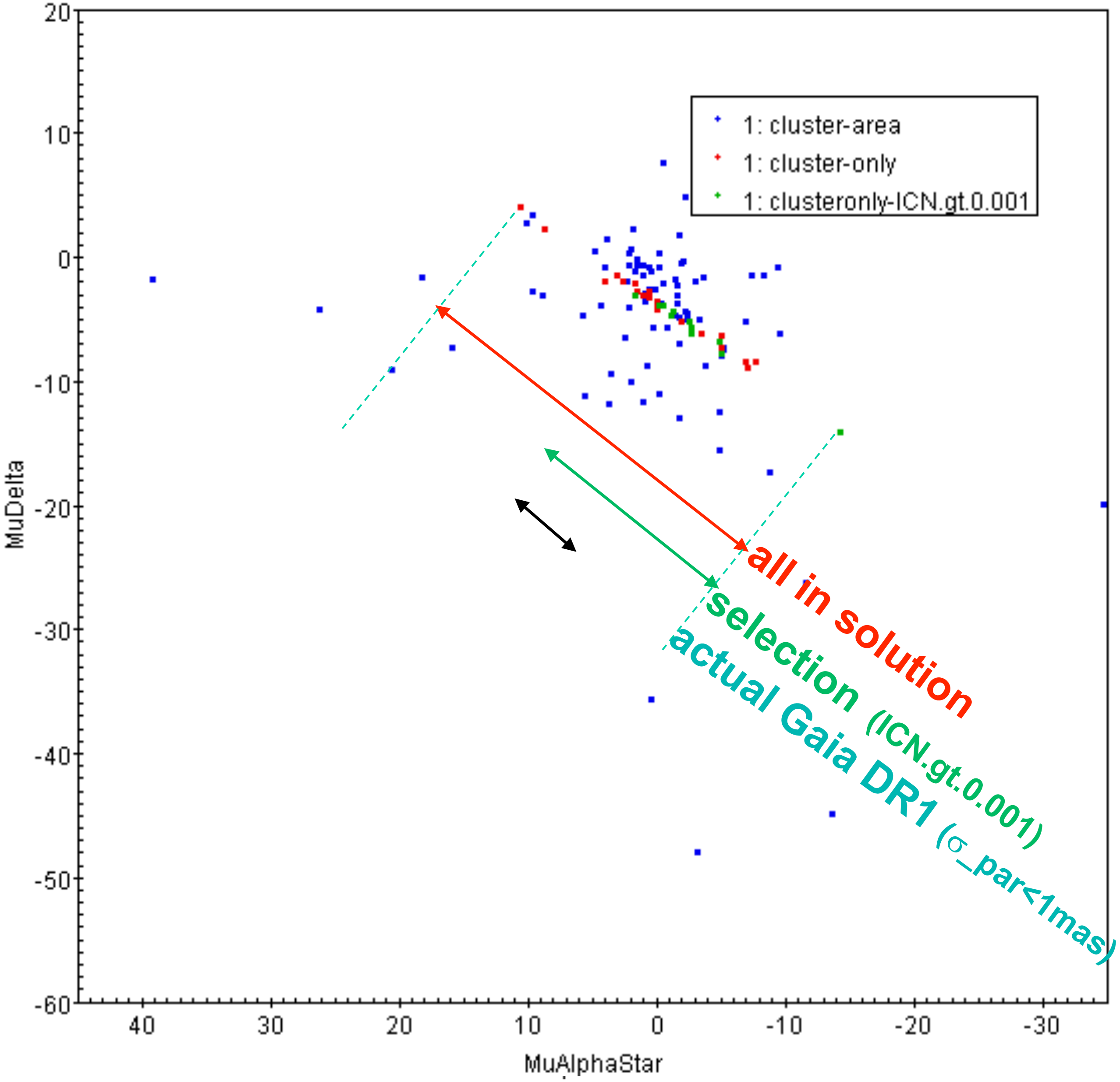
M11; lessons to be learned



Variances/mean errors
Covariances/Correlations

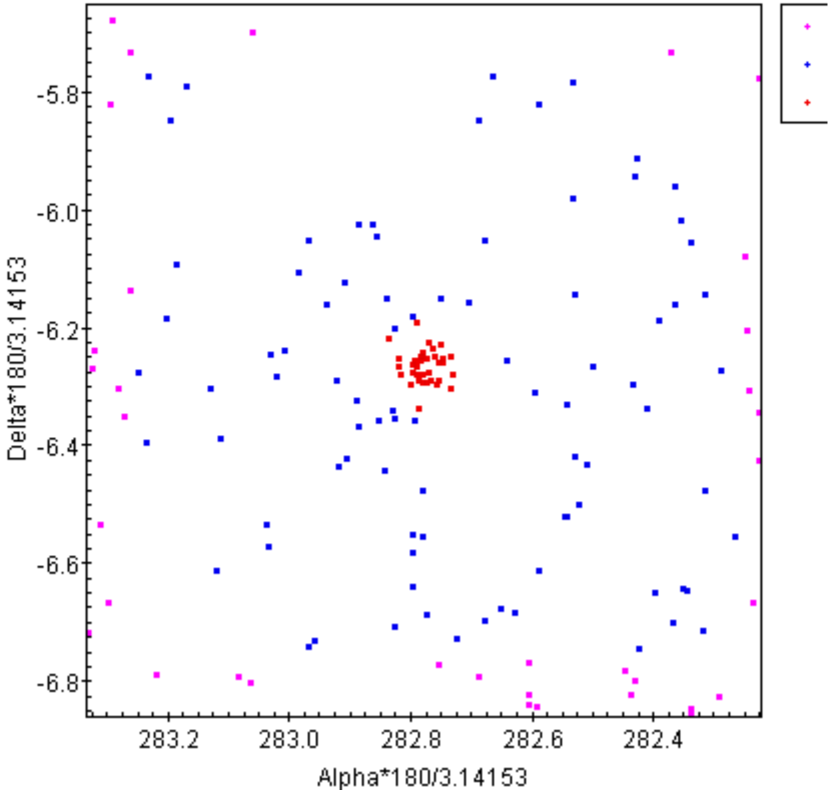
GoF (F2)
Source excess noise

M11; reasonable selection improves things

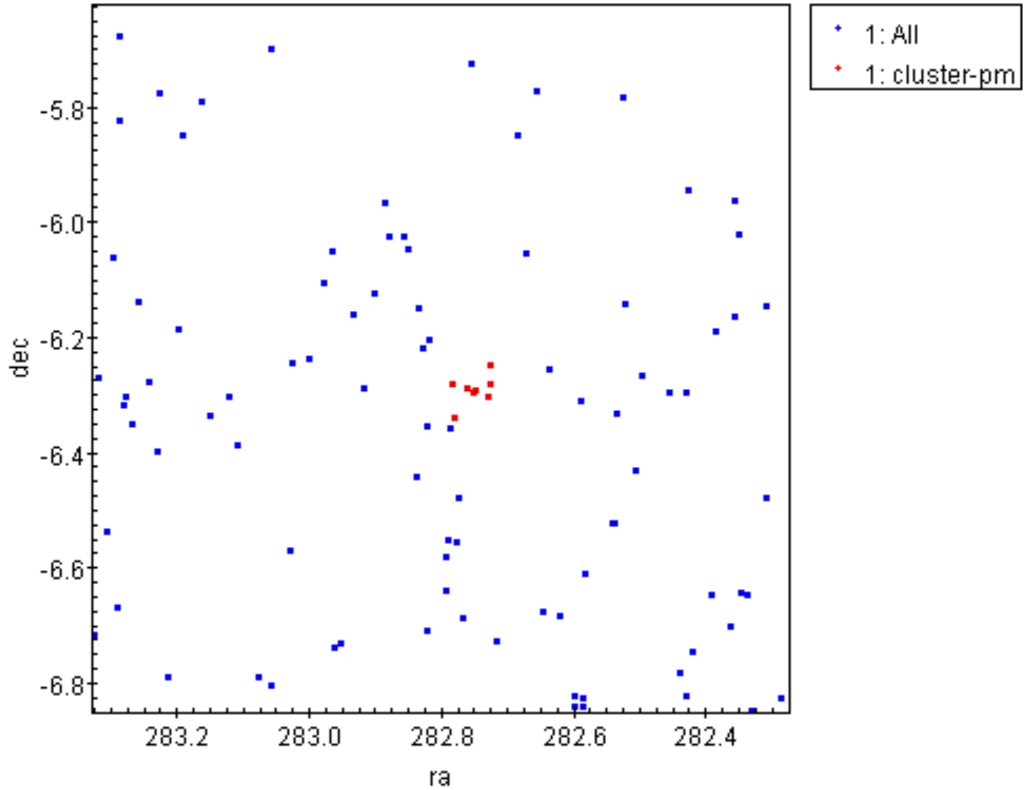


But there's always a price to be payed:

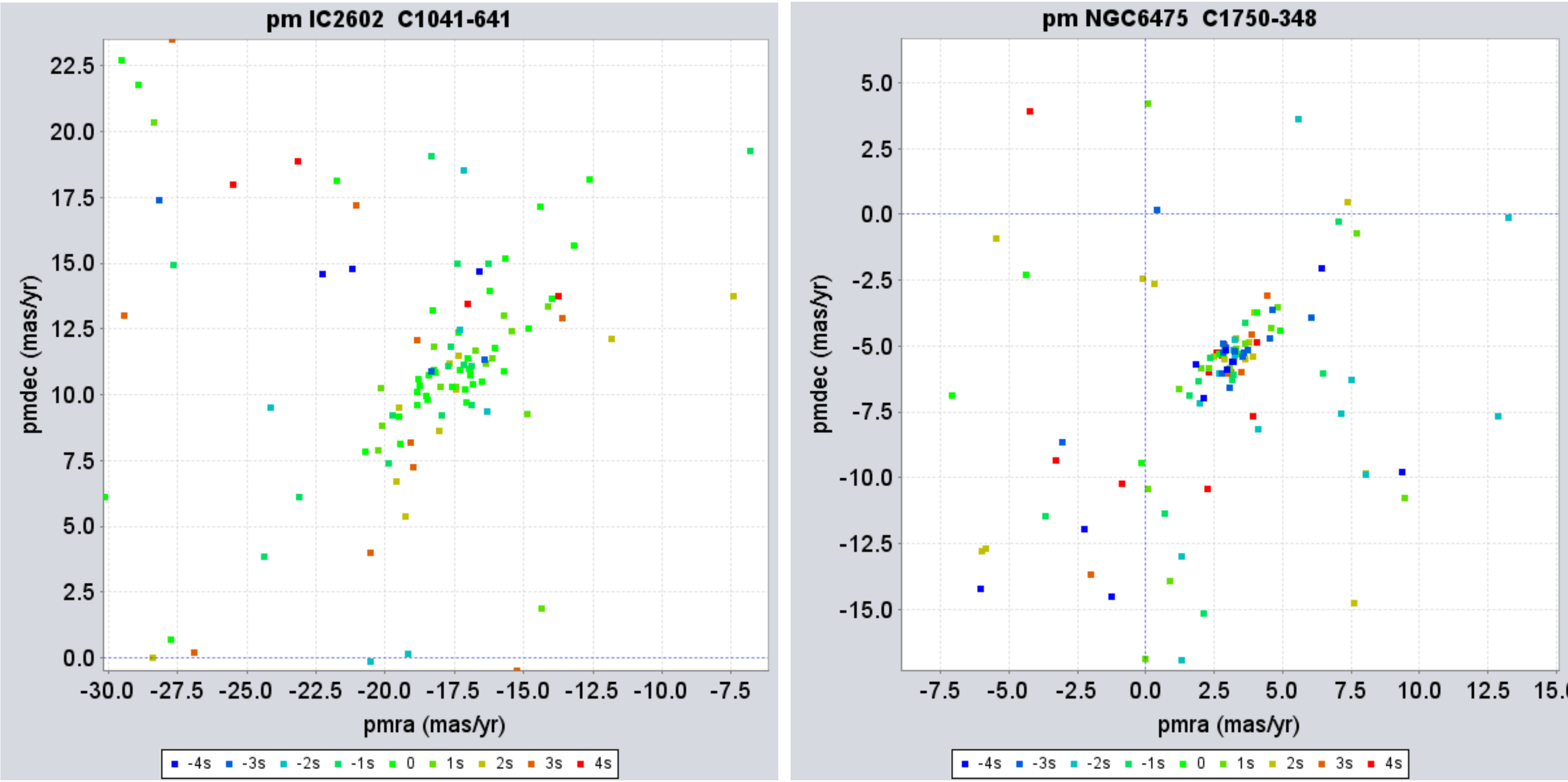
all in TGAS solution



actually in Gaia DR1



M11 is an extreme case, but ...



Two less extreme but still clearcut cases; using public DR1 data.

Note: the scales of the two figures are equal. NGC 6475 measured much more precisely.

Chapter 4: Transformations

Transformations:

when the quantity you want to study
is not the quantity you observe

- Usually you want distances, not parallaxes
- Usually you want spatial velocities, not proper motions

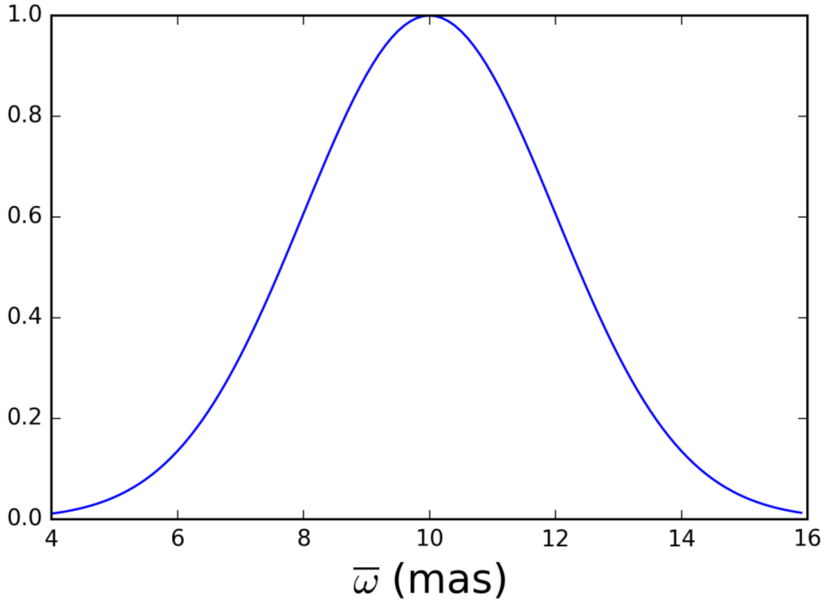
Warning:

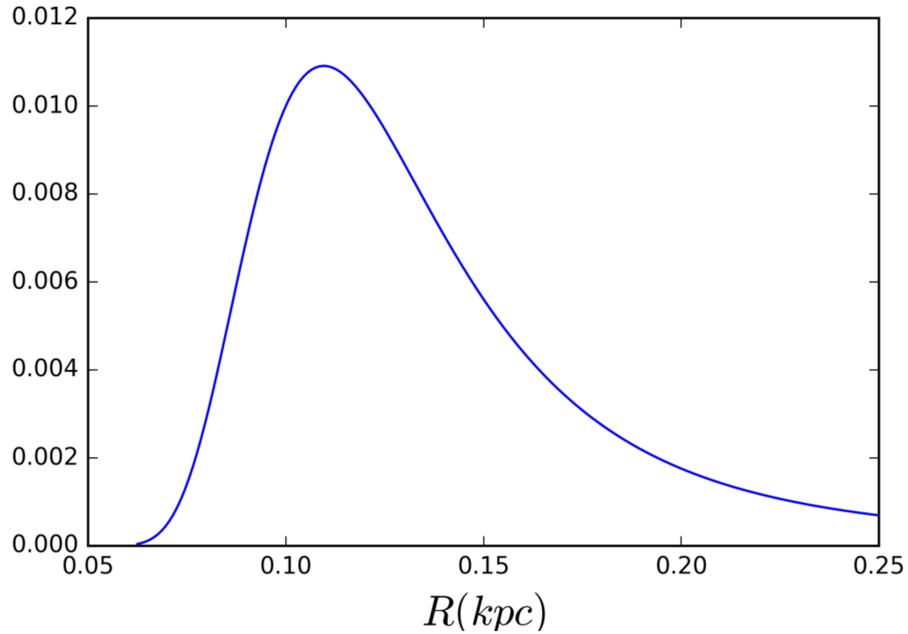
when using a transformed quantity the error distribution also is transformed

- This is especially crucial for the calculation of distances from parallaxes
- And even more so for the calculation of luminosities from parallaxes
- A symmetrical, well behaved error in parallax is transformed into an asymmetrical error in distance

Error distribution comparison: star at 100pc and parallax error 2mas

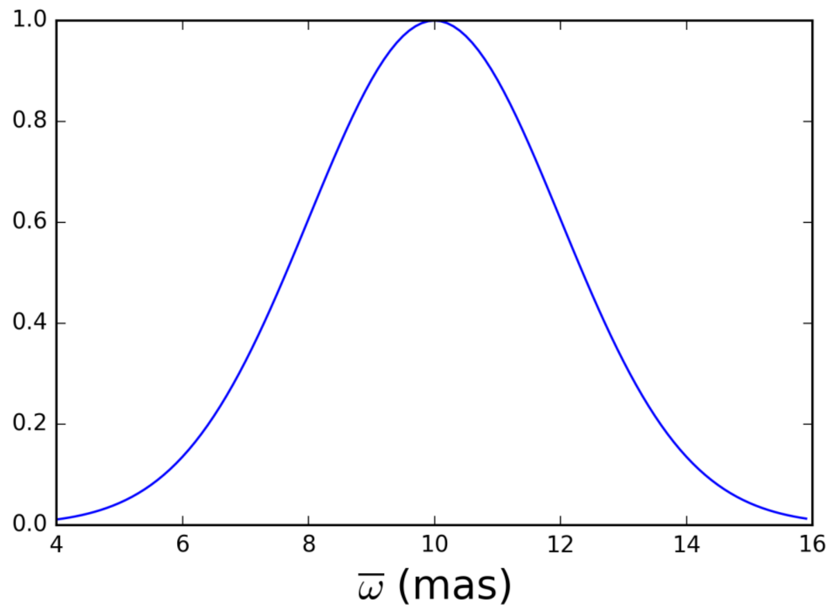
parallax and distance (schematic; non normalised)

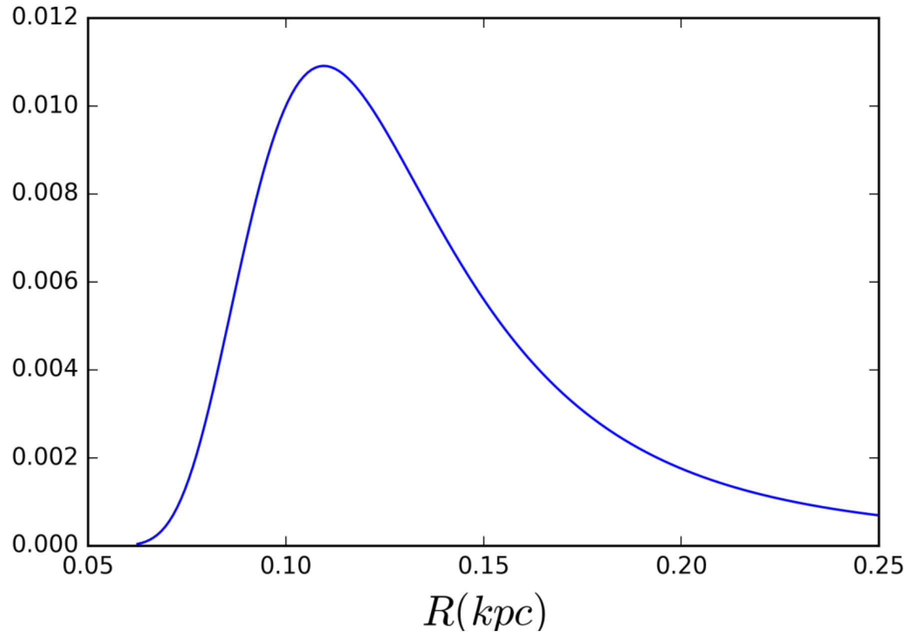




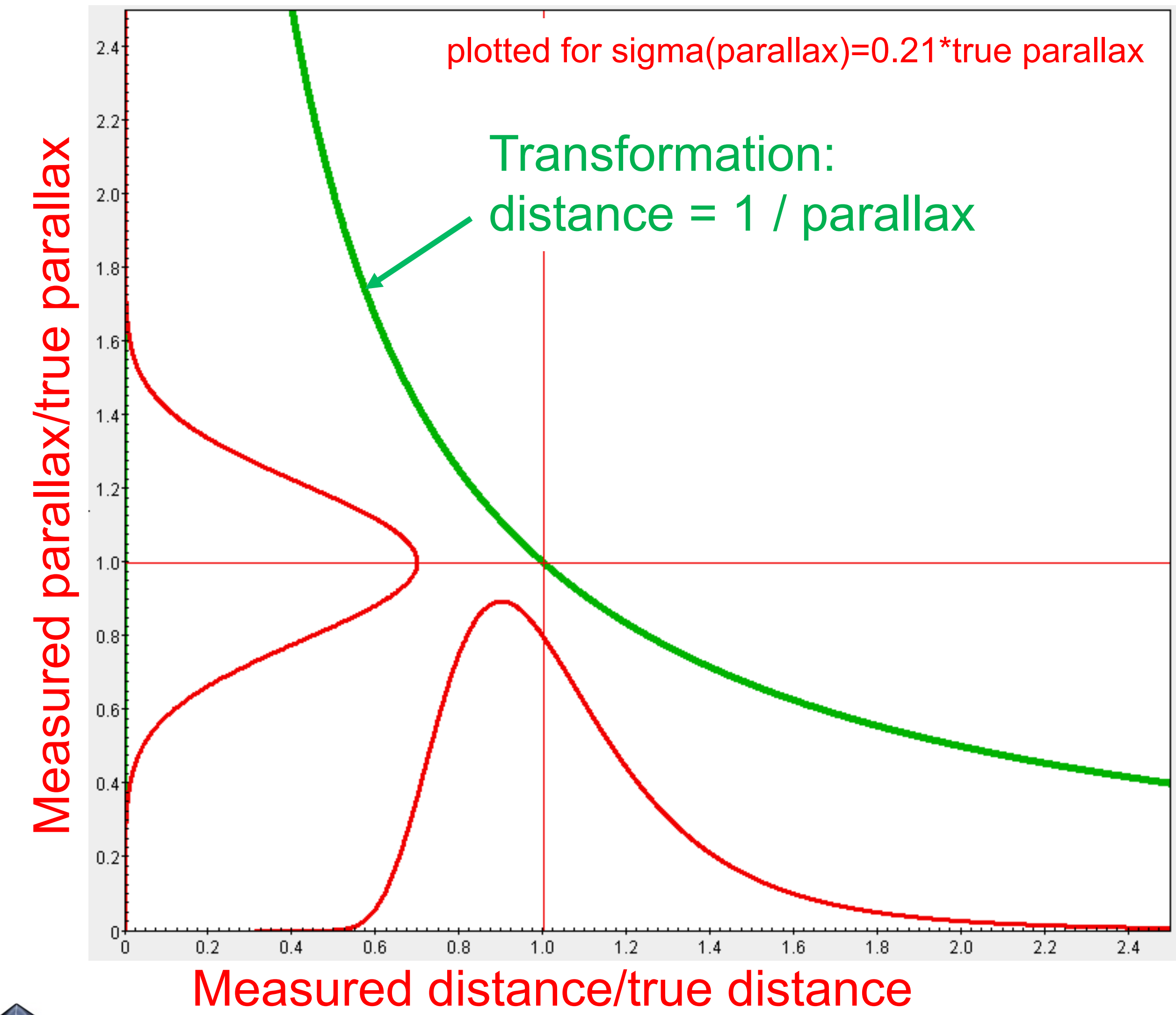
Error distribution comparison:
star at 100pc and parallax error 2mas

parallax and distance (non normalised)

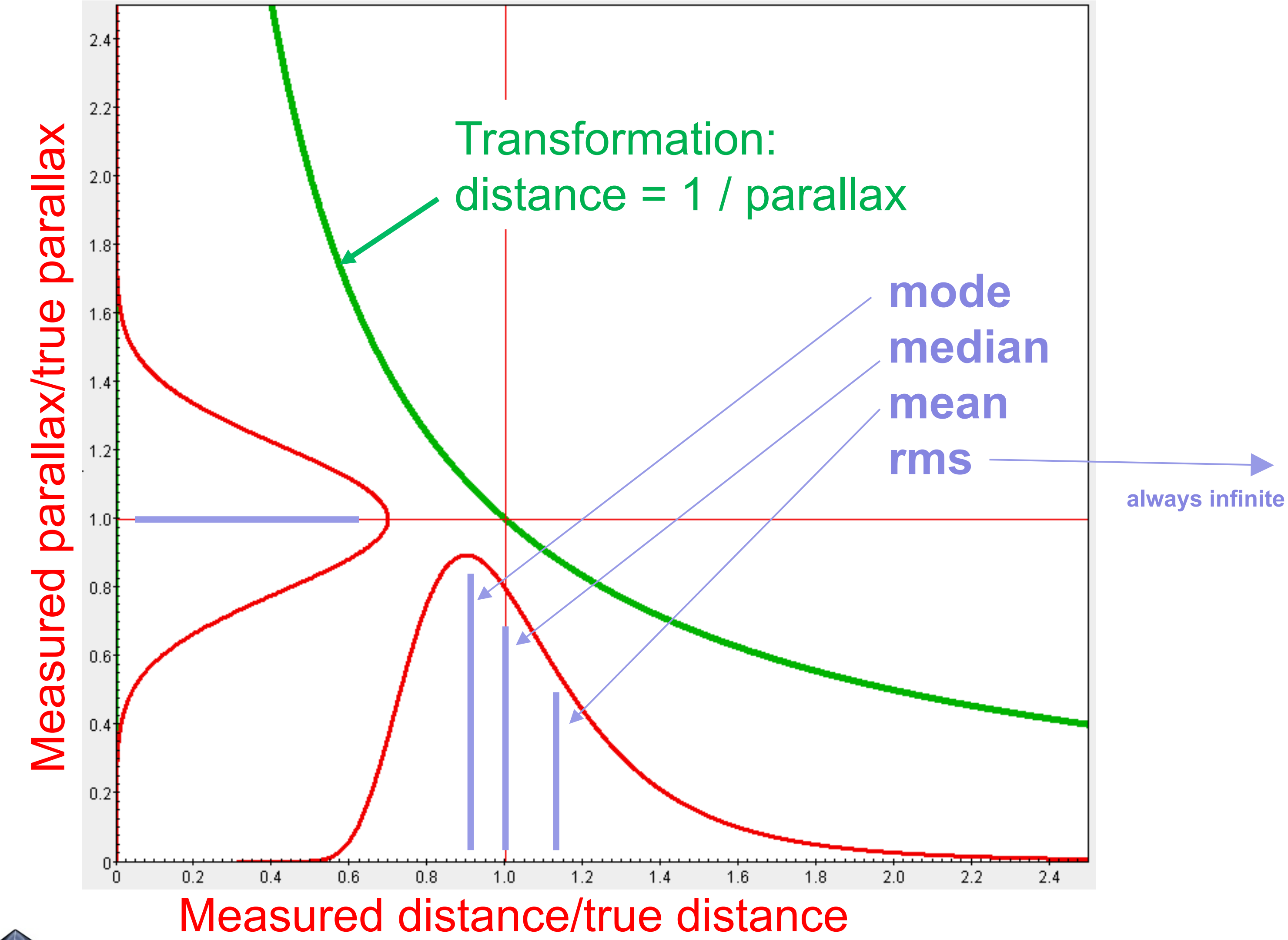




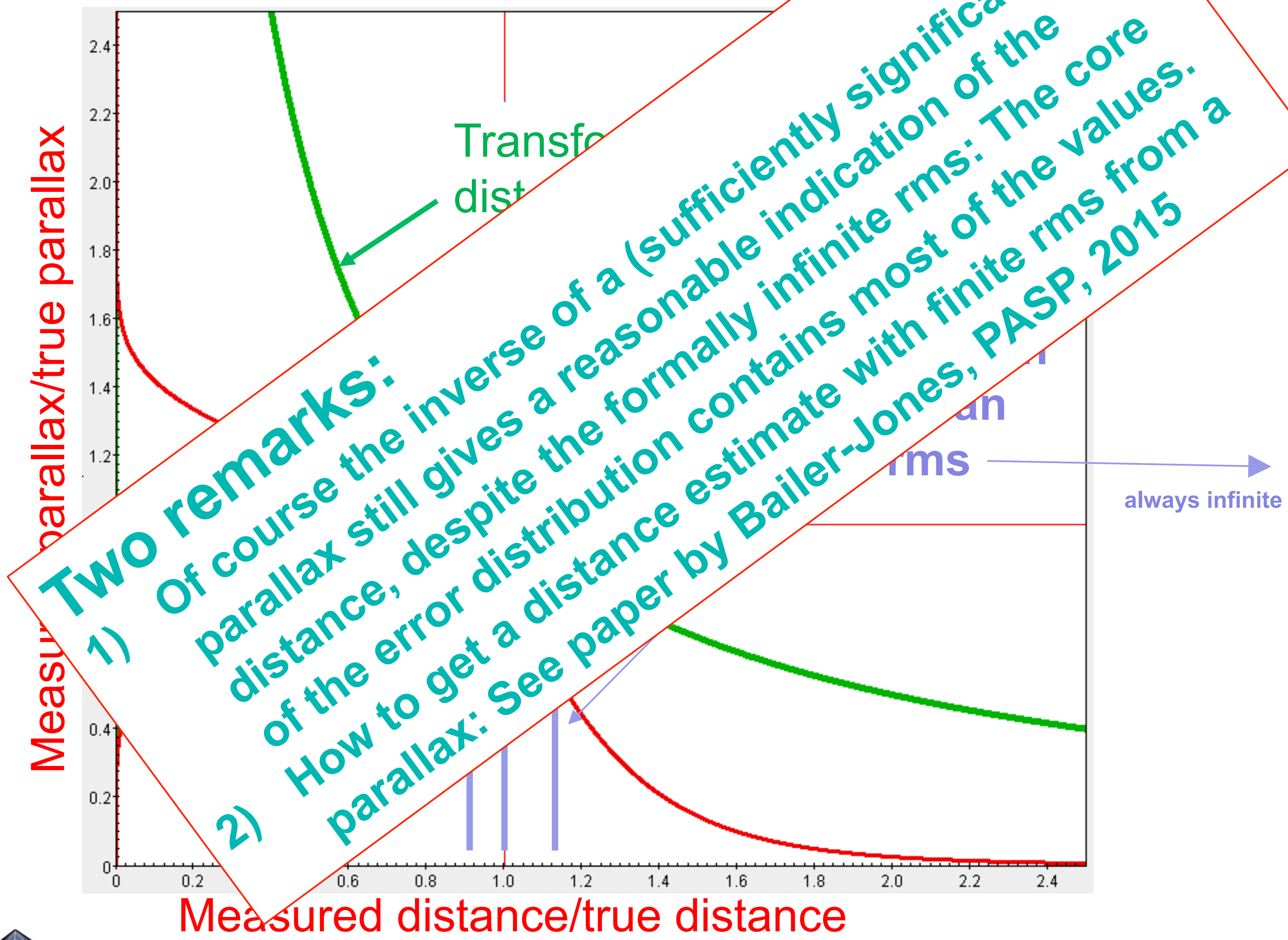
Error distribution comparison:
parallax versus distance



Error distribution comparison:
parallax versus distance



Error distribution comparison:
parallax versus distance



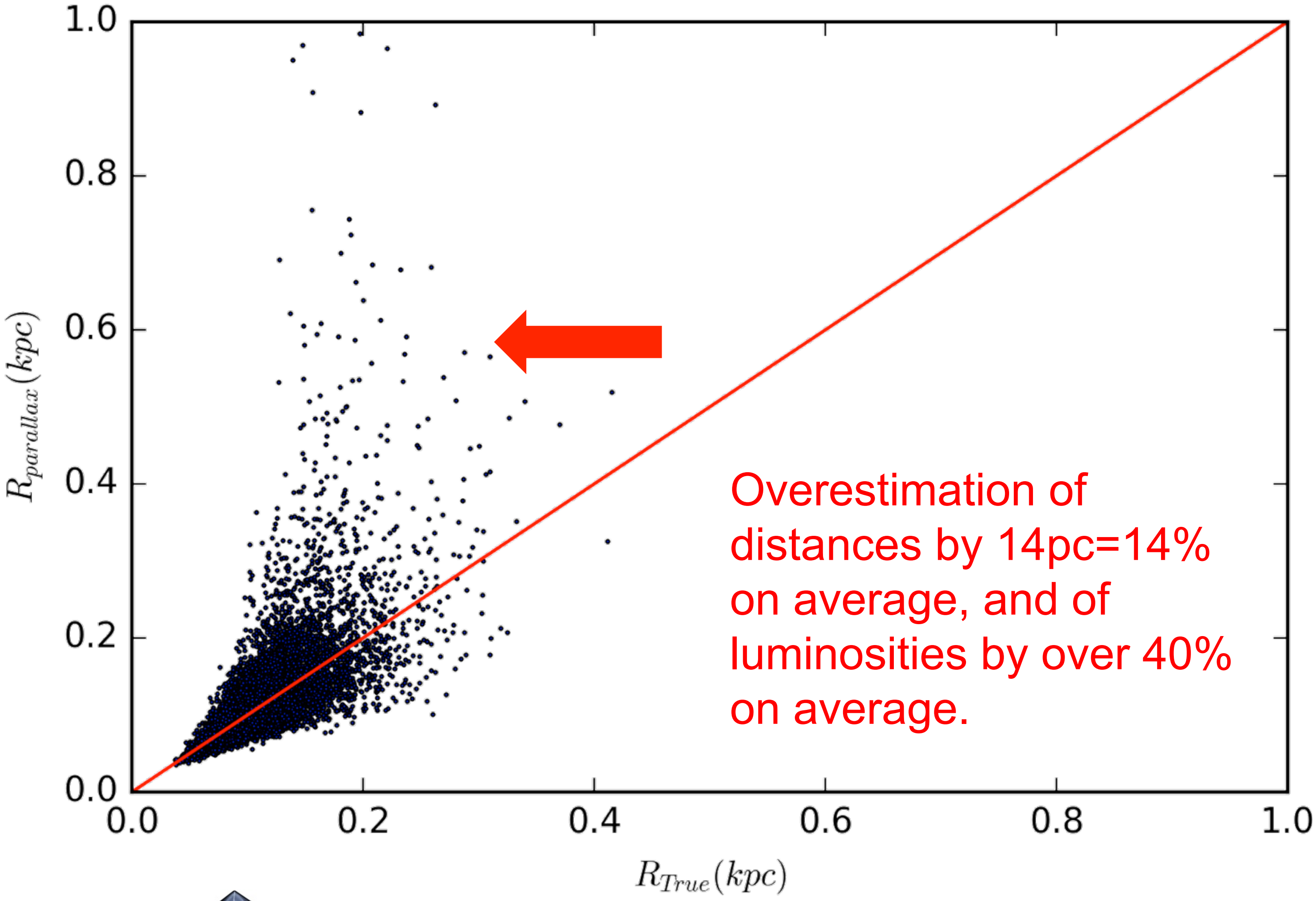
Two remarks:

1) Of course the inverse of a (sufficiently significant) parallax still gives a reasonable indication of the distance, despite the formally infinite rms: The core of the error distribution contains most of the values.

2) How to get a distance estimate with finite rms from a parallax: See paper by Bailer-Jones, PASP, 2015

Sample simulation with a parallax error of 2mas

True distance vs. distance from parallax



How to take this into account

- Avoid using transformations as much as possible
- If unavoidable:
  - Do fits in the plane of parallaxes (e.g. PL relations using ABL method*) where errors are well behaved
  - Do any averaging in parallaxes and then do the transformation (e.g. distance to an open cluster)
  - Always estimate the remaining effect (analytically or with simulations)

*Astrometry-Based Luminosity (ABL) method

$$a_V = 10^{0.2M_V} = \pi 10^{\frac{m_V+5}{5}}$$

This quantity is:
- related to luminosity
  (sqrt of inverse luminosity)
- a linear function of parallax
- thus nicely behaved
- thus can be averaged safely

Also beware of additional assumptions

- For instance about the absorption when calculating absolute magnitudes from parallaxes

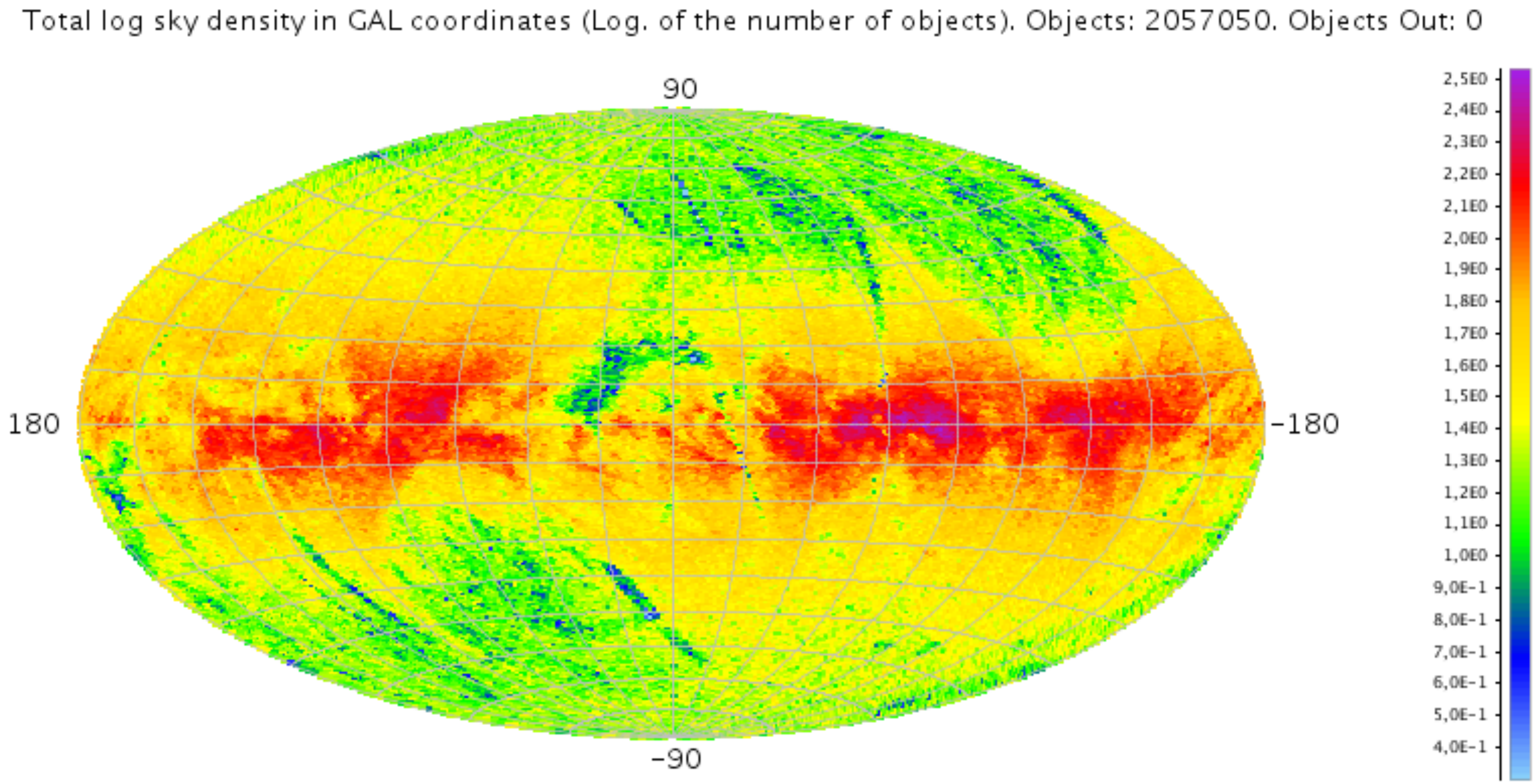
Chapter 5: Sample censorships

Completeness/representativeness:

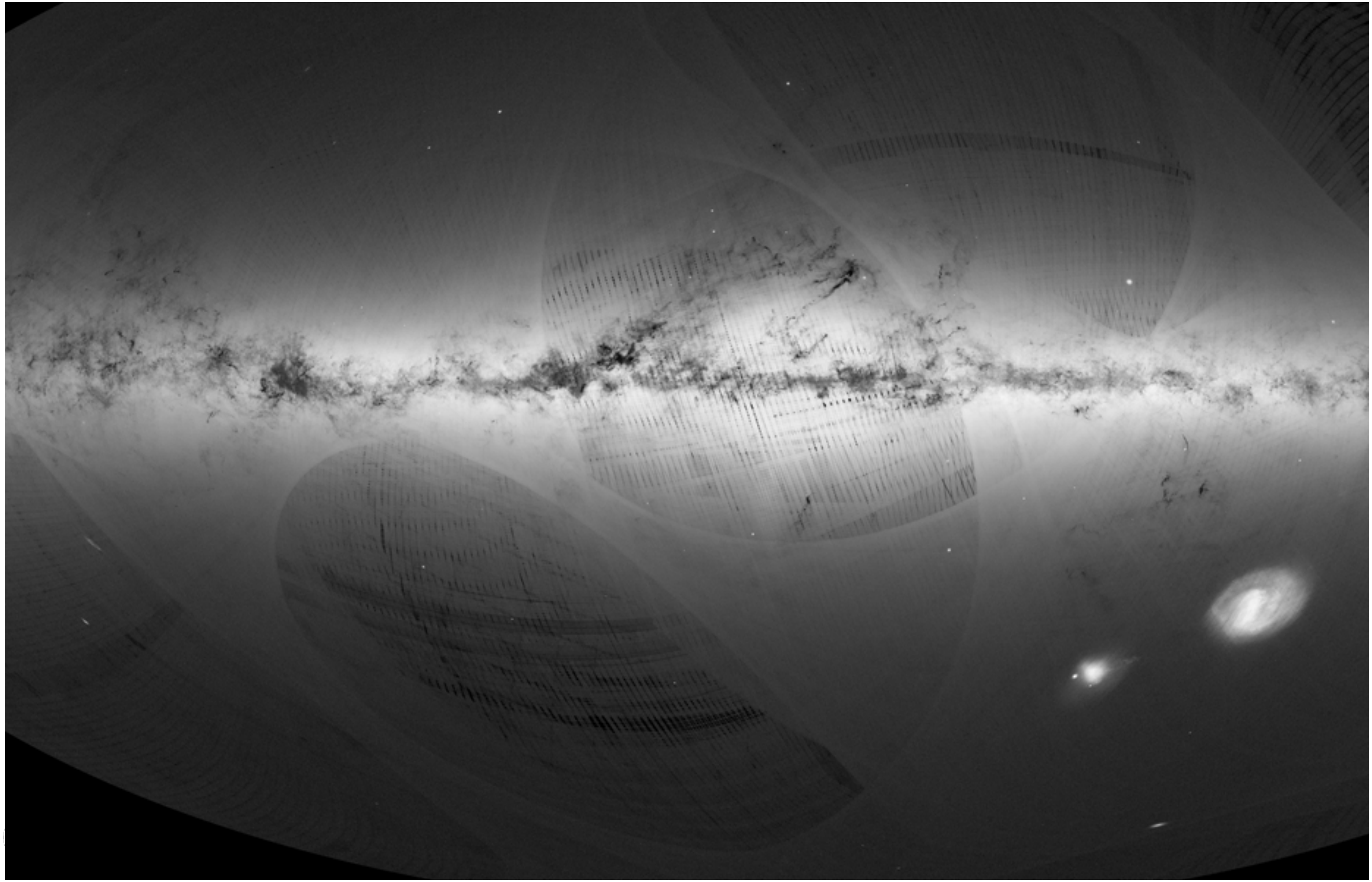
we have the complete population of objects or at least a subsample which is representative for a given purpose

- DR1 is a very complex dataset, its completeness or representativeness can not be guaranteed for any specific purpose

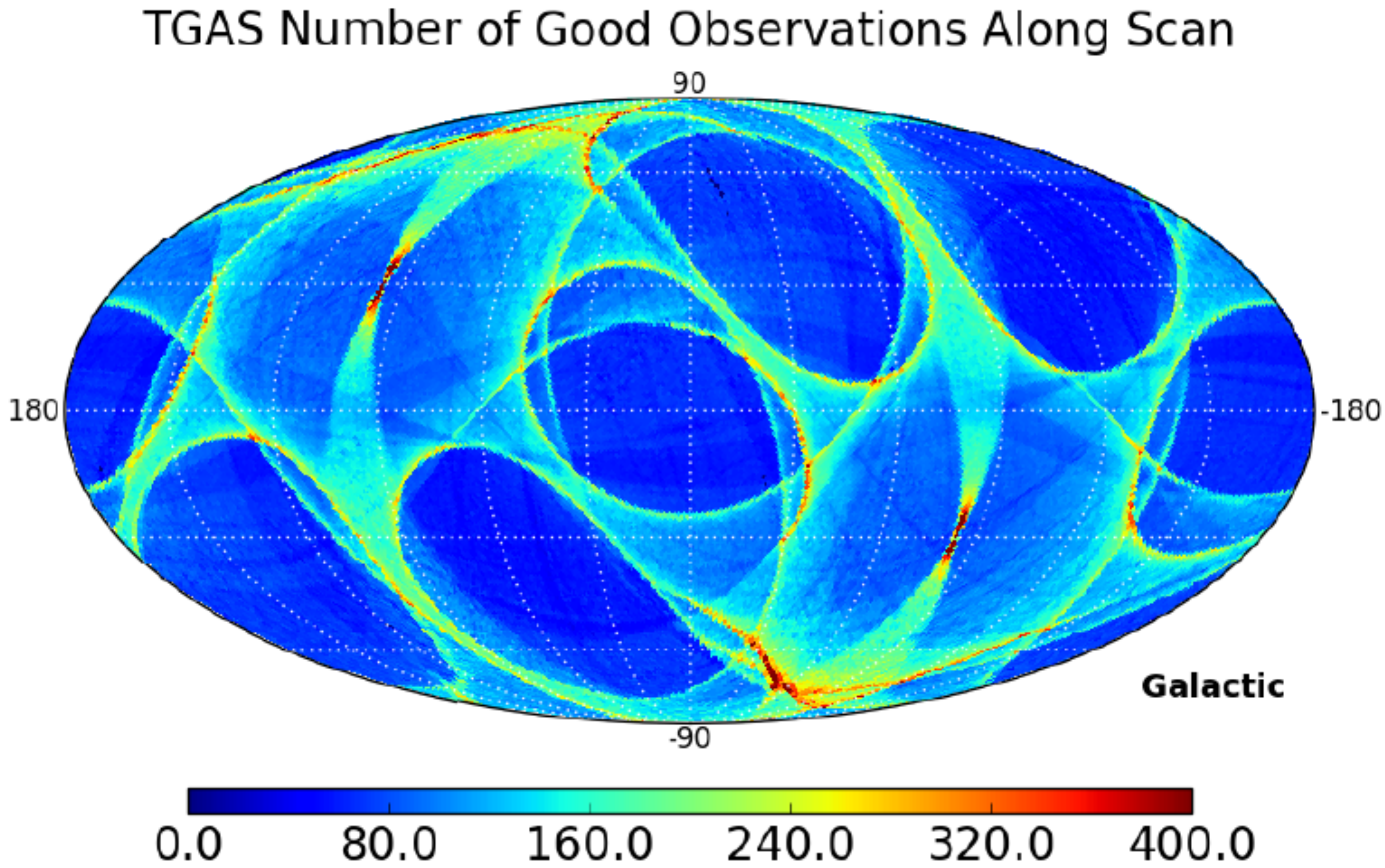
Significant completeness variations as a function of the sky position



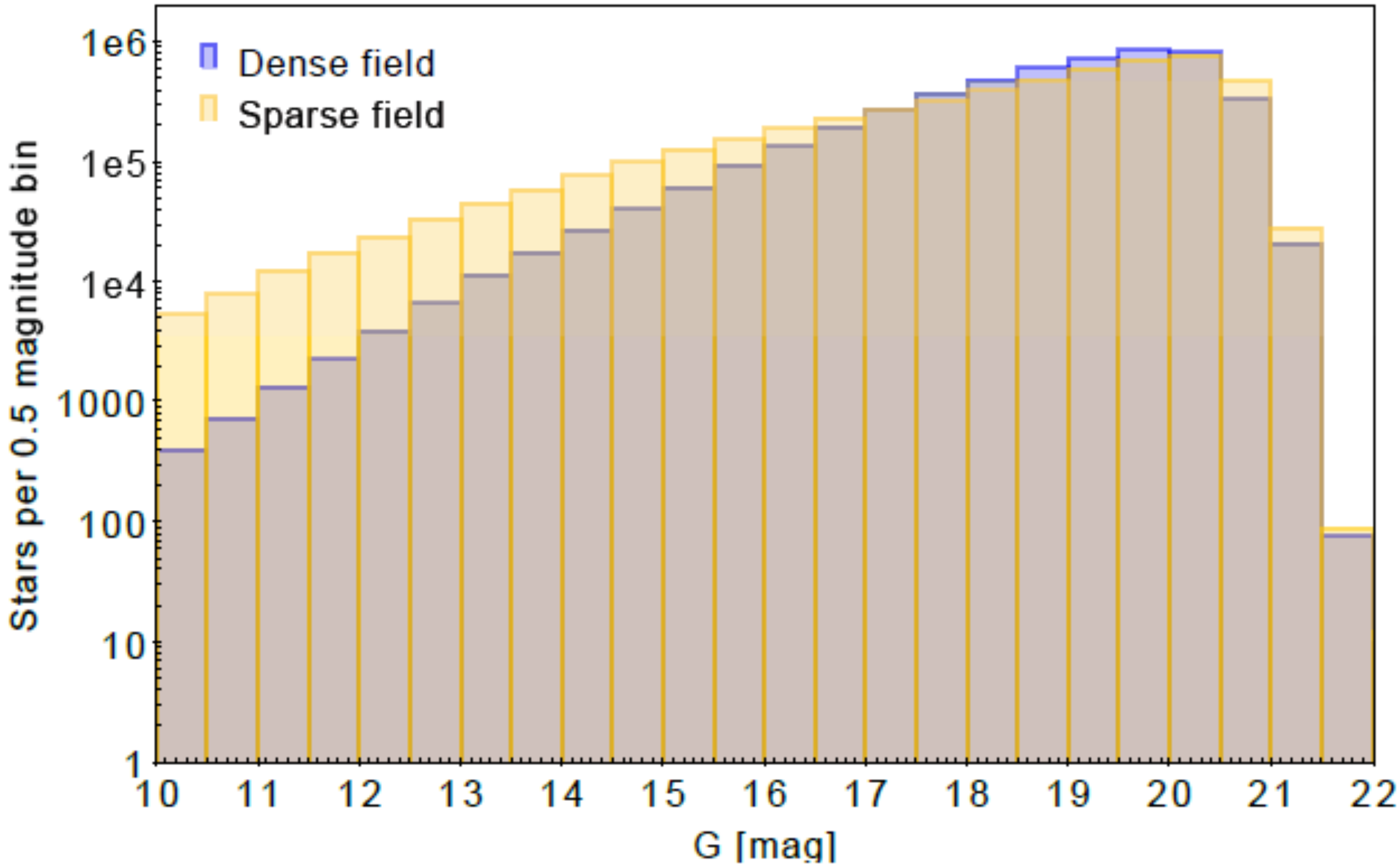
Significant completeness variations as a function of the sky position



Complex selection of astrometry (e.g. Nobs)



Not complete in magnitude or color



How to take this into account

- **Very difficult,** will depend on your specific purpose

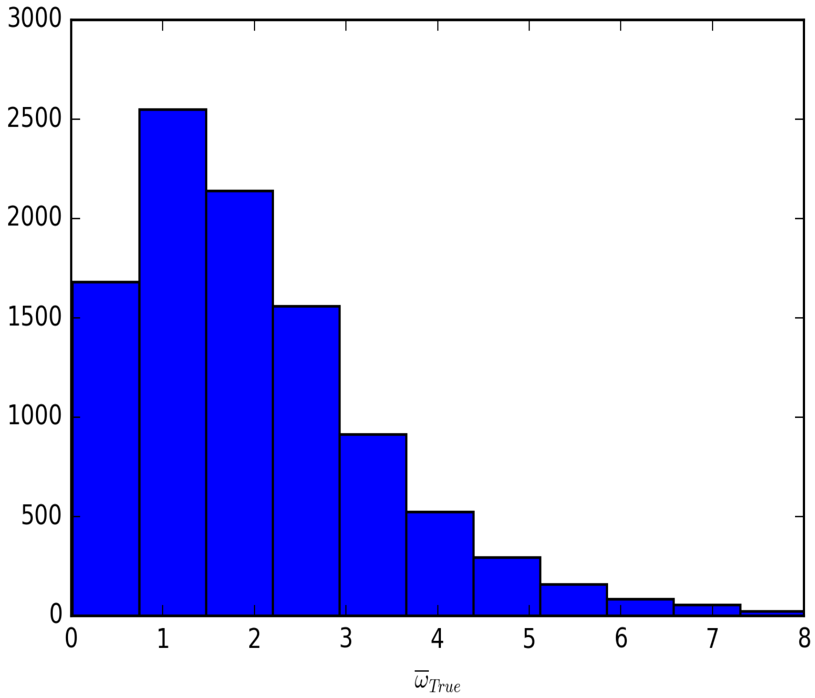
- Analyze if the problem exists, and try to determine if the known censorships are correlated with the parameter you are analyzing (see validation paper)

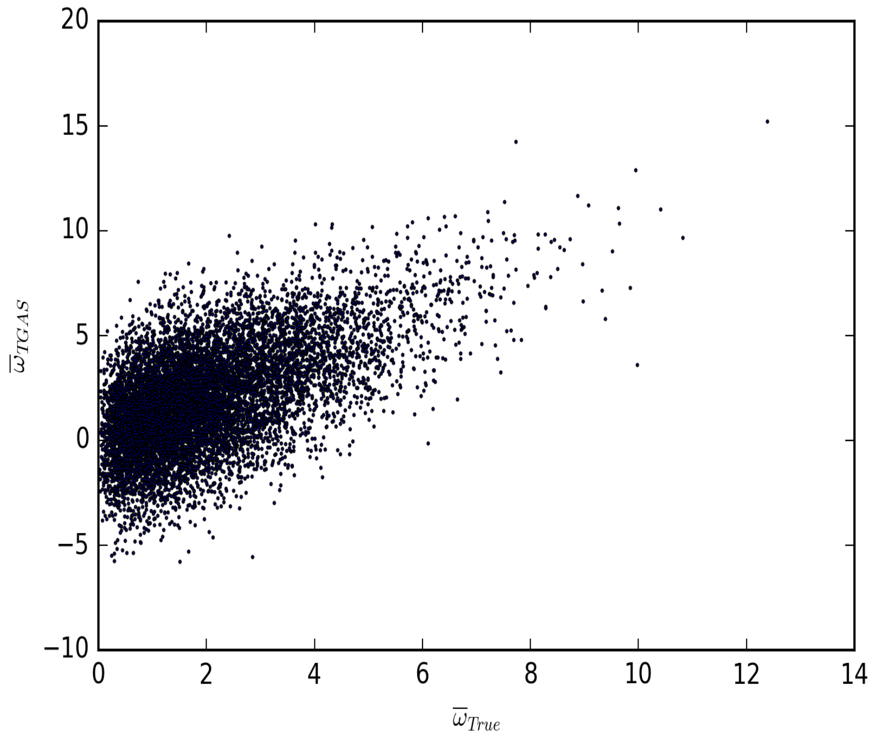
- At least do some simulations to evaluate the possible effects

IMPORTANT: do not make things worse by adding your own additional censorships

- **This is specially important for parallaxes**
- Avoid removing negative parallaxes; this removes information and biases the sample for distant stars
- Avoid selecting subsamples on parallax relative error. This also removes information and biases the sample for distant stars
- **Use instead fitting methods able to use all available data (e.g. Bayesian methods) and always work on the observable space (e.g. on parallaxes, not on distances or luminosities)**

Example: Original (complete) dataset
(errors in parallax of 2mas)



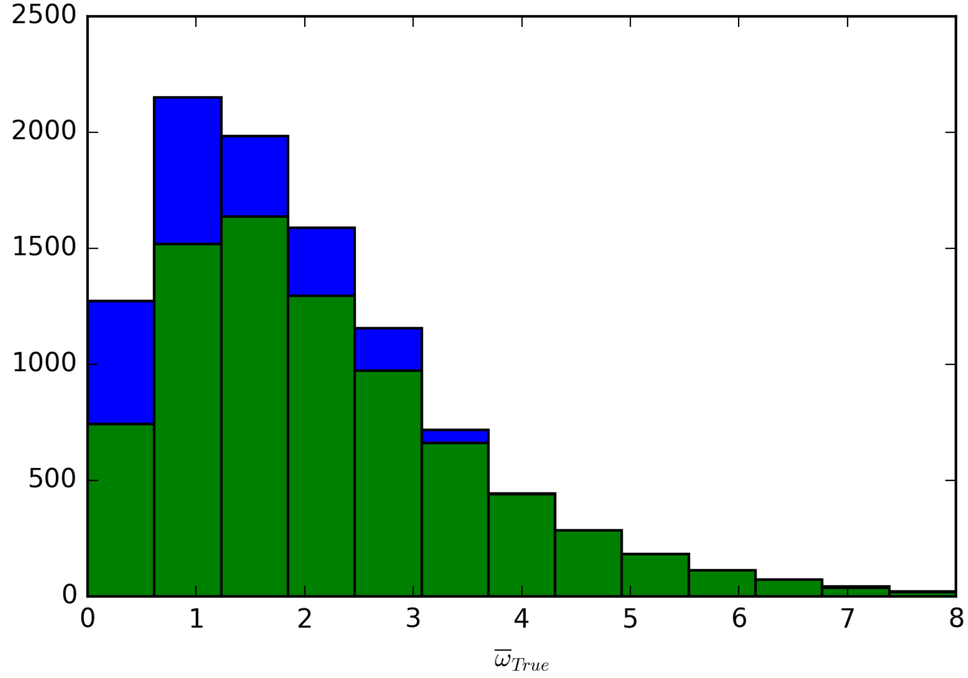


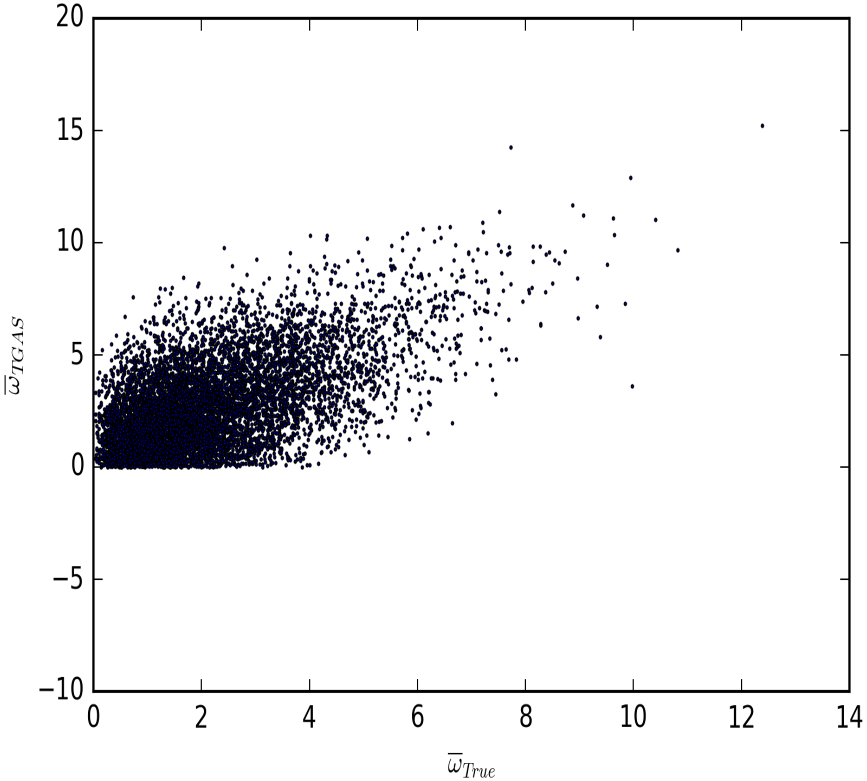
Simulation !

Average diff. of parallaxes = 0.002 mas

Example: removing negative parallaxes

Favours large parallaxes



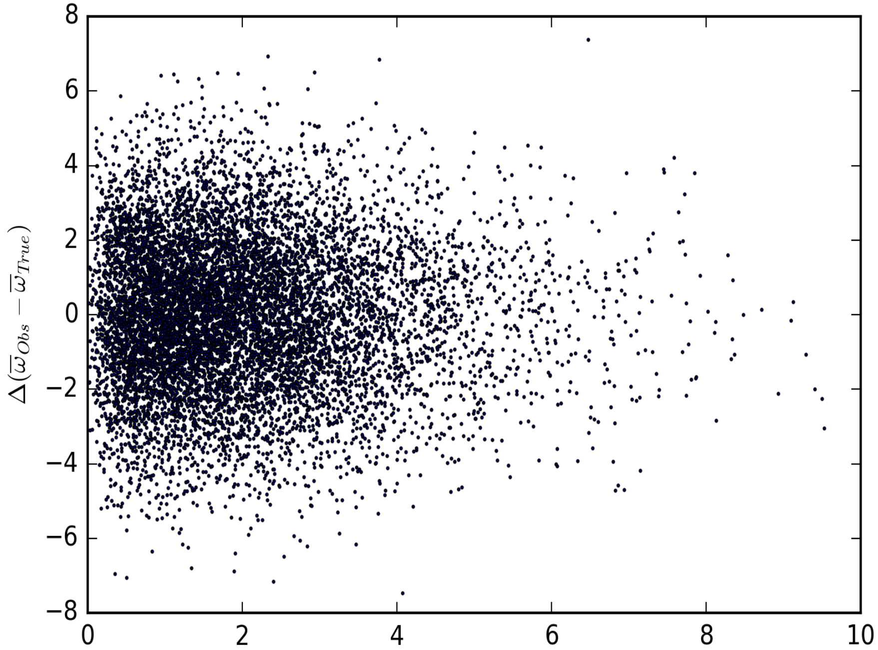


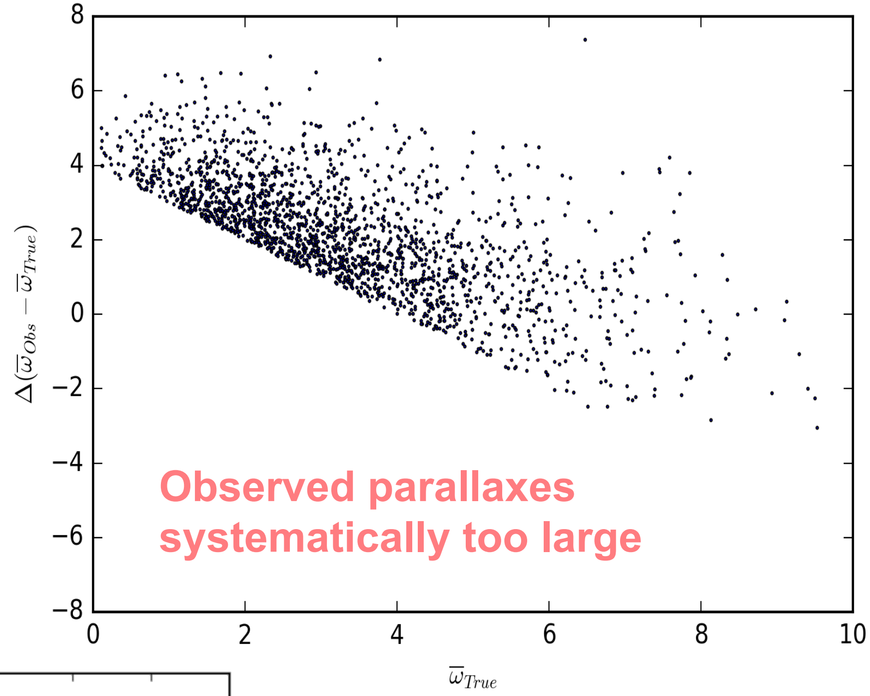
Simulation !

Average diff. of parallaxes = 0.65 mas

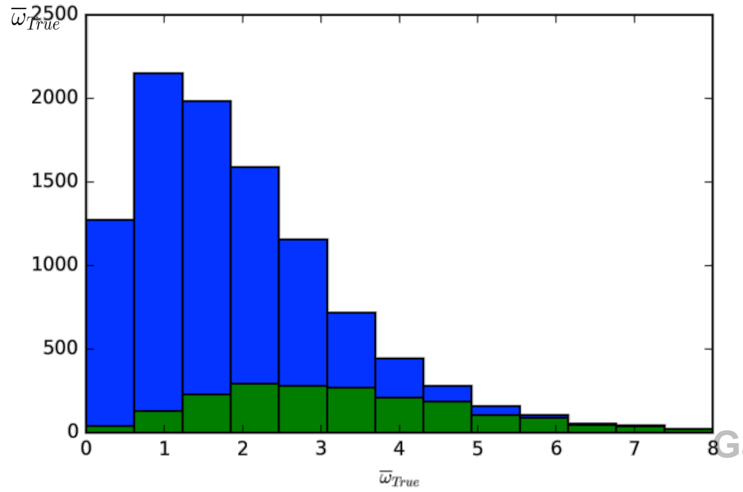
Example: removing sigmaPar/Par > 50%

Favours errors making parallax larger





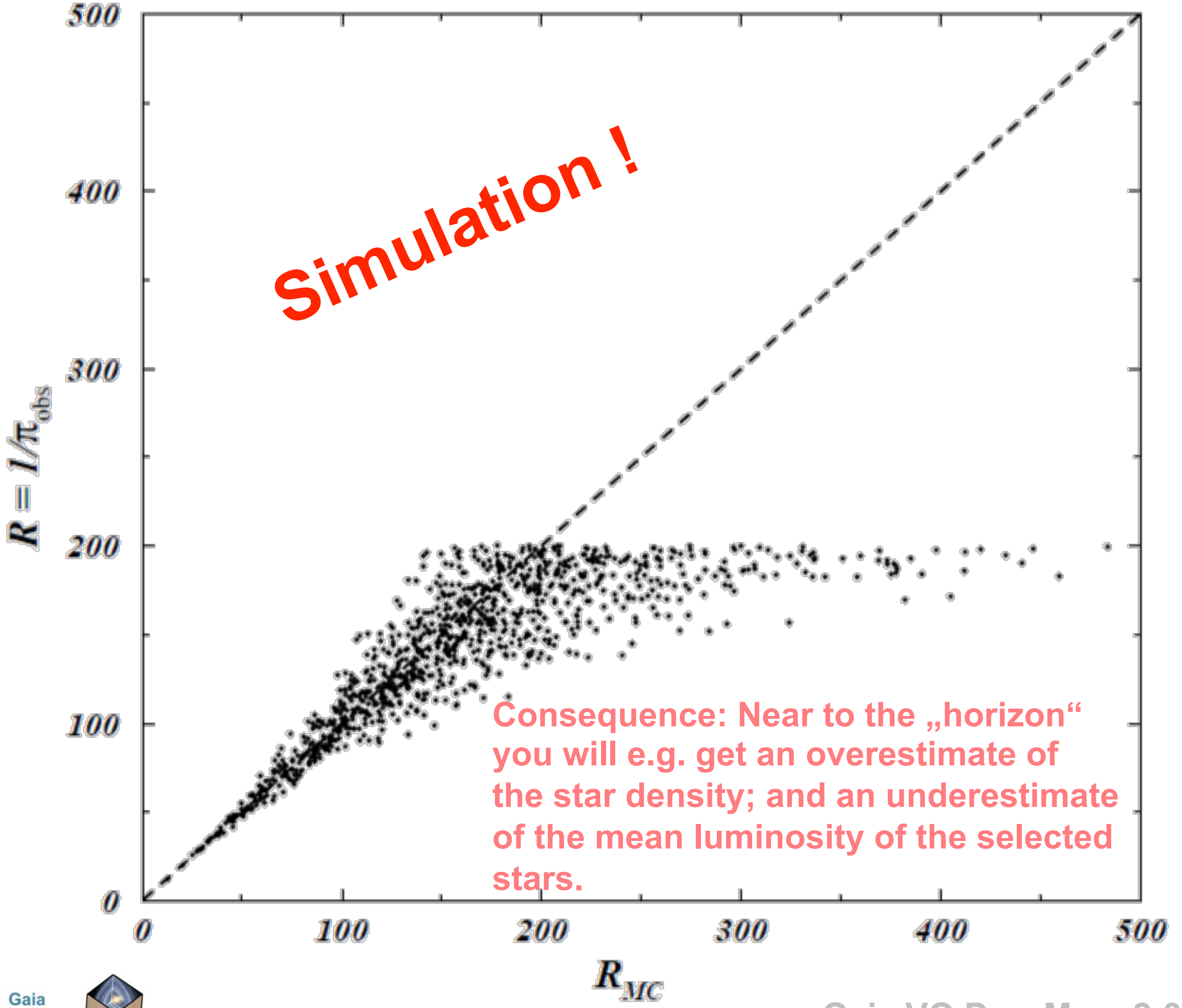
Simulation !



Average diff. of parallaxes = 2.2 mas

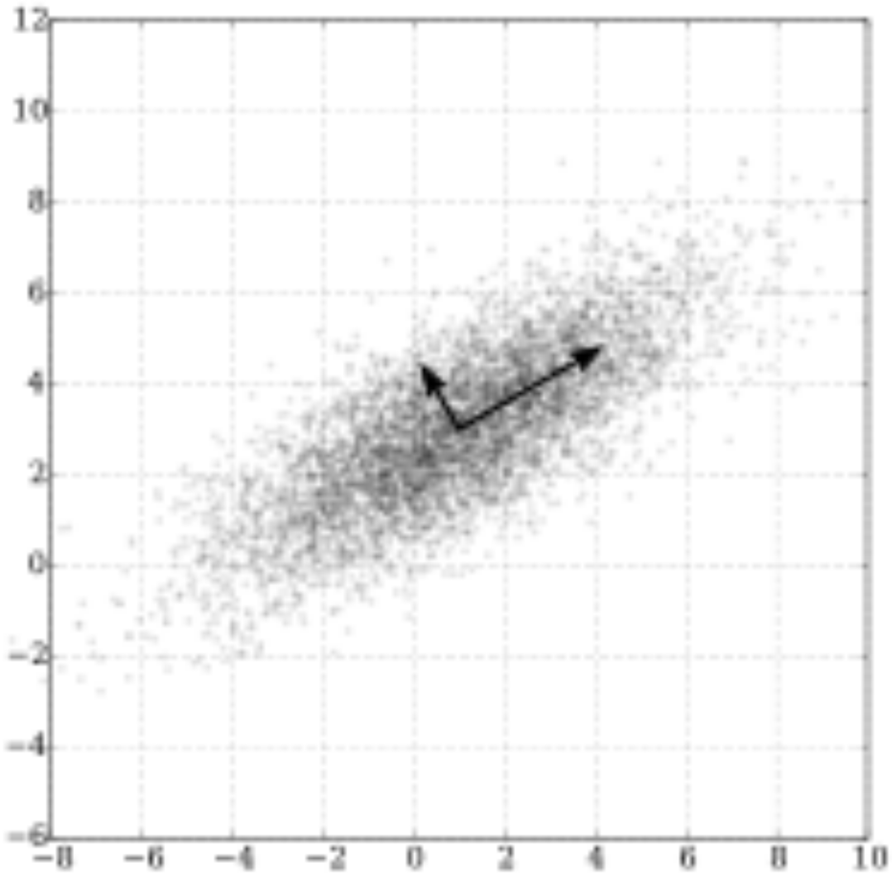
Example: truncation by observed parallax

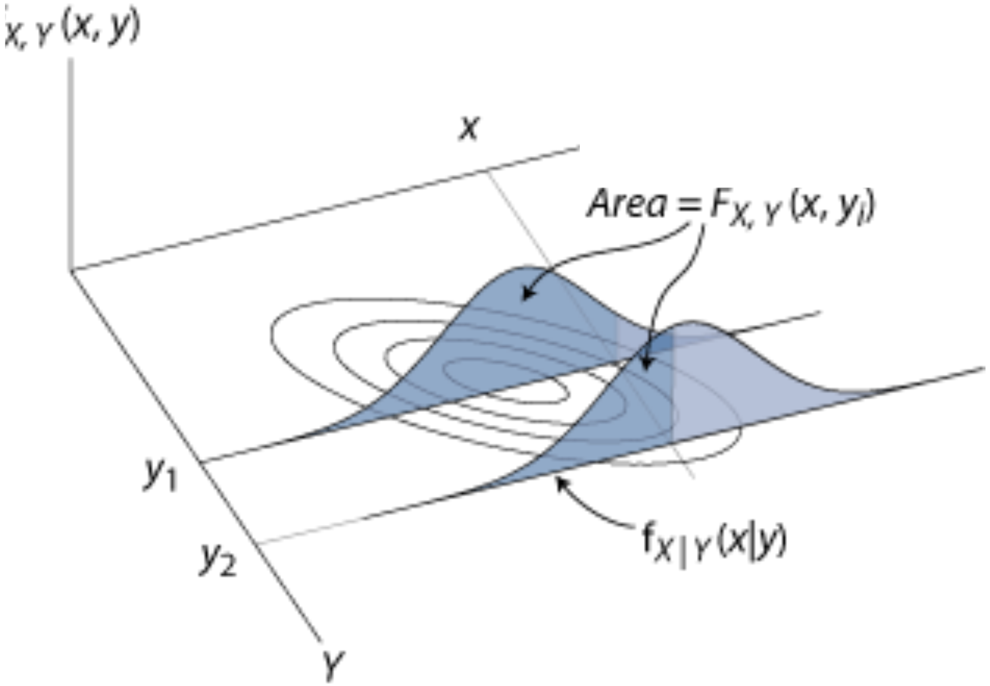
Favours objects at large distances (small true parallax)



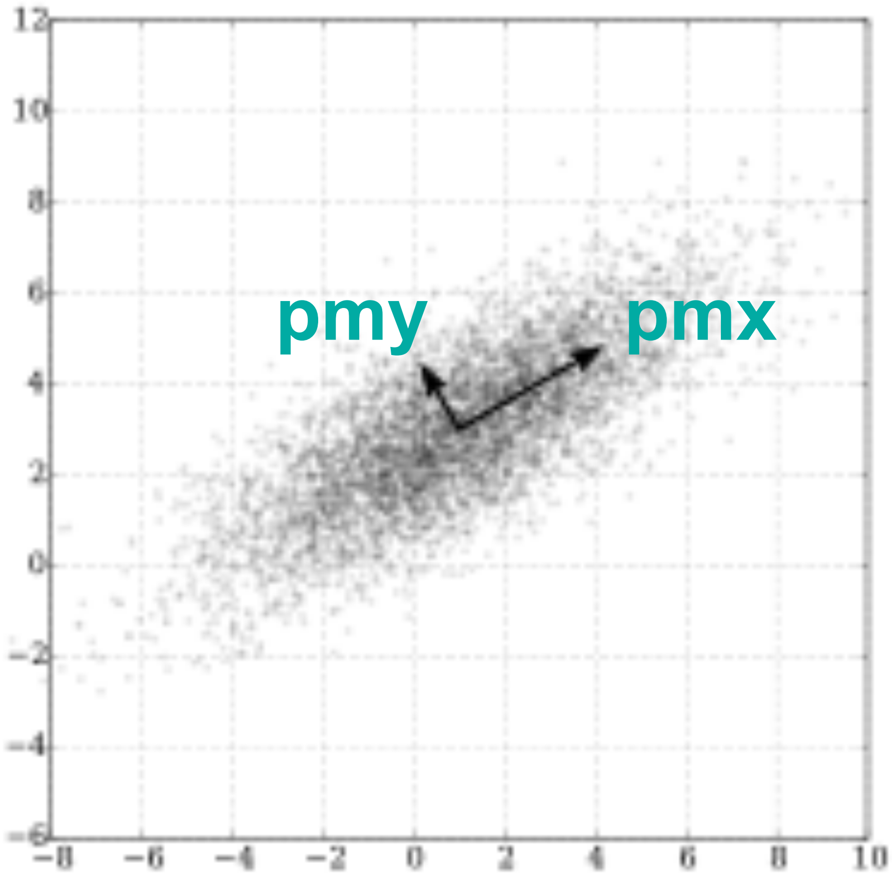
Appendix

Uncorrelated quantities from correlated catalogue values





Uncorrelated quantities from correlated catalogue values



Given:
pma, pmd,
sigma(pma),sigma(pmd), corr(pma,pmd)

Wanted: orientation and principal axes of the error ellipse

Go to rotated coordinate system x,y. The two proper-motion components pmx and pmy are uncorrelated:

pmx= pmd*cos(theta) + pma* sin(theta)
pmy= -pmd*sin(theta) + pma*cos(theta)

Question:

Which theta?
And which sigma(pmx), sigma(pmy) ?

Uncorrelated quantities from correlated catalogue values

Keyword: Eigenvalue decomposition

(of the relevant covariance matrix part)

Example for the "looks" of a covariance matrix (2 × 2, proper motions only):

$$\begin{pmatrix} \sigma^2_{\mu_{\alpha*}} & \mathrm{cov}_{\mu_{\alpha*},\mu_\delta} \\ \mathrm{cov}_{\mu_{\alpha*},\mu_\delta} & \sigma^2_{\mu_\delta} \end{pmatrix}$$

Note: $\mathrm{cov}_{\mu_{\alpha*},\mu_\delta} = \mathrm{corr}_{\mu_{\alpha*},\mu_\delta}\,\sigma_{\mu_{\alpha*}}\,\sigma_{\mu_\delta}$

Solution of the Eigenvalue decomposition for 2 dimensions:
The maxima and minima of the variance (the eigenvalues of the matrix) are:

$$\begin{aligned} \sigma^2_{\mu_x} &= \frac{1}{2}\left(\sigma^2_{\mu_{\alpha*}} + \sigma^2_{\mu_\delta} + \sqrt{\left[\sigma^2_{\mu_{\alpha*}} + \sigma^2_{\mu_\delta}\right]^2 - 4\mathrm{cov}^2_{\mu_{\alpha*},\mu_\delta}}\right) \\ \sigma^2_{\mu_y} &= \frac{1}{2}\left(\sigma^2_{\mu_{\alpha*}} + \sigma^2_{\mu_\delta} - \sqrt{\left[\sigma^2_{\mu_{\alpha*}} + \sigma^2_{\mu_\delta}\right]^2 - 4\mathrm{cov}^2_{\mu_{\alpha*},\mu_\delta}}\right) \\ \tan(\theta) &= \frac{\sigma^2_{\mu_{\alpha*}} - \sigma^2_{\mu_\delta}}{\mathrm{cov}_{\mu_{\alpha*},\mu_\delta}} \end{aligned}$$

Note 1: the $\pm 180°$ ambiguity of the tangent does not matter in this case

Note 2: for $\mathrm{cov}_{\mu_{\alpha*},\mu_\delta} = 0$, then $\theta = 0$ if $\sigma_{\mu_\delta} > \sigma_{\mu_{\alpha*}}$, else $\theta = 90°$ and the values are trivial

Even more tedious formulae for 3 dimensions; better use matrix routines for 3d and higher dimensions.

Uncorrelated quantities from correlated catalogue values

Keyword: Eigenvalue decomposition

(of the relevant covariance matrix part)

Example for the “looks” of a covariance matrix (2 by 2, proper motions only):

sigma^2(pma)	cov (pma, pmd)
cov (pma, pmd)	sigma^2(pmd)

Note: cov(pma,pmd) = corr(pma, pmd)* sigma(pma) * sigma(pmd)

Solution of the Eigenvalue decomposition for 2 dimensions: (promised during the talk to be added here)

The maxima and minima of the variance (the eigenvalues of the matrix) are:

sigma^2(pmx) = 1/2* (sigma^2(pma)+sigma^2(pmd) + sqrt((sigma^2(pma)+sigma^2(pmd))^2-4cov^2(pma,pmd)))

sigma^2(pmy) = 1/2* (sigma^2(pma)+sigma^2(pmd) - sqrt((sigma^2(pma)+sigma^2(pmd))^2-4cov^2(pma,pmd)))

tan(theta) = (sigma^2(pma) - sigma^2(pmd)) / cov(pma,pmd) ; note 1: the +/- 180 deg ambiguity of the tangens does not matter in this case.
note 2: for cov(pma,pmd)=0, then theta=0 if sigma(pmd)>sigma(pma), else theta=90deg, and the values are trivial

Even more tedious formulae for 3 dimensions; better use matrix routines for 3d and higher dimensions.

Sorry for the clumsy formula notation, but I didn’t find the time to typeset them more nicely. Volunteers are invited to email me ☺

Thank you