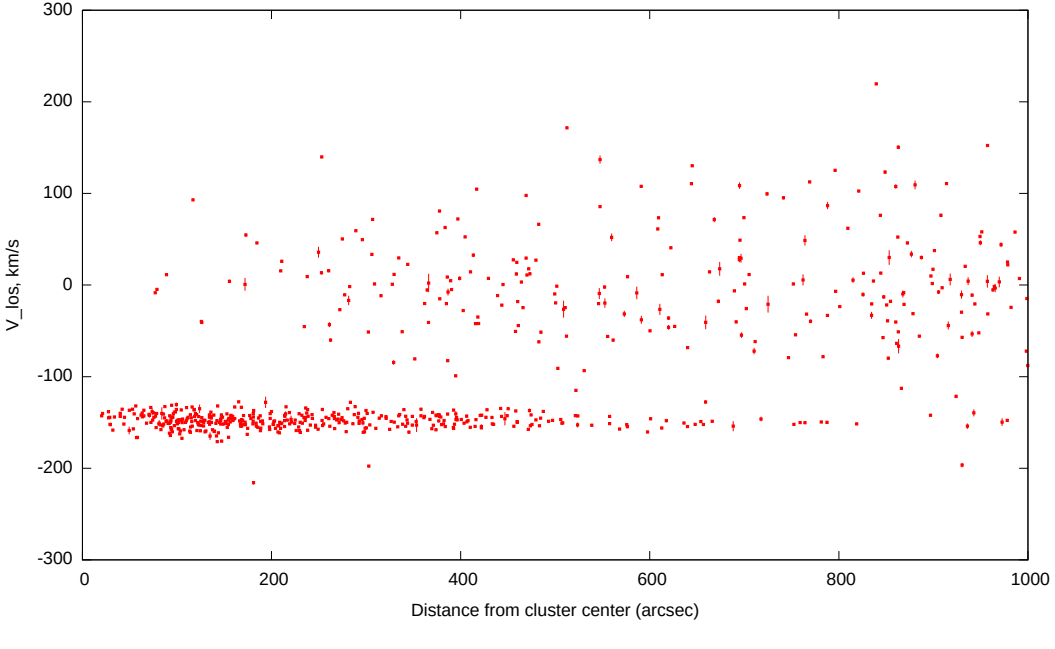
Fitting models to data

Eugene Vasiliev

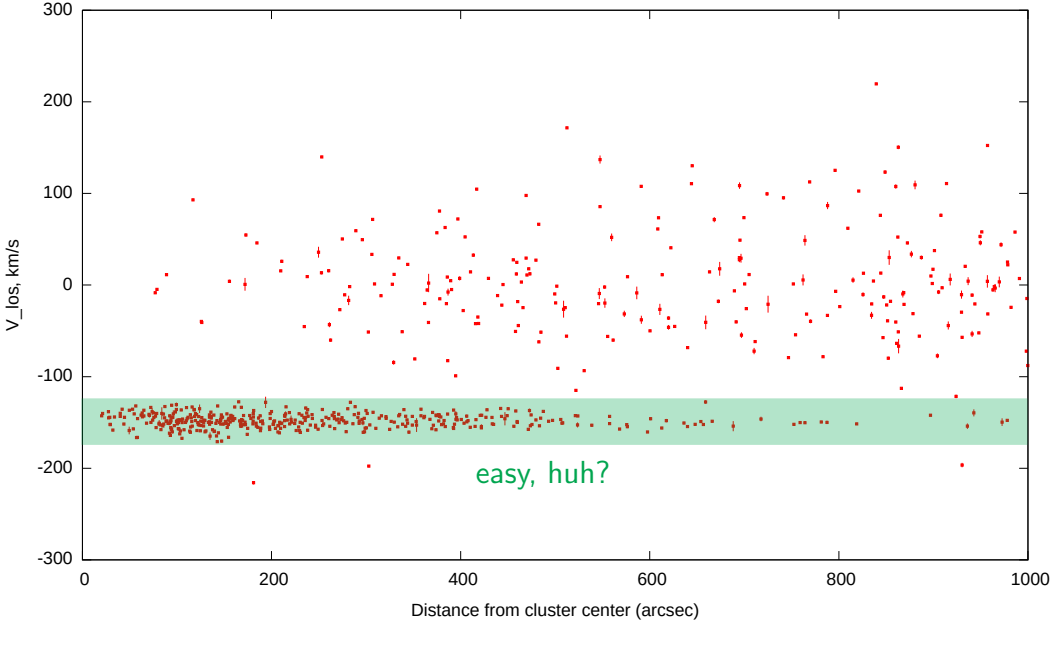
Institute of Astronomy, Cambridge

Summer School on Galactic Dynamics, Shanghai, June 2019

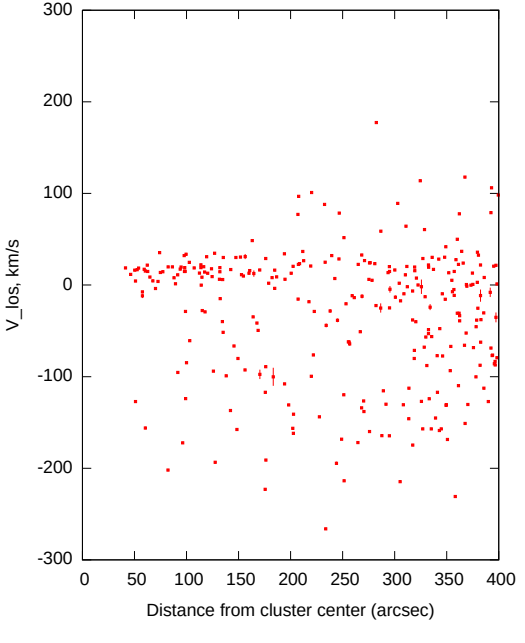
Line-of-sight velocities of globular clusters

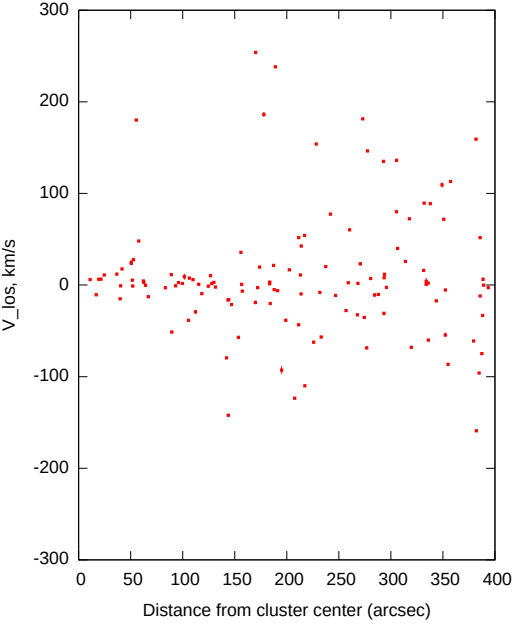


Line-of-sight velocities of globular clusters



Line-of-sight velocities of globular clusters





Conventional approach: https://en.wikipedia.org/wiki/Weighted_arithmetic_mean

Computing the mean value of noisy measurements:

$$\text{measured values:} \quad x_i$$

$$\text{errors (uncertainties):} \quad \delta x_i, \quad i = 1..N$$

$$\text{error-weighted mean value:} \quad \mu = \frac{\sum_{i=1}^{N} x_i/{\delta x_i}^2}{\sum_{i=1}^{N} 1/{\delta x_i}^2}$$

$$\text{uncertainty of the mean value:} \quad \delta\mu = \left(\sum_{i=1}^{N} 1/{\delta x_i}^2\right)^{-1/2}$$

Rejection of outliers (3σ–clipping):

or some other threshold

- if $|x_i - \mu| \geq 3\,\delta x_i$: eliminate this datapoint
- recompute μ from remaining datapoints
- repeat until the list of remaining points doesn't change

Maximum-likelihood computation of the error-weighted mean

1. Assume a model: e.g., all datapoints have the same true value μ, but are measured with some error which is normally distributed:

$$x_i \sim \mathcal{N}(\mu, \delta x_i) \equiv \frac{1}{\sqrt{2\pi}\,\delta x_i} \exp\left[-\frac{(x_i-\mu)^2}{2\,{\delta x_i}^2}\right]$$

2. Write down the likelihood function for the observed dataset given the model:

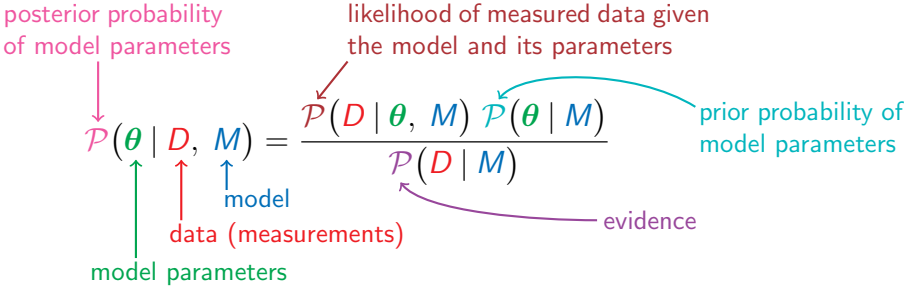
$$\mathscr{L} = \prod_{i=1}^{N} \mathcal{N}(x_i \mid \mu, \delta x_i), \quad \text{or} \quad \ln \mathscr{L} = -\tfrac{1}{2} \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{{\delta x_i}^2} - \sum_{i=1}^{N} \ln \delta x_i - \tfrac{N}{2} \ln 2\pi$$

3. Vary the parameters of the model (in this case, only μ) to maximize $\mathscr{L}$:

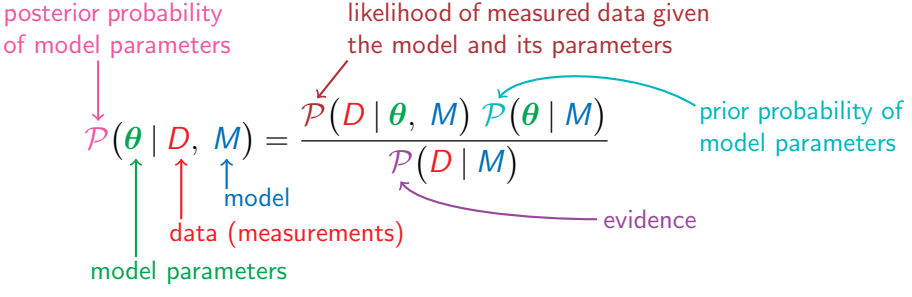
$$\frac{d \ln \mathscr{L}}{d\mu} = \sum_{i=1}^{N} \frac{x_i - \mu}{{\delta x_i}^2} = 0 \quad \Longrightarrow \quad \mu = \frac{\sum_{i=1}^{N} x_i/\delta x_i^2}{\sum_{i=1}^{N} 1/{\delta x_i}^2}$$

4. $\ln \mathscr{L}$ is a parabola near the best-fit μ: $\dfrac{d^2 \ln \mathscr{L}}{d\mu^2} = -\sum_{i=1}^{N} \dfrac{1}{{\delta x_i}^2}$;
confidence interval for μ: $\ln \mathscr{L}$ decreases by 1 from the best-fit value

Bayesian formulation



Bayesian formulation



$$\mathcal{P}(\boldsymbol{\theta} \mid D,\ M) = \frac{\mathcal{P}(D \mid \boldsymbol{\theta},\ M)\ \mathcal{P}(\boldsymbol{\theta} \mid M)}{\mathcal{P}(D \mid M)}$$

Posterior is a normalized probability distribution: $\int \mathcal{P}(\boldsymbol{\theta} \mid D,\ M)\ d\boldsymbol{\theta} = 1$,

hence the evidence is a "normalization factor":

$$\mathcal{P}(D \mid M) = \int \mathcal{P}(D \mid \boldsymbol{\theta},\ M)\ \underbrace{\mathcal{P}(\boldsymbol{\theta} \mid M)}_{\text{often this is a flat prior, i.e. constant [in some range]}}\ d\boldsymbol{\theta}$$

Treatment of measurement errors

D are measured (observed) data;

T are "true" (intrinsic) values predicted by the model M with parameters $\boldsymbol{\theta}$: $\mathcal{P}\left(T \mid \boldsymbol{\theta}, M\right)$ is the predicted distribution of true values;

$\mathcal{P}\left(D \mid T\right)$ is the measurement model: predicted distribution of D given T;

predicted distribution of observables is a marginalization over the [unknown] true values: $\mathcal{P}\left(D \mid \boldsymbol{\theta}, M\right) = \underbrace{\int \mathcal{P}\left(D \mid T\right) \mathcal{P}\left(T \mid \boldsymbol{\theta}, M\right) \, dT}_{\text{convolution with error distribution}}$

In the previous example, the model prediction was a single number μ: $\mathcal{P}\left(T \mid \mu, M\right) = \delta(T - \mu)$,

and the measurement model was a normal distribution with width δx_i.

Inferring the intrinsic dispersion

1. Make the model slightly more complicated:
   the true values are drawn from a Gaussian with mean μ and width σ,
   the observed values x_i are further perturbed by measurement errors δx_i

$$x_i \sim \underbrace{\mathcal{N}(\mu,\sigma) * \mathcal{N}(0,\delta x_i)}_{\text{convolution of two Gaussians is also a Gaussian}} = \mathcal{N}(\mu,\sigma_i), \quad \sigma_i = \sqrt{\sigma^2 + {\delta x_i}^2}$$

2. Write down the likelihood function $\mathscr{L}\left(\mu,\sigma \mid \{x_i,\delta x_i\}\right)$:

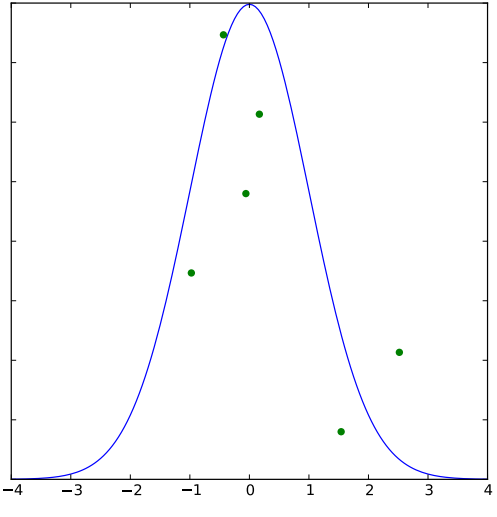
$$\ln \mathscr{L} = -\tfrac{1}{2}\sum_{i=1}^{N} \frac{(x_i-\mu)^2}{\sigma^2+{\delta x_i}^2} - \tfrac{1}{2}\sum_{i=1}^{N} \ln\left(\sigma^2+{\delta x_i}^2\right) - \tfrac{N}{2}\ln 2\pi$$

3. Vary the parameters (μ,σ) to maximize $\ln\mathscr{L}$: solve $\left\{\dfrac{\partial\mathscr{L}}{\partial\mu}=0, \dfrac{\partial\mathscr{L}}{\partial\sigma}=0\right\}$

4. The covariance matrix of the uncertainties on model parameters is

$$C = \begin{pmatrix} \delta\mu^2 & \rho\,\delta\mu\,\delta\sigma \\ \rho\,\delta\mu\,\delta\sigma & \delta\sigma^2 \end{pmatrix} = -\begin{pmatrix} \dfrac{\partial^2 \ln\mathscr{L}}{\partial\mu^2} & \dfrac{\partial^2 \ln\mathscr{L}}{\partial\mu\,\partial\sigma} \\ \dfrac{\partial^2 \ln\mathscr{L}}{\partial\mu\,\partial\sigma} & \dfrac{\partial^2 \ln\mathscr{L}}{\partial\sigma^2} \end{pmatrix}^{-1}$$

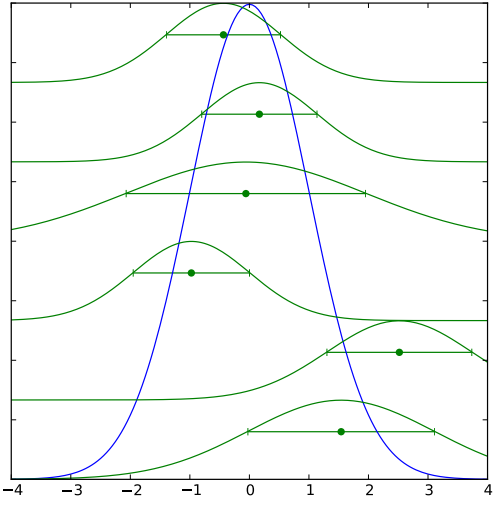
Inferring the intrinsic dispersion



blue: intrinsic distribution

green: true values of sampled points

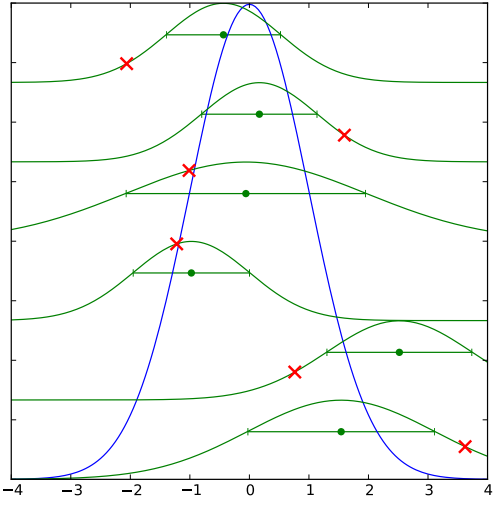
Inferring the intrinsic dispersion



blue: intrinsic distribution

green: true values of sampled points with measurement uncertainties

Inferring the intrinsic dispersion

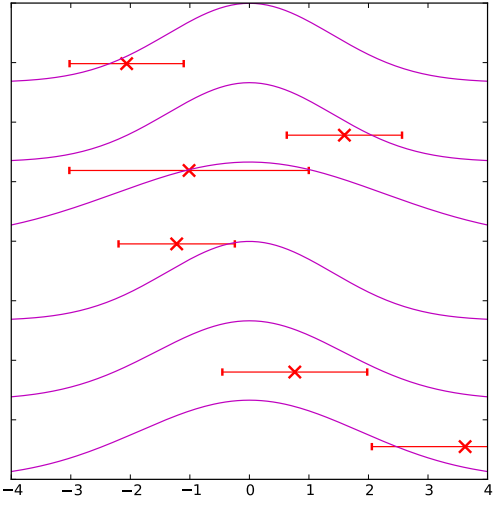


blue: intrinsic distribution

green: true values of sampled points
with measurement uncertainties

red: measured values (perturbed by errors)

Inferring the intrinsic dispersion



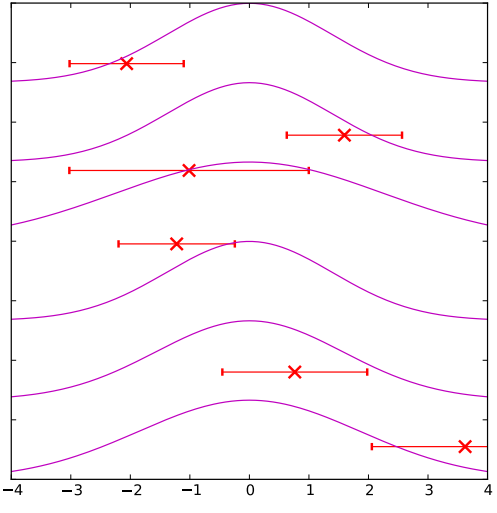
blue: intrinsic distribution

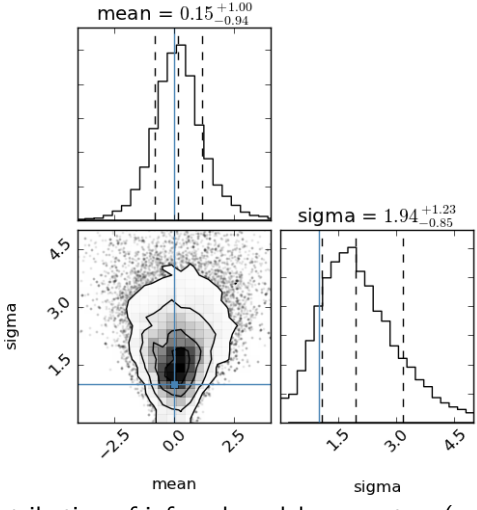
green: true values of sampled points
with measurement uncertainties

red: measured values (perturbed by errors)

magenta: prob.distrib. for each measured point

Inferring the intrinsic dispersion: deconvolution





distribution of inferred model parameters (μ, σ) obtained from a Markov Chain Monte Carlo run

Treatment of outliers: mixture models

1. Assume a two-component model specified by two intrinsic distributions, convolved with individual measurement errors for each datapoint:
   - points belonging to the object of interest: $x_i \sim \mathcal{F}_{\rm int}(\boldsymbol{\theta}) * \mathcal{N}(0, \delta x_i)$
   - ouliers: $x_i \sim \mathcal{F}_{\rm out}(\boldsymbol{\zeta}) * \mathcal{N}(0, \delta x_i)$

   could be a Gaussian, but not necessarily

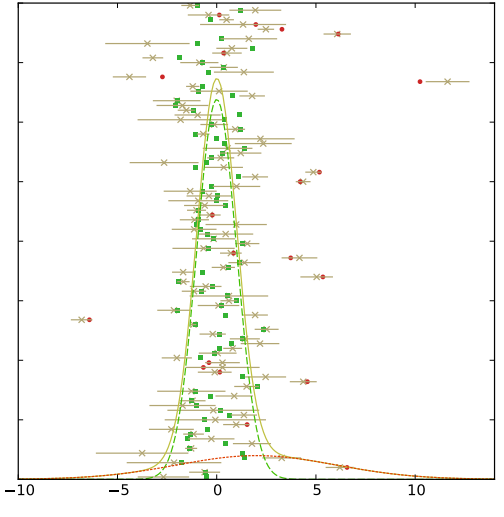
   $\boldsymbol{\theta}, \boldsymbol{\zeta}$ are the parameters of the intrinsic distributions (e.g., mean and width of Gaussians)

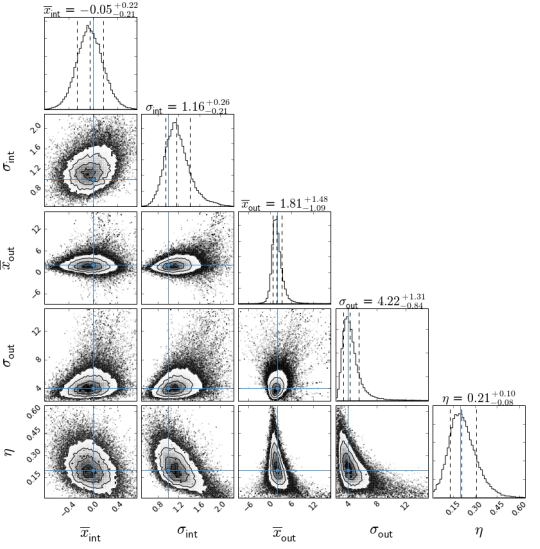
2. Assume that an (unknown) fraction η of all datapoints are outliers: the probability distribution of the mixture model for datapoint x_i is $x_i \sim \mathcal{F}_{{\rm mix},i} \equiv \big[(1-\eta)\,\mathcal{F}_{\rm int}(\boldsymbol{\theta}) + \eta\,\mathcal{F}_{\rm out}(\boldsymbol{\zeta})\big] * \mathcal{N}(0, \delta x_i)$, and the likelihood of the entire model is $\mathscr{L}\big(\boldsymbol{\theta}, \boldsymbol{\zeta}, \eta \mid \{x_i, \delta x_i\}\big) = \prod_{i=1}^{N} \mathcal{F}_{{\rm mix},i}$

3. Define suitable priors $\mathcal{P}$ for the nuisance parameters $\boldsymbol{\zeta}, \eta$

4. Obtain the posterior probability distribution for the parameters of interest $\boldsymbol{\theta}$ by marginalizing over the nuisance parameters $\boldsymbol{\zeta}, \eta$: $\mathscr{P}\big(\boldsymbol{\theta} \mid \{x_i\}\big) = \int \mathscr{L}\big(\boldsymbol{\theta}, \boldsymbol{\zeta}, \eta \mid \{x_i\}\big)\, \mathcal{P}(\boldsymbol{\zeta}, \eta)\, d\boldsymbol{\zeta}\, d\eta$; determine the confidence intervals for $\boldsymbol{\theta}$ from this posterior distribution

Treatment of outliers: mixture models





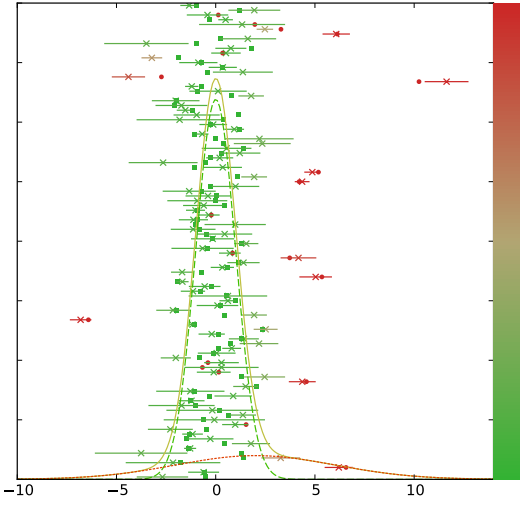
Mixture models and classification

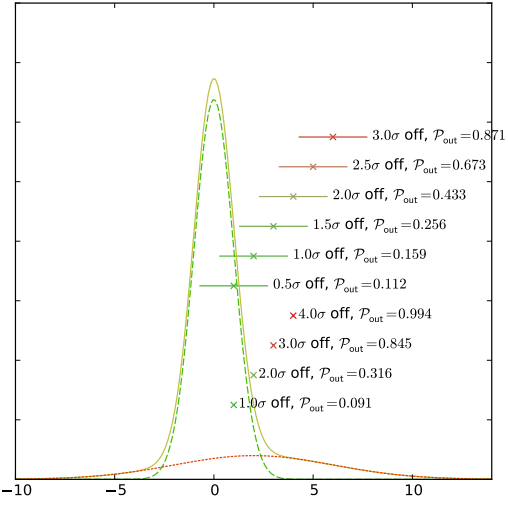
- Each datapoint has the same prior probability η of being an outlier; however, the posterior probability does depend on the measured value x_i and its uncertainty δx_i, as well as the model parameters $\boldsymbol{\theta}$, $\boldsymbol{\zeta}$:

$$\mathcal{P}_{\mathrm{out}}(x_i, \delta x_i \mid \boldsymbol{\theta}, \boldsymbol{\zeta}, \eta) = \frac{\left[\eta\, \mathcal{F}_{\mathrm{out}}(\boldsymbol{\zeta}) * \mathcal{N}(0, \delta x_i)\right](x_i)}{\left[\left\{(1-\eta)\, \mathcal{F}_{\mathrm{int}}(\boldsymbol{\theta}) + \eta\, \mathcal{F}_{\mathrm{out}}(\boldsymbol{\zeta})\right\} * \mathcal{N}(0, \delta x_i)\right](x_i)}$$

- For the best-fit values of parameters, $\sum_{i=1}^{N} \mathcal{P}_{\mathrm{out}}(x_i, \delta x_i) = N\,\eta$
- There is no single "N-σ" criterion: if there were no model for outliers, one couldn't reject a point even when it is 10σ off!
- The dataset should contain enough contaminants to reliably infer their fraction η and the parameters $\boldsymbol{\zeta}$ for $\mathcal{F}_{\mathrm{out}}$
- Probabilistic membership classification should be carried onward to subsequent modelling procedures, if possible.

Mixture models, classification and rejection of outliers



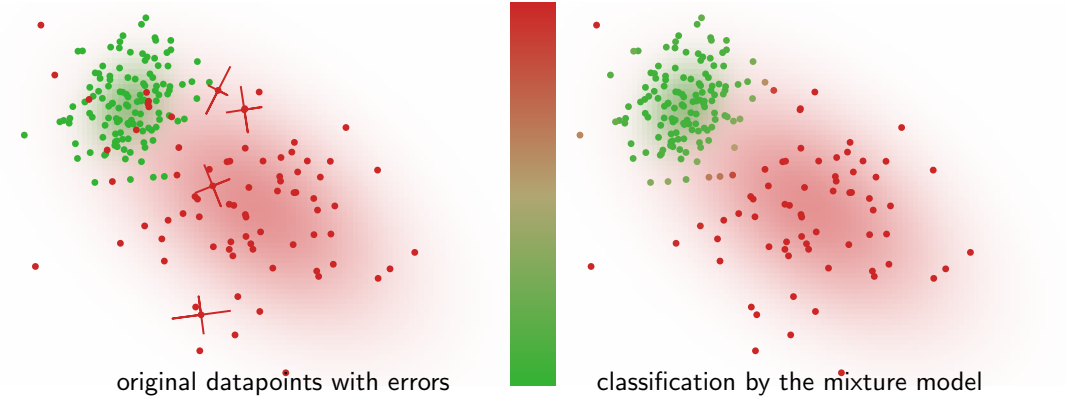


Multidimensional case

- D-dimensional Gaussian with mean $\boldsymbol{\mu}$ and covariance matrix Σ:

$$\mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, \Sigma) = \frac{1}{\sqrt{(2\pi)^D \det \Sigma}} \exp\left[-\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu}) \right]$$

- Measurement errors for i-th datapoint described by error covariance matrix δx_i
- Convolution of two Gaussians is also a Gaussian with covariance $\Sigma + \delta x_i$

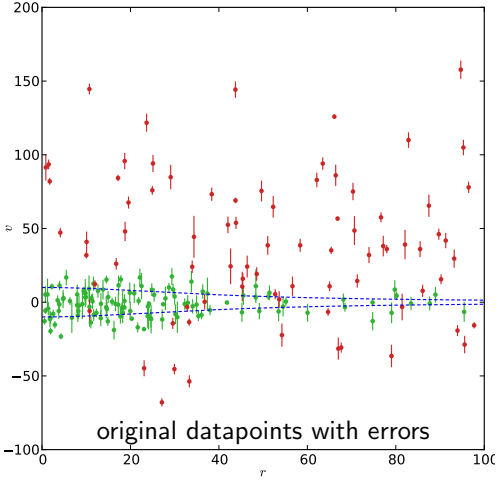


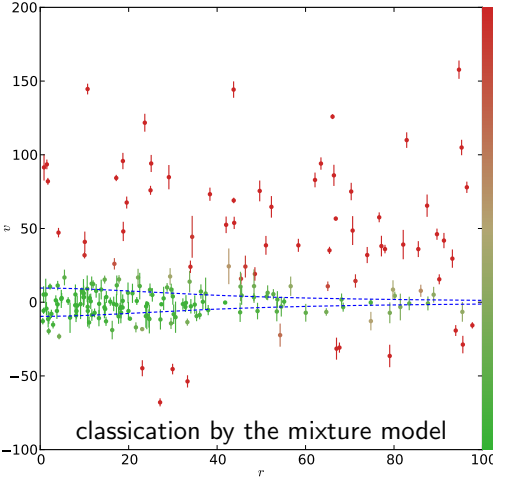
original datapoints with errors

classification by the mixture model

Additional parameters in the model

The probability distribution $\mathcal{F}_{\text{int}}$ and the fraction of outliers η
may depend on some additional parameters ς and measured properties $\{\xi_i\}$
(e.g., scale radius a, the distance R_i of a star from the cluster center, etc.)





Fitting dynamical models to discrete-kinematic data

Example: Plummer-like model for the cluster, uniform contamination:

$$\mathcal{F}_{\rm int}\big(\{R_i, v_i, \delta v_i\} \mid a, \overline{v}, \sigma_0\big) = \mathcal{N}\Big(v_i \mid \overline{v}, \sqrt{\sigma^2(R_i) + \delta v_i^2}\Big)\, \frac{1 + (R_{\rm max}/a)^2}{\big[1 + (R_i/a)^2\big]^2}$$

scale radius — cluster mean velocity — central velocity dispersion

$$\sigma(R_i \mid a, \sigma_0) \equiv \frac{\sigma_0}{\big[1 + (R_i/a)^2\big]^{1/4}}$$

normalized surface density

mean and dispersion of contaminants

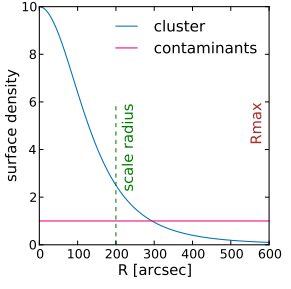
$$\mathcal{F}_{\rm out}\big(\{v_i, \delta v_i\} \mid \overline{v}_{\rm out}, \sigma_{\rm out}\big) = \mathcal{N}\Big(v_i \mid \overline{v}_{\rm out}, \sqrt{\sigma_{\rm out}^2 + \delta v_i^2}\Big)$$

Distribution function of the mixture model:

$$\mathcal{F}_{\rm mix} = (1 - \eta)\, \mathcal{F}_{\rm int} + \eta\, \mathcal{F}_{\rm out}$$

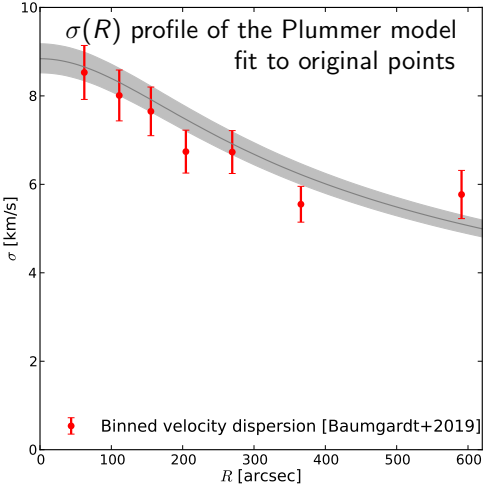
parameters: $a, \overline{v}, \sigma_0, \overline{v}_{\rm out}, \sigma_{\rm out}, \eta$

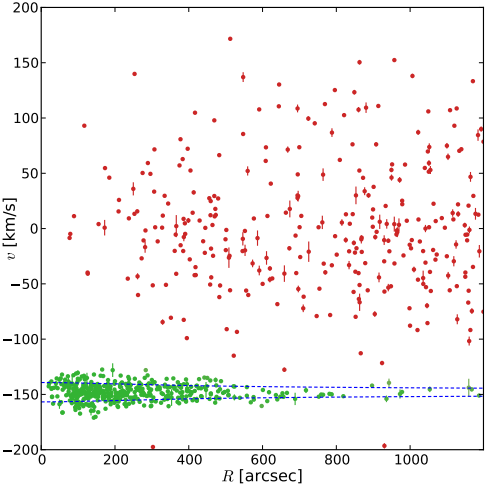
fraction of contaminants



Fitting dynamical models to discrete-kinematic data

Example: globular cluster NGC 6656

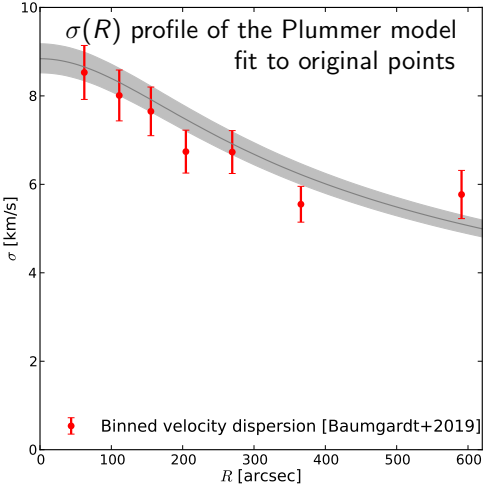


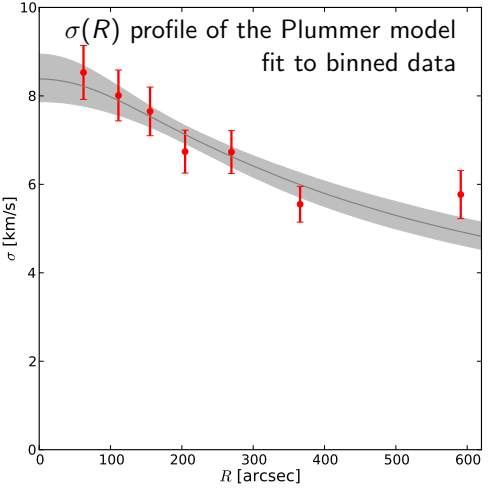


Fitting dynamical models to discrete-kinematic data

Example: globular cluster NGC 6656

General problem with parametric models:
inferred uncertainty intervals are too small

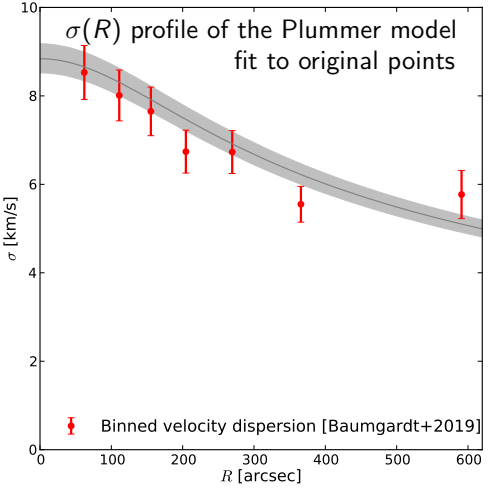


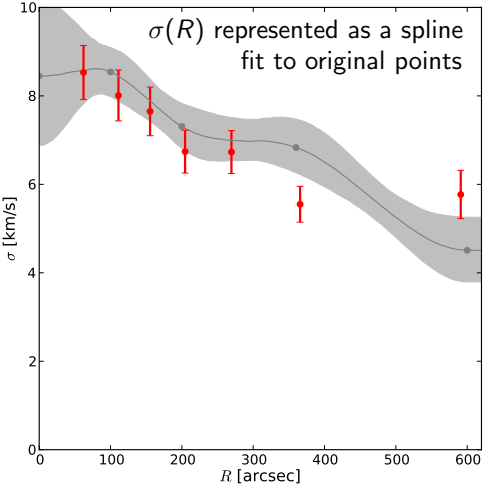


Fitting "nonparametric" models to discrete-kinematic data

Consider a more flexible (but not dynamically motivated) dispersion profile: $\sigma(R)$ represented as a cubic spline with fixed nodes and free coefficients.

Similar uncertainties as binned data, but in a mathematically consistent model.

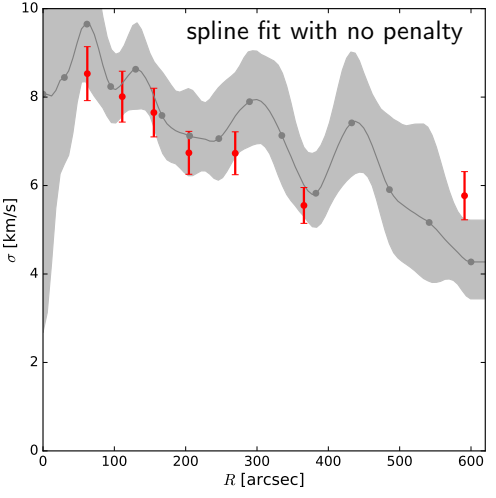


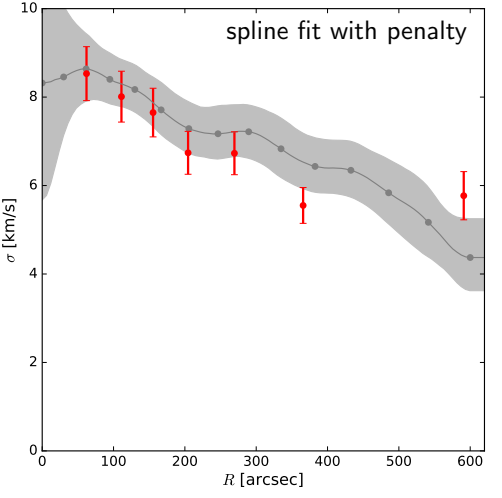


Penalization of overly flexible models

If one has too many free parameters, the models would overfit the data.

To prevent this, one needs to impose stronger priors on the parameters, e.g., penalizing large variations between adjacent spline values.





Penalization of overly flexible models

If one has too many free parameters, the models would overfit the data.

To prevent this, one needs to impose stronger priors on the parameters, e.g., penalizing large variations between adjacent spline values.

There are objective methods for determining the optimal value of penalty (smoothing parameter), based on cross-validation:

- adopt a particular value of the smoothing parameter λ;
- split the data sample into two parts – training set and validation set;
- find the best-fit model for the training set;
- evaluate the goodness-of-fit of this model for the validation set;
- average this over different choices of training and validation subsets;
- adjust the parameter λ to maximize this validation score;

The same approach is used in machine learning to prevent overfitting

Binned vs. non-binned data

Traditional (binned) approach:

- Clean up the sample
- Bin datapoints (e.g., in radius R)
- Compute the mean and dispersion $\overline{v}, \sigma$ and their standard deviations in each bin
- Subtract the measurement uncertainty ("error") in quadrature from σ
- Fit $\sigma(R)$ from the model to the binned data $\{\sigma_b \pm \delta\sigma_b\}\big|_{b=1}^{N_{\rm bin}}$

Non-binned approach:

- Assume models (parametric or free-form) for the distribution of objects of interest and outliers
- Write down a mixture model: $\mathcal{F} = (1-\eta)\mathcal{F}_{\rm int} + \eta\mathcal{F}_{\rm out}$
- Find the best-fit parameters and their uncertainties from a Monte Carlo simulation
- Compute the membership probability of each datapoint

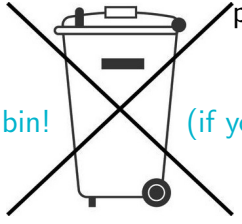
Binned vs. non-binned data

Traditional (binned) approach:

- Clean up the sample
- Bin datapoints (e.g., in radius R)
- Compute the mean and dispersion $\overline{v}, \sigma$ and their standard deviations in each bin
- Subtract the measurement uncertainty ("error") in quadrature from σ
- Fit $\sigma(R)$ from the model to the binned data $\{\sigma_b \pm \delta\sigma_b\}\big|_{b=1}^{N_{\rm bin}}$

Non-binned approach:

- Assume models (parametric or free-form) for the distribution of objects of interest and outliers
- Write down a mixture model: $\mathcal{F} = (1-\eta)\mathcal{F}_{\rm int} + \eta\mathcal{F}_{\rm out}$
- Find the best-fit parameters and their uncertainties from a Monte Carlo simulation
- Compute the membership probability of each datapoint



Pros and cons of fitting models directly to discrete data

+ outlier rejection or fitting multiple populations are straightforward
+ easy to account for selection function and measurement uncertainties
+ use all available information (no coarse-graining)
– computationally demanding in case of large datasets
– marginalization and convolution often non-analytic $\Rightarrow$ expensive
– difficult to fit discrete models (Schwarzschild, M2M) to discrete data

Applications

- Membership determination for dSph [Walker+ 2009]
- Jeans models of globular clusters and dSph [Watkins+ 2013; Zhu+ 2016]
- DF-block model of Milky Way nuclear star cluster [Magorrian 2019]
- Proof-of-concept Schwarzschild and M2M methods for discrete data
  [Chanamé+ 2008; Breddels 2013 (thesis); Hunt & Kawata 2013; Bovy+ 2017]
- Parametric DF models of the Milky Way disk
  [McMillan & Binney 2013; Bovy & Rix 2013; Trick+ 2016]