Bayes vs the virial theorem

John Magorrian

Gaia Challenge Workshop,
Surrey, 20 August 2013

Science fiction!

“superGaia”: Gaia w/o observational errors, selection effects.

An idealised problem

superGaia gives us $(\mathbf{x}, \mathbf{v})$ for every star in the Galaxy.
What’s the gravitational potential in which they move?

Assume: collisionless, $\partial_t = 0 \rightarrow$ 2 functions + Jeans thm

$$\Phi(\mathbf{x}), \qquad f(\mathbf{x}, \mathbf{v}) = f(\mathbf{J}), \qquad \mathbf{J} = \mathbf{J}(\mathbf{x}, \mathbf{v}|\Phi)$$

Want this

observations are a **discrete realisation** of this

Science fiction!

“superGaia”: Gaia w/o observational errors, selection effects.

An idealised problem

superGaia gives us $(\mathbf{x}, \mathbf{v})$ for every star in the Galaxy.
What’s the gravitational potential in which they move?

Assume: collisionless, $\partial_t = 0 \rightarrow$ 2 functions + Jeans thm

$$\Phi(\mathbf{x}), \qquad f(\mathbf{x}, \mathbf{v}) = f(\mathbf{J}), \qquad \mathbf{J} = \mathbf{J}(\mathbf{x}, \mathbf{v}|\Phi)$$

Want this

observations are a **discrete realisation** of this

Standard methods don't work (well)

I. Parametrized DF

- Guess some form for $f(\mathbf{J}|\alpha)$, parametrised by α
- Try a Φ
- Marginalise or fiddle parameters α
- Assign likelihood to Φ.

But what if your guess for $f(\mathbf{J}|\alpha)$ is wrong?
(Ex: real galaxy has streams/substructure...)

II. Schwarzschild's method

- Pick a trial potential Φ
- Launch orbit from each observed $(\mathbf{x}_n^\star, \mathbf{v}_n^\star)$ in Φ
- ermm..

All Φ equally likely!

Standard methods don't work (well)

I. Parametrized DF

- Guess some form for $f(\mathbf{J}|\alpha)$, parametrised by α
- Try a Φ
- Marginalise or fiddle parameters α
- Assign likelihood to Φ.

But what if your guess for $f(\mathbf{J}|\alpha)$ is wrong?
(Ex: real galaxy has streams/substructure...)

II. Schwarzschild's method

- Pick a trial potential Φ
- Launch orbit from each observed $(\mathbf{x}_n^\star, \mathbf{v}_n^\star)$ in Φ
- ermm..

All Φ equally likely!

Counterpoint

But wait, surely this problem is trivial...?

CBE is

$$\mathbf{v} \cdot \frac{\partial f}{\partial \mathbf{x}} = \frac{\partial \Phi}{\partial \mathbf{x}} \cdot \frac{\partial f}{\partial \mathbf{v}}.$$

If we knew $f(\mathbf{x}, \mathbf{v})$ perfectly then moment-based methods (Jeans, virial) would cough up Φ very easily.

But

All we have is N stars drawn from f:

- We don't know $f(\mathbf{x}, \mathbf{v})$ itself,
- nor do we know its moments particularly well!
- Want to avoid taking derivatives of moments, or solving awkward implicit relns among moments.

Another way of tackling the problem

Motivation: Shirley, there must be a Bayesian version of the (generalised) virial theorem...

(unreadable `arXiv:1303.6099`)

Broad-brush restatement of problem

(JM, `arXiv:1303.6099`)

Problem:

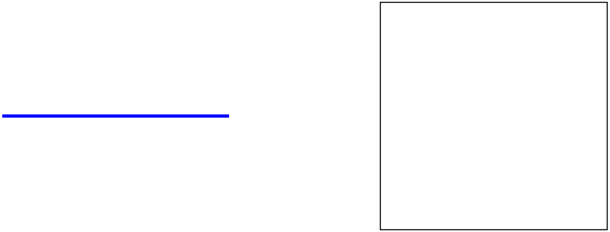
- We have two unknown functions, $\Phi(\mathbf{x})$ and $f(\mathbf{x}, \mathbf{v})$.
- Our data D is a discrete sample of f, but want Φ.
- Jeans: $f(\mathbf{x}, \mathbf{v}) = f(\mathbf{J})$, where $\mathbf{J} = \mathbf{J}(\mathbf{x}, \mathbf{v}|\Phi)$.

Solution: Marginalise f!

- But how to represent it?

new idea alert!

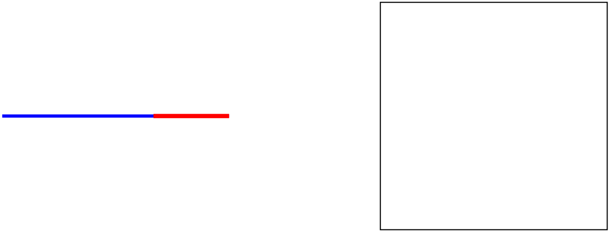
A Dirichlet process mixture model for the DF



Have: unit stick of starlight + big action-space box.

1. Snap off a chunk π_k of starlight
    - Draw $\beta_k \sim \text{Beta}(1, \alpha)$ and set $\pi_k = \beta_k \times$ [stick length]
2. Toss this chunk π_k into the action-space box
    - where it lands given by uniform prior in **J**
3. Smear out this π_k in action space:
    - Spike of height π_k at some point $\mathbf{J}_{\text{land}}$
    - convolve with Gaussian having randomly chosen inverse covariance matrix drawn from uninformative prior (Wishart)
4. Apply steps 1–4 to remainder of stick.

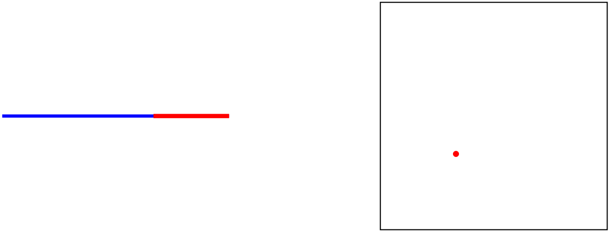
A Dirichlet process mixture model for the DF



Have: unit stick of starlight + big action-space box.

1. Snap off a chunk π_k of starlight
    - Draw $\beta_k \sim \text{Beta}(1, \alpha)$ and set $\pi_k = \beta_k \times$ [stick length]
2. Toss this chunk π_k into the action-space box
    - where it lands given by uniform prior in **J**
3. Smear out this π_k in action space:
    - Spike of height π_k at some point $\mathbf{J}_{\rm land}$
    - convolve with Gaussian having randomly chosen inverse covariance matrix drawn from uninformative prior (Wishart)
4. Apply steps 1–4 to remainder of stick.

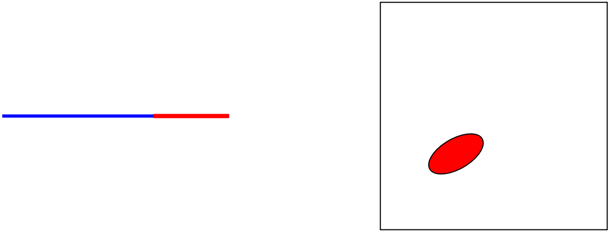
A Dirichlet process mixture model for the DF



Have: unit stick of starlight + big action-space box.

1. Snap off a chunk π_k of starlight
   - Draw $\beta_k \sim \text{Beta}(1, \alpha)$ and set $\pi_k = \beta_k \times$ [stick length]
2. Toss this chunk π_k into the action-space box
   - where it lands given by uniform prior in **J**
3. Smear out this π_k in action space:
   - Spike of height π_k at some point $\mathbf{J}_{\text{land}}$
   - convolve with Gaussian having randomly chosen inverse covariance matrix drawn from uninformative prior (Wishart)
4. Apply steps 1–4 to remainder of stick.

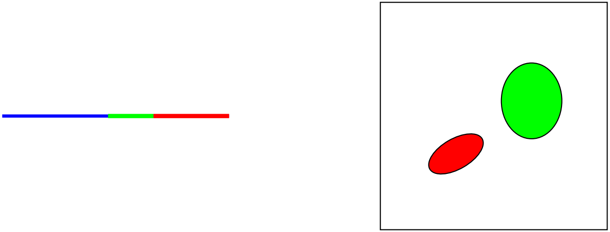
A Dirichlet process mixture model for the DF



Have: unit stick of starlight + big action-space box.

1. Snap off a chunk π_k of starlight
    - Draw $\beta_k \sim \mathrm{Beta}(1, \alpha)$ and set $\pi_k = \beta_k \times$ [stick length]
2. Toss this chunk π_k into the action-space box
    - where it lands given by uniform prior in $\mathbf{J}$
3. Smear out this π_k in action space:
    - Spike of height π_k at some point $\mathbf{J}_{\mathrm{land}}$
    - convolve with Gaussian having randomly chosen inverse covariance matrix drawn from uninformative prior (Wishart)
4. Apply steps 1–4 to remainder of stick.

A Dirichlet process mixture model for the DF



Have: unit stick of starlight + big action-space box.

1. Snap off a chunk π_k of starlight
    - Draw $\beta_k \sim \text{Beta}(1, \alpha)$ and set $\pi_k = \beta_k \times$ [stick length]
2. Toss this chunk π_k into the action-space box
    - where it lands given by uniform prior in **J**
3. Smear out this π_k in action space:
    - Spike of height π_k at some point $\mathbf{J}_{\text{land}}$
    - convolve with Gaussian having randomly chosen inverse covariance matrix drawn from uninformative prior (Wishart)
4. Apply steps 1–4 to remainder of stick.

Dirichlet process mixture = distribution of blobs

DF is composed of K blobs (“streams”?), each having

- some probability mass π_k, ($k = 1...K$)
- centred on some $\mathbf{J}_k$, and
- with inverse covariance Λ_k (size/shape).

Single free hyperparameter α controls clumpiness of DF.

Different ways of thinking about Dirichlet process:

1. Stick-breaking construction
2. Chinese Restuarant process
3. Limit of Dirichlet distribution as cell size → 0.

Natural way of representing distribution over distributions (i.e., prior on DF).

Dirichlet process mixture = distribution of blobs

DF is composed of K blobs (“streams”?), each having

- some probability mass π_k, ($k = 1...K$)
- centred on some $\mathbf{J}_k$, and
- with inverse covariance $\boldsymbol{\Lambda}_k$ (size/shape).

Single free hyperparameter α controls clumpiness of DF.

Different ways of thinking about Dirichlet process:

1. Stick-breaking construction
2. Chinese Restuarant process
3. Limit of Dirichlet distribution as cell size $\to$ 0.

Natural way of representing distribution over distributions (i.e., prior on DF).

How to apply it to constraining Φ

Suppose we have a big list of potentials Φ to try.
Given a trial Φ, each $(\mathbf{x}_n^\star, \mathbf{v}_n^\star) \to (\mathbf{J}_n^\star, \theta_n^\star)$.
Then given a DF $f(\mathbf{J})$ the likelihood is simply

$$\mathrm{p}(D|\Phi, f) = \prod_{n=1}^{N} f(\mathbf{J}_n^\star).$$

Easy! Sum $\mathrm{p}(D|\Phi, f)$ over all f generated by the stick-breaking procedure! Obtain **marginalised likelihood** $\mathrm{p}(D|\Phi)$.

$$\begin{aligned}\mathrm{p}(D|\Phi) &= \mathrm{p}(\mathbf{J}_1^\star, ..., \mathbf{J}_N^\star) \\ &= \sum_f \mathrm{p}(D|\Phi, f) = \int \mathrm{dp}(f) \prod_{n=1}^{N} f(\mathbf{J}_n^\star) \\ &= \text{ginormous sum over partitions of } N \text{ points}\end{aligned}$$

NB: $\mathrm{p}(D|\Phi)$ depends only on how the $\mathbf{J}_n^\star$ are distributed!

How to apply it to constraining Φ

Suppose we have a big list of potentials Φ to try.
Given a trial Φ, each $(\mathbf{x}_n^\star, \mathbf{v}_n^\star) \to (\mathbf{J}_n^\star, \theta_n^\star)$.
Then given a DF $f(\mathbf{J})$ the likelihood is simply

$$\mathrm{p}(D|\Phi, f) = \prod_{n=1}^{N} f(\mathbf{J}_n^\star).$$

Easy! Sum $\mathrm{p}(D|\Phi, f)$ over all f generated by the stick-breaking procedure! Obtain **marginalised likelihood** $\mathrm{p}(D|\Phi)$.

$$\begin{aligned}\mathrm{p}(D|\Phi) &= \mathrm{p}(\mathbf{J}_1^\star, ..., \mathbf{J}_N^\star) \\ &= \sum_f \mathrm{p}(D|\Phi, f) = \int \mathrm{dp}(f) \prod_{n=1}^{N} f(\mathbf{J}_n^\star) \\ &= \text{ginormous sum over partitions of } N \text{ points}\end{aligned}$$

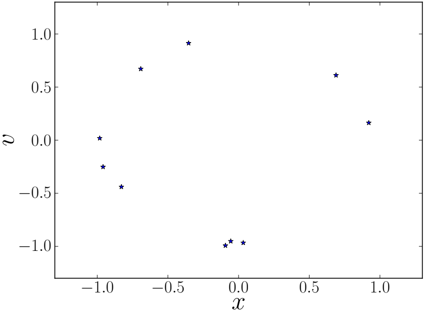
NB: $\mathrm{p}(D|\Phi)$ depends only on how the $\mathbf{J}_n^\star$ are distributed!

A really really toy problem

Ptles moving in 1d SHO potential, $\Phi(x) = \frac{1}{2}\omega^2 x^2$.

(x, v) phase plane:

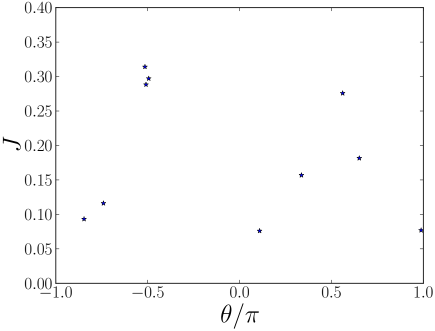
Any guesses about ω?



A really really toy problem

Ptles moving in 1d SHO potential, $\Phi(x) = \frac{1}{2}\omega^2 x^2$.

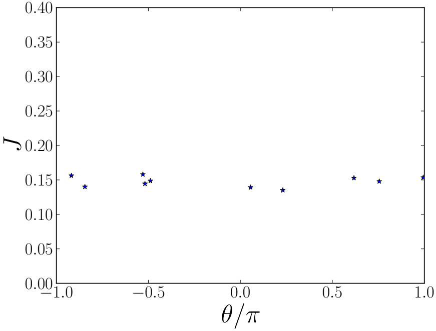
Different choices of ω give different (J, θ) distns:
$\omega = 0.5$



A really really toy problem

Ptles moving in 1d SHO potential, $\Phi(x) = \frac{1}{2}\omega^2 x^2$.

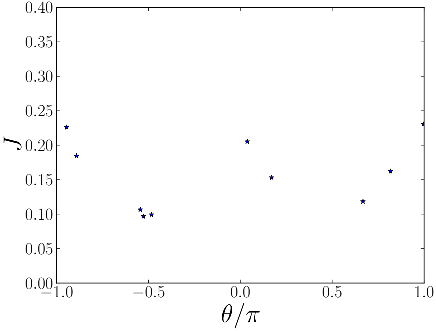
Different choices of ω give different (J, θ) distns:
$\omega = 1$



A really really toy problem

Ptles moving in 1d SHO potential, $\Phi(x) = \frac{1}{2}\omega^2 x^2$.

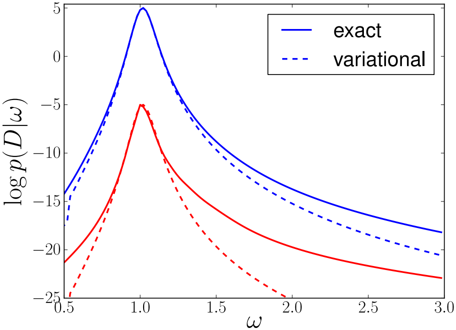
Different choices of ω give different (J, θ) distns:
$\omega = 1.5$



A really really toy problem

Ptles moving in 1d SHO potential, $\Phi(x) = \frac{1}{2}\omega^2 x^2$.

$\mathrm{p}(D|\omega)$ using DPM prior on DF



$\mathrm{p}(D|\Phi)$ is biggest for Φ that make **J** distn “sharpest”:

- **DPM naturally “likes” sharply peaked, clumpy DFs!**

Plot shows two different schemes for calculating $\mathrm{p}(D|\Phi)$.

Aside: relation to "minimum entropy" idea

(Peñarrubia, Koposov, Walker 2012, also Helmi talk)

A galaxy's entropy,

$$S[f] = -\int f \log f \, \mathrm{d}^3\mathbf{x} \, \mathrm{d}^3\mathbf{v},$$

is *independent* of the assumed potential Φ.

We may average the DF over angles in an assumed Φ:

$$\bar{f}(\mathbf{J}) = \frac{1}{(2\pi)^3} \int f(\mathbf{J}, \theta) \mathrm{d}^3\theta.$$

Note that:

- $\bar{f} = f$ if assumed Φ is true potential (because θ flat)
- $S[\bar{f}] \geq S[f]$ because $S[f]$ indep of Φ.
- So, correct Φ minimises entropy of orbit-avgd DF.

Aside: relation to “minimum entropy” idea

(Peñarrubia, Koposov, Walker 2012, also Helmi talk)

A galaxy’s entropy,

$$S[f] = -\int f \log f \, \mathrm{d}^3\mathbf{x} \, \mathrm{d}^3\mathbf{v},$$

is *independent* of the assumed potential Φ.
We may average the DF over angles in an assumed Φ:

$$\bar{f}(\mathbf{J}) = \frac{1}{(2\pi)^3} \int f(\mathbf{J}, \theta) \mathrm{d}^3\theta.$$

Note that:

- $\bar{f} = f$ if assumed Φ is true potential (because θ flat)
- $S[\bar{f}] \geq S[f]$ because $S[f]$ indep of Φ.
- So, correct Φ minimises entropy of orbit-avgd DF.

But does it work?

Application to Gaia challenge problems

Apply to $(\mathbf{x}, \mathbf{v})$ from Matt W/Jorge P's spherical models in `spherical_df.tar.gz`.

I assume a parametrised mass density of the form

$$\rho(r) = \rho_0 r^{-\gamma}(1 + r)^{\gamma-3}.$$

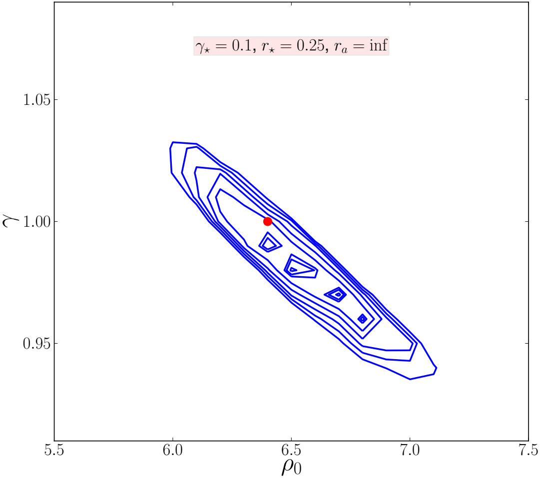
Matt's $(\mathbf{x}, \mathbf{v})$ "data" come from models having

- $\rho_0 = 6.4$ and $\gamma = 1$, but
- a range of different stellar DFs (10^4 stars each).

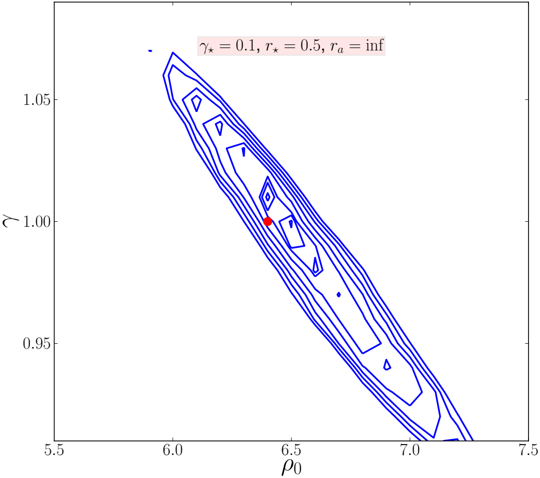
For each dataset D I **estimate** $\mathrm{p}(D|\rho_0, \gamma)$ using VB method.

(NB: Following plots updated during the meeting: still not perfect, but VB is an *approximate* method...)

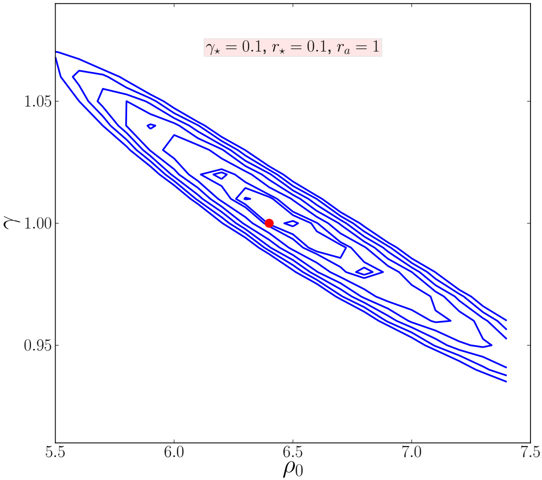
Results: various stellar DFs in NFW potential



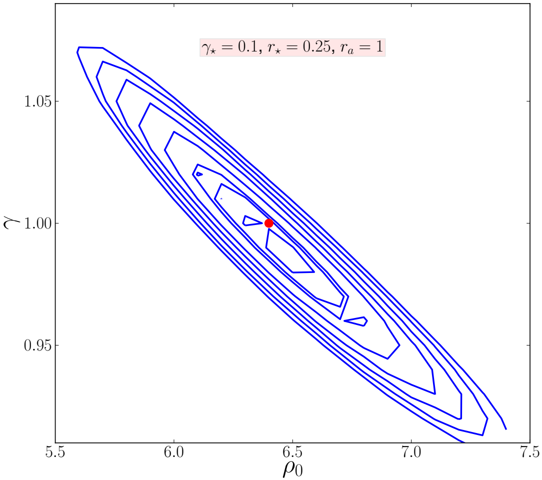
Results: various stellar DFs in NFW potential



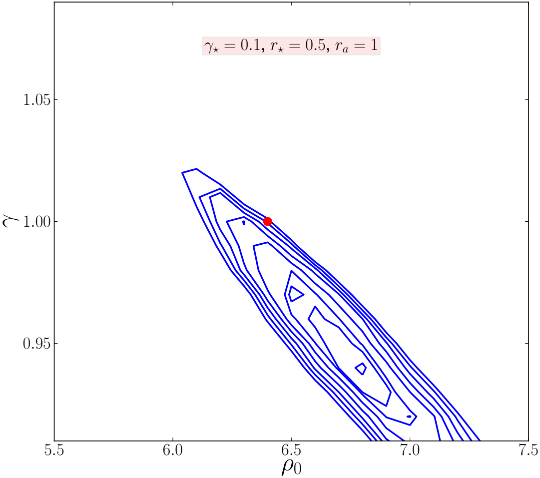
Results: various stellar DFs in NFW potential



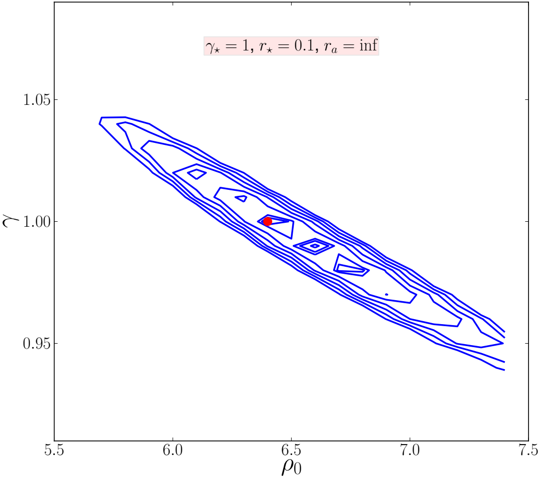
Results: various stellar DFs in NFW potential



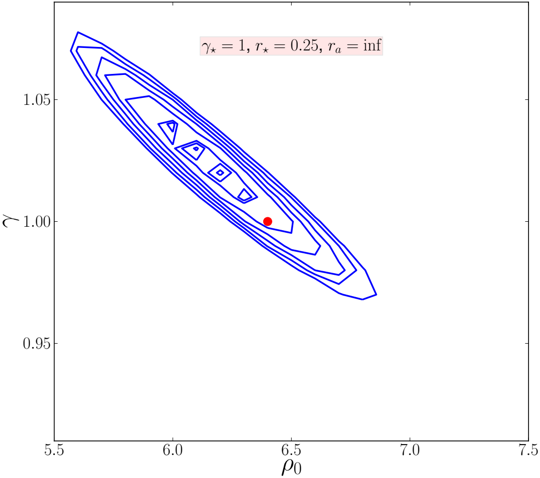
Results: various stellar DFs in NFW potential



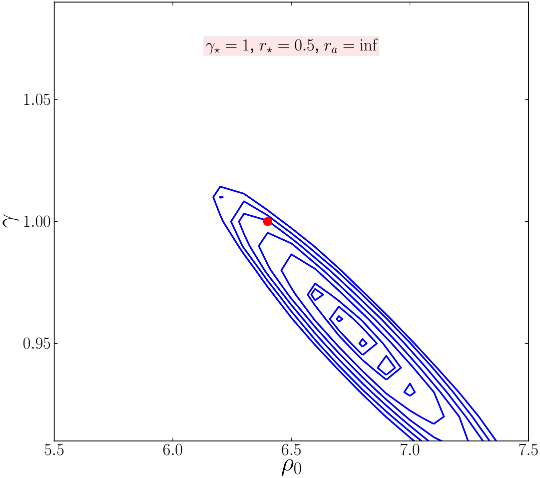
Results: various stellar DFs in NFW potential



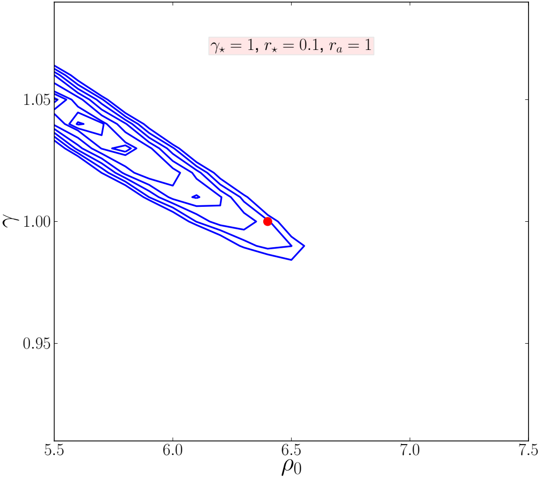
Results: various stellar DFs in NFW potential



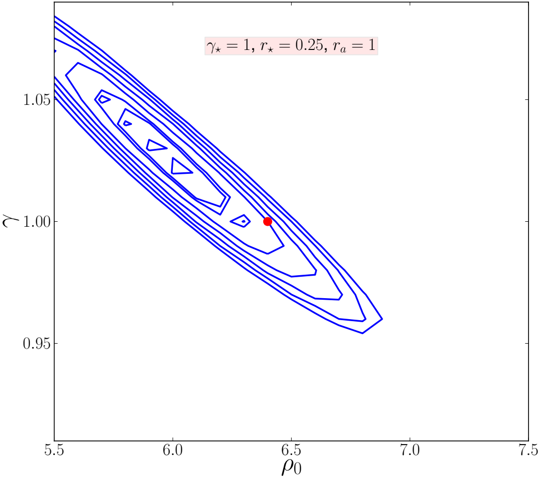
Results: various stellar DFs in NFW potential



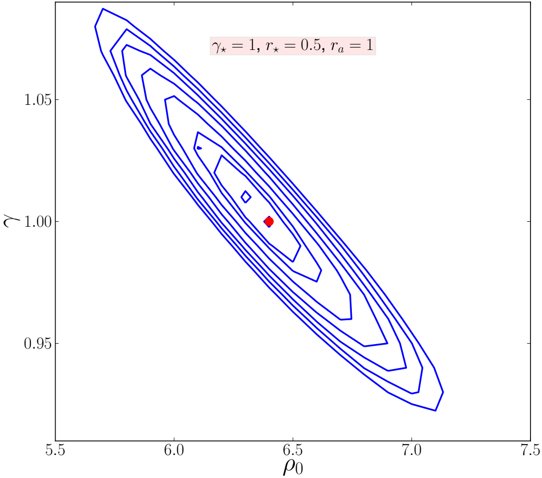
Results: various stellar DFs in NFW potential



Results: various stellar DFs in NFW potential



Results: various stellar DFs in NFW potential



Summary

DPM is sensible[1] way of representing prior on DFs:

- Models the DF “nonparametrically” (in true sense)
  - could view as unholy marriage of ∞-resolution Schwarzschild method in which δ-fn orbits replaced by parametrised DFs
  - (but that might not be helpful)

- Gives broadly correct results for $\mathrm{p}(D|\rho_0, \gamma)$
  - exact calculation possible only for $N \leq 10$
  - approximate VB fast, but still room for improvement
  - beguiling interpretation in terms of streams.

- Straightforward **in principle** to include in likelihood
  - measurement errors (e.g., missing radial vels)
  - survey selection functions.

- Finding an effective practical method to do this is challenging.

---

[1] Worth thinking about: do we *really* want to marginalise?