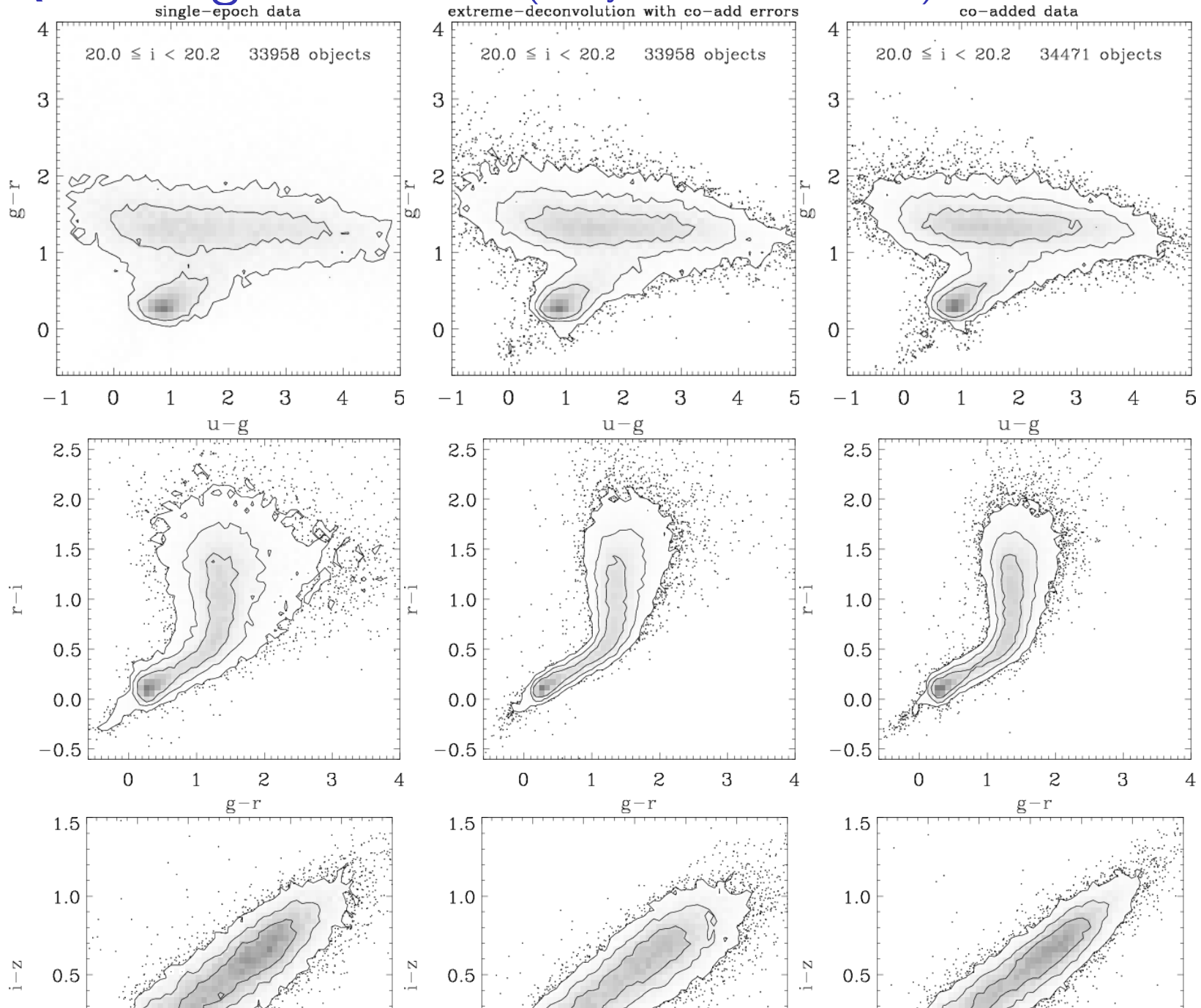
Hierarchical modeling

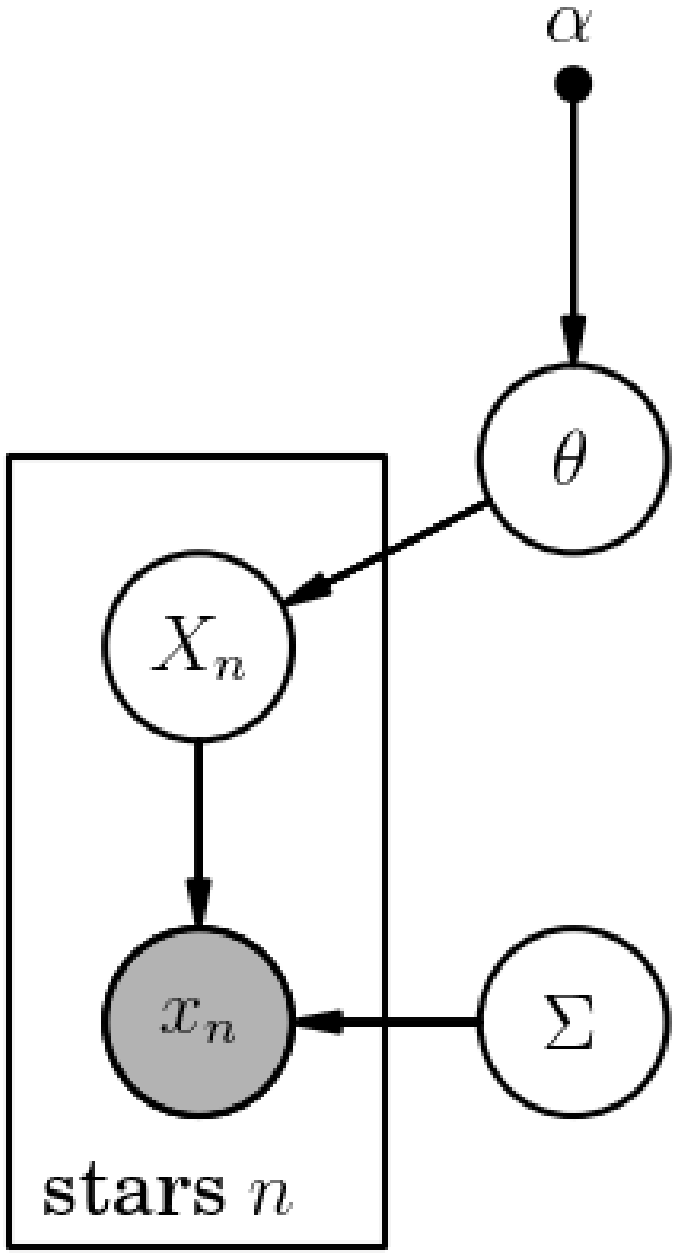
David W. Hogg
Center for Cosmology and Particle Physics,
New York University
and
Max-Planck-Institut für Astronomie, Heidelberg

2013 July 26

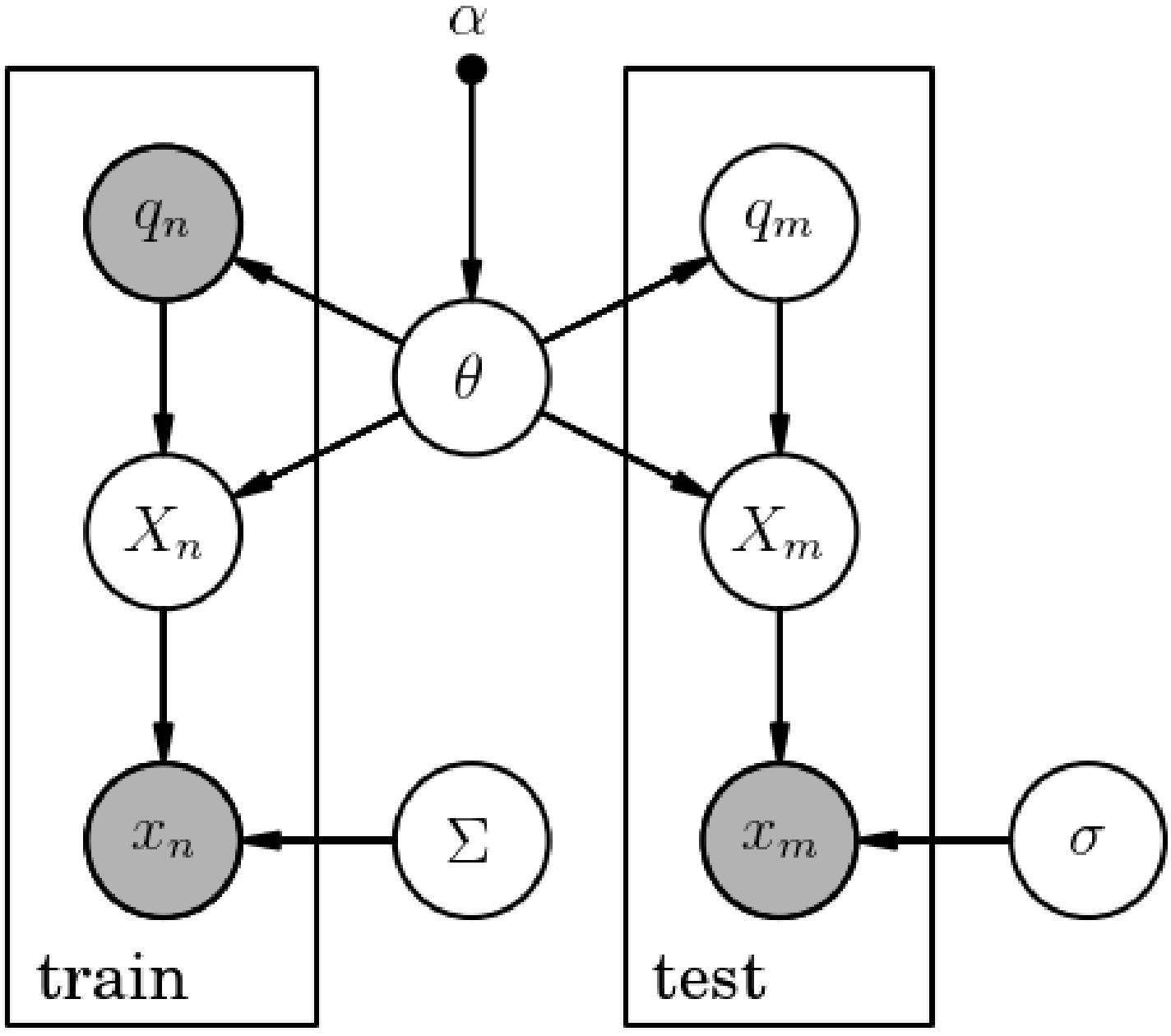
XDQSO target selection (Bovy *et al.*, 1011.6392)



XDQSO target selection (Bovy *et al.*, 1011.6392)



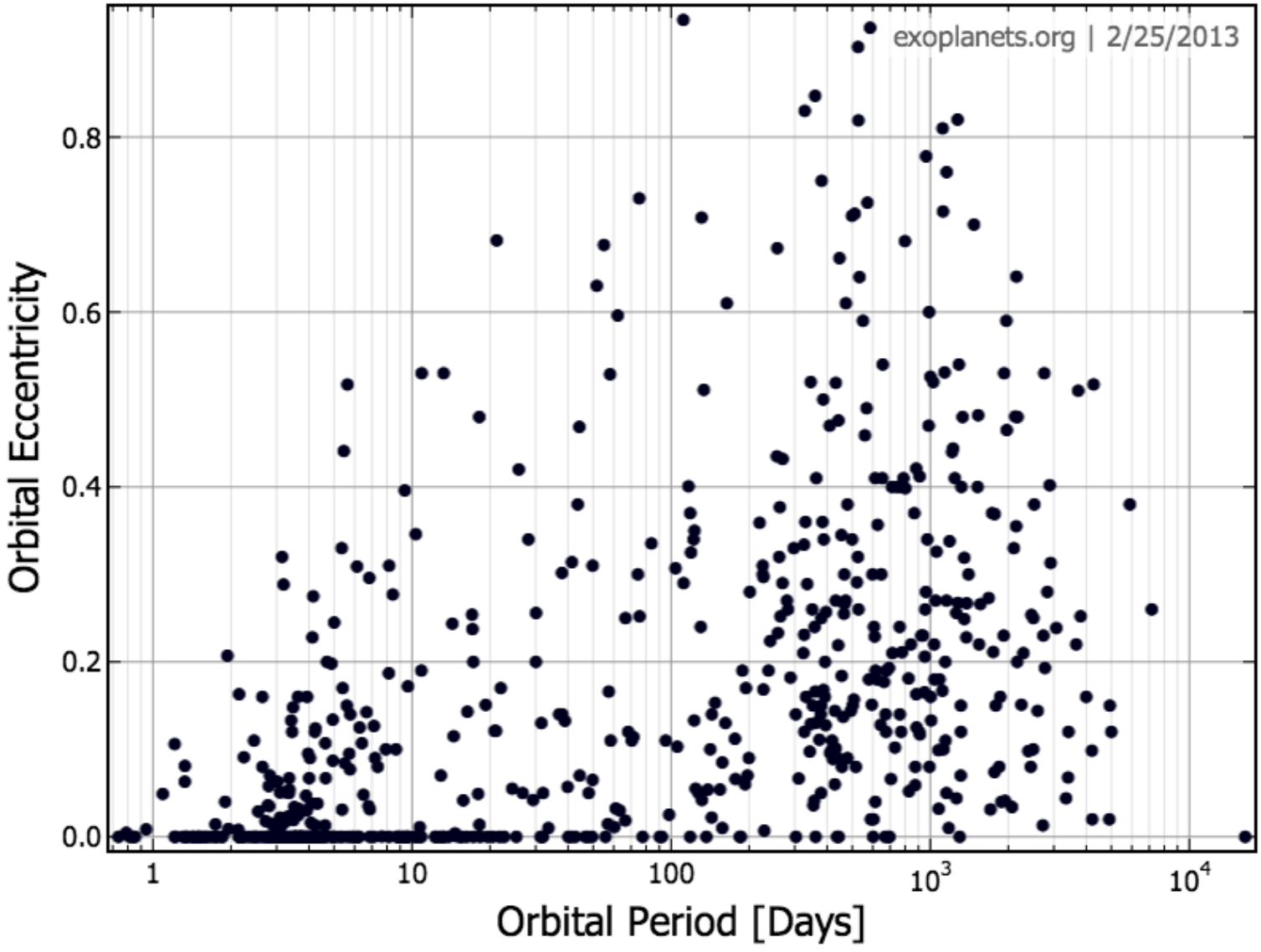
XDQSO target selection (Bovy *et al.*, 1011.6392)

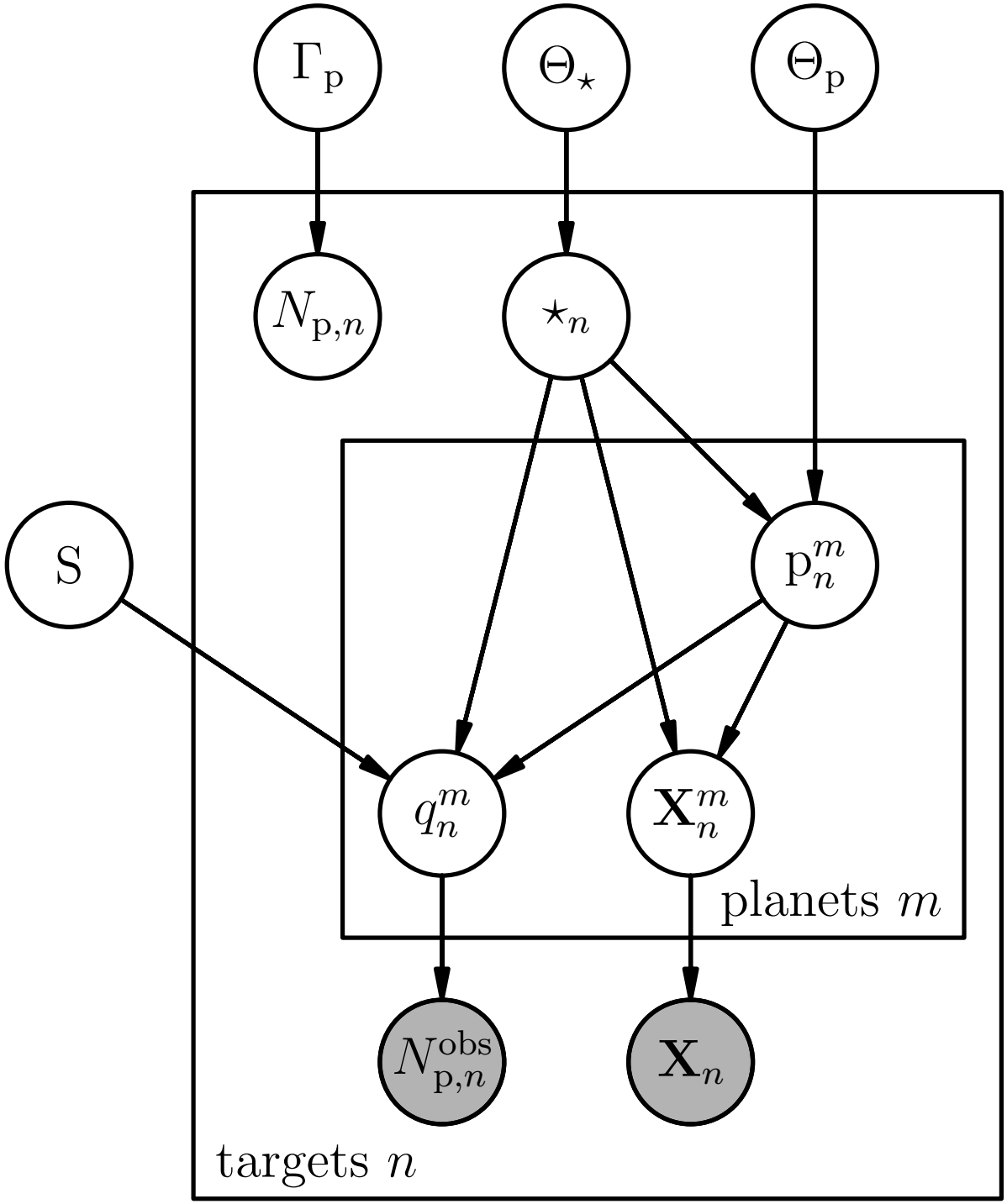


collaborators

- Jo Bovy (IAS)
- Brendon Brewer (Auckland)
- Rob Fergus (NYU)
- **Dan Foreman-Mackey** (NYU)
- Jonathan Goodman (NYU)
- Dustin Lang (CMU)

eccentricities

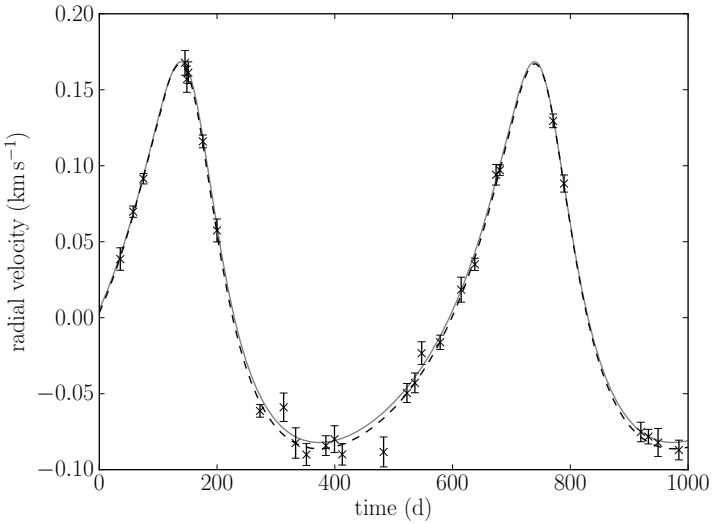


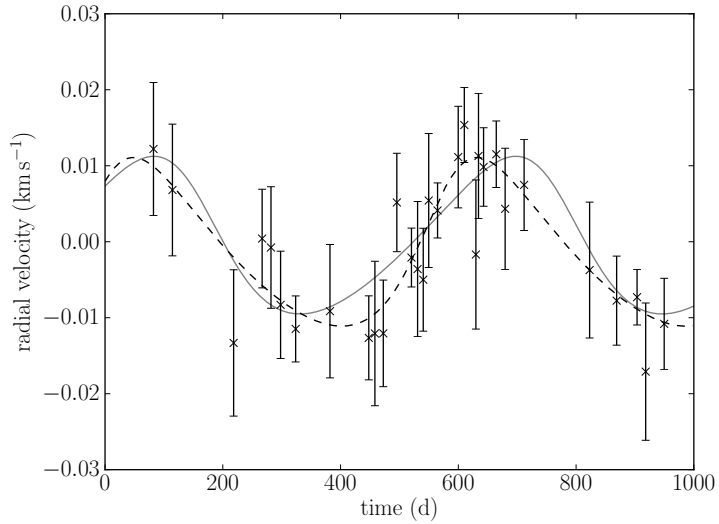

Γ_{p}
$\Theta_{\star}$
Θ_{p}
$N_{\mathrm{p},n}$
$\star_n$
S
p_n^m
q_n^m
$\mathbf{X}_n^m$
planets m
$N_{\mathrm{p},n}^{\mathrm{obs}}$
$\mathbf{X}_n$
targets n

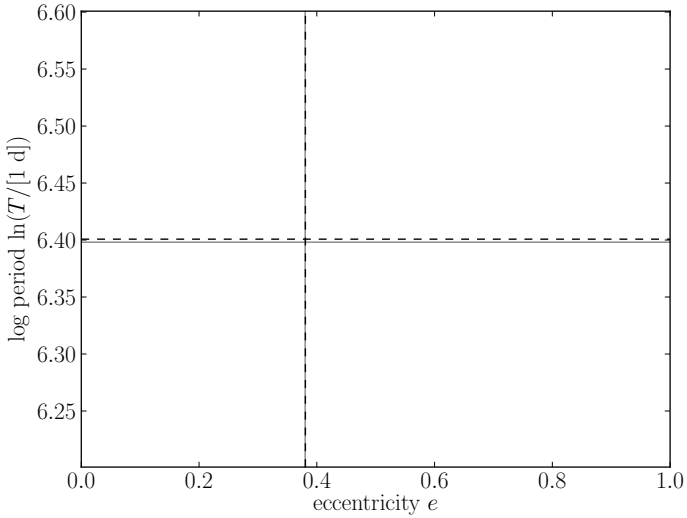
eccentricity inference, usual story

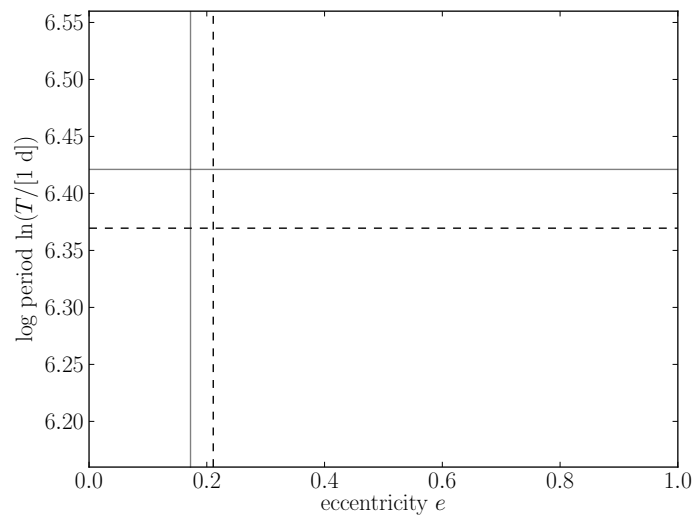
$$
\begin{aligned}
\boldsymbol{\omega}_n &\equiv (\kappa_n, T_n, \phi_n, e_n, \varpi_n) \\
v_{nj} &= V_n + g_{\boldsymbol{\omega}_n}(t_{nj}) + E_{nj} \\
-2 \ln p(\mathbf{D}_n|\boldsymbol{\omega}_n) &= Q + \sum_{j=1}^{M_n} \ln(\sigma_{nj}^2 + S_n^2) + \sum_{j=1}^{M_n} \frac{[V_n + g_{\boldsymbol{\omega}_n}(t_{nj}) - v_{nj}]^2}{\sigma_{nj}^2 + S_n^2} \\
p(\boldsymbol{\omega}_n|\mathbf{D}_n) &= \frac{1}{Z_n}\, p(\mathbf{D}_n|\boldsymbol{\omega}_n)\, p_0(\boldsymbol{\omega}_n) \quad ,
\end{aligned}
$$

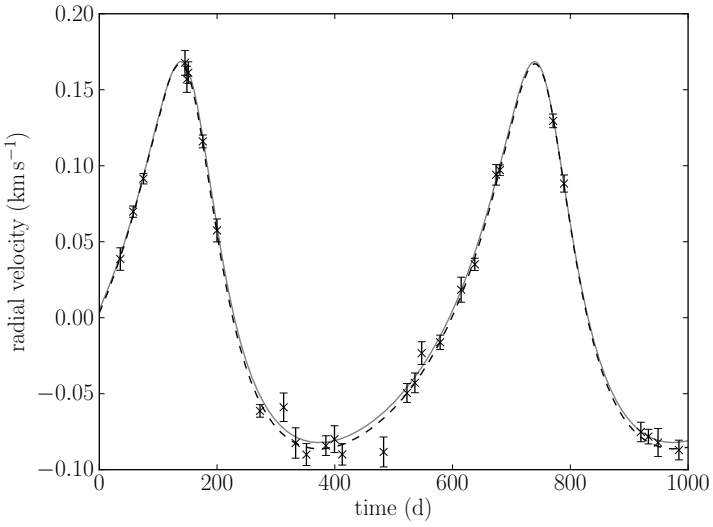
where $p_0(\boldsymbol{\omega}_n)$ is some "uninformative" prior like flat in some parameters, $1/x$ in others.

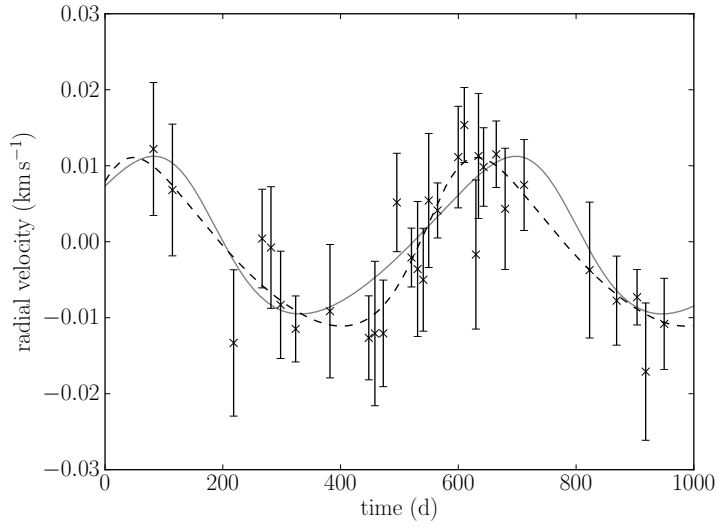

radial velocity (km s^{-1})
time (d)

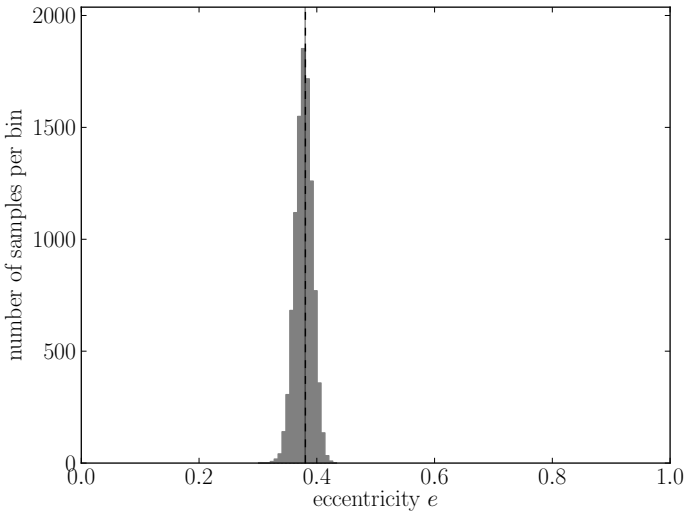

radial velocity (km s^{-1})
time (d)

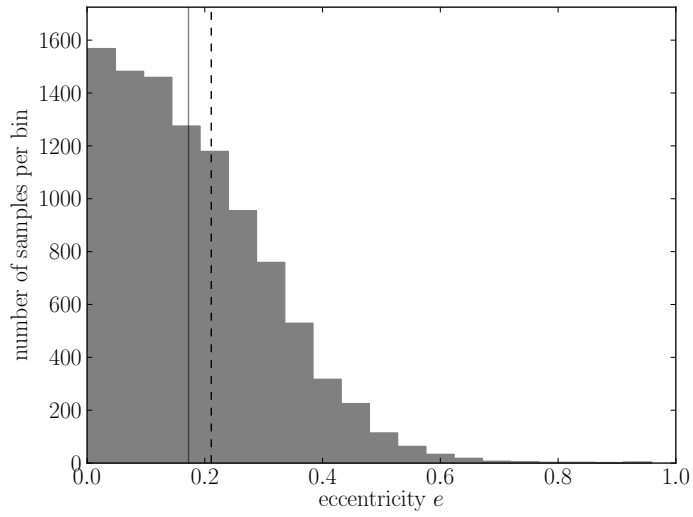

log period ln(T/[1 d])
eccentricity e


log period ln(T/[1 d])
eccentricity e


radial velocity (km s^{-1})
time (d)


radial velocity (km s^{-1})
time (d)


number of samples per bin
eccentricity e


number of samples per bin
eccentricity e

eccentricity distribution inference (1008.4146)

Of course you don't know the priors on exoplanet properties! What if you think there might be some family of priors $p(\omega_n|\alpha)$, parameterized by some α? Could you find the best α?

$$p(\{\mathbf{D}_n\}_{n=1}^N|\alpha) = \prod_{n=1}^N \int \mathrm{d}\omega_n \, p(\mathbf{D}_n|\omega_n) \, p(\omega_n|\alpha) \quad .$$

This is still a likelihood, but we have marginalized out the properties of every exoplanet—these are "nuisance" parameters in this formulation.

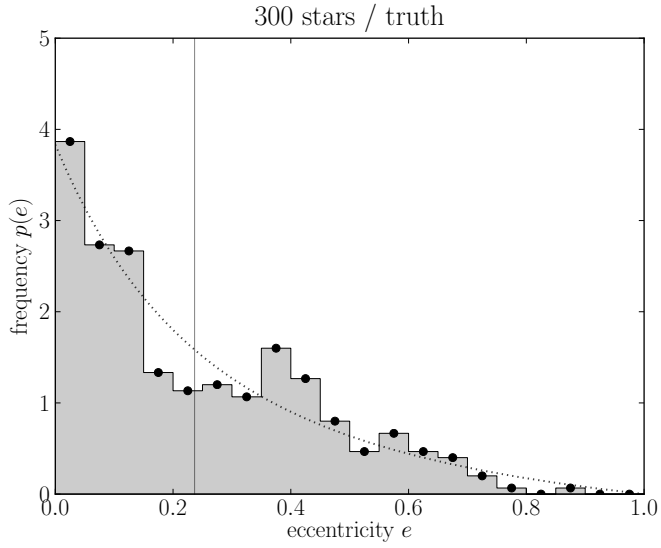
eccentricity distribution inference (1008.4146)

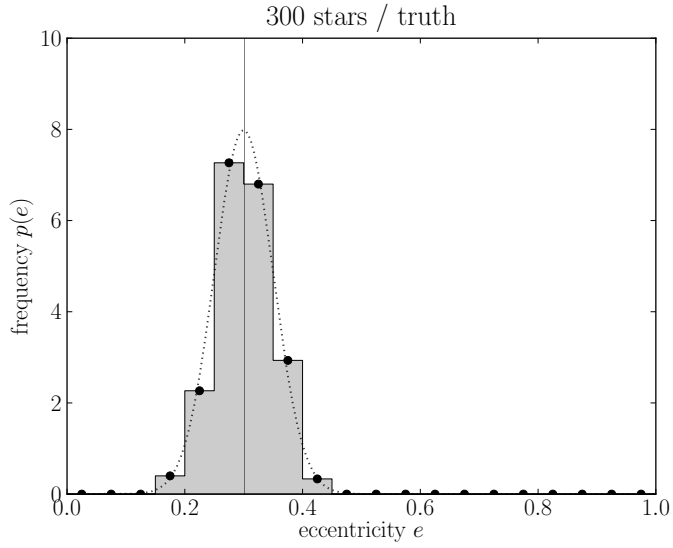
Say all you get, for each exoplanet, are K samples drawn from an uninformative prior. What then? Importance sampling.

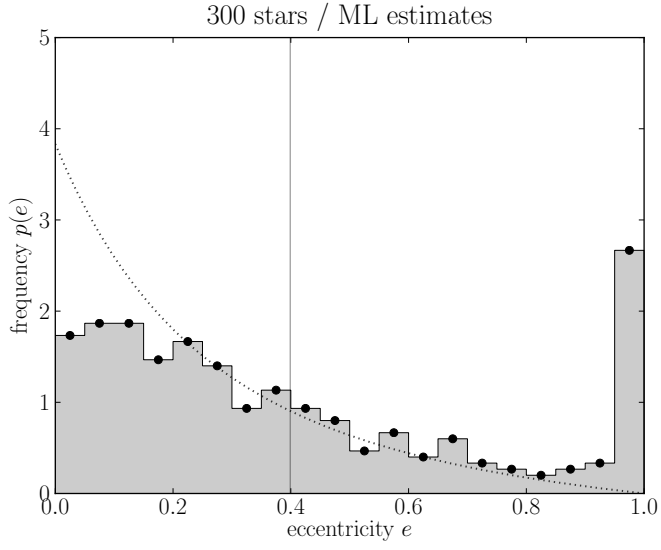
$$
\begin{aligned}
p(\omega_n|\alpha) &\equiv \frac{f_\alpha(e_n)\, p_0(\omega_n)}{p_0(e_n)} \\
\int \mathrm{d}\omega_n\, p_0(\omega_n|\mathbf{D}_n)\, F(\omega_n) &\approx \frac{1}{K}\sum_{k=1}^{K} F(\omega_{nk}) \\
p(\{\mathbf{D}_n\}_{n=1}^{N}|\alpha) &\approx \prod_{n=1}^{N}\frac{1}{K}\sum_{k=1}^{K}\frac{f_\alpha(e_{nk})}{p_0(e_{nk})}
\end{aligned}
$$

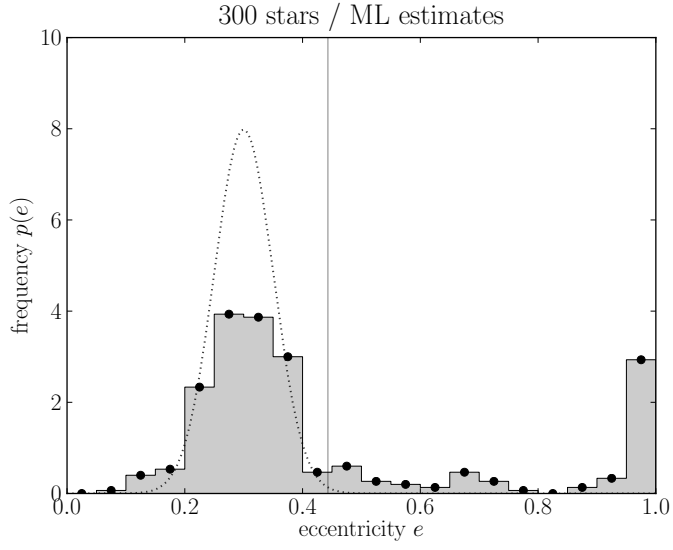
(Concept of an "interim prior".)

distribution inference demo: ML estimates—bad

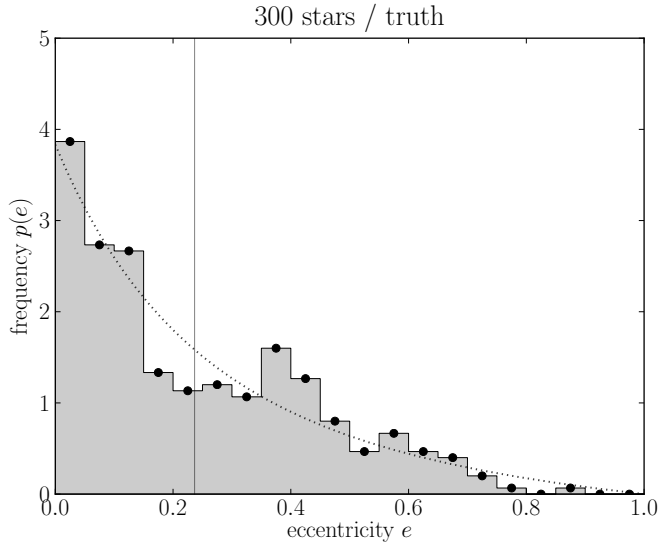


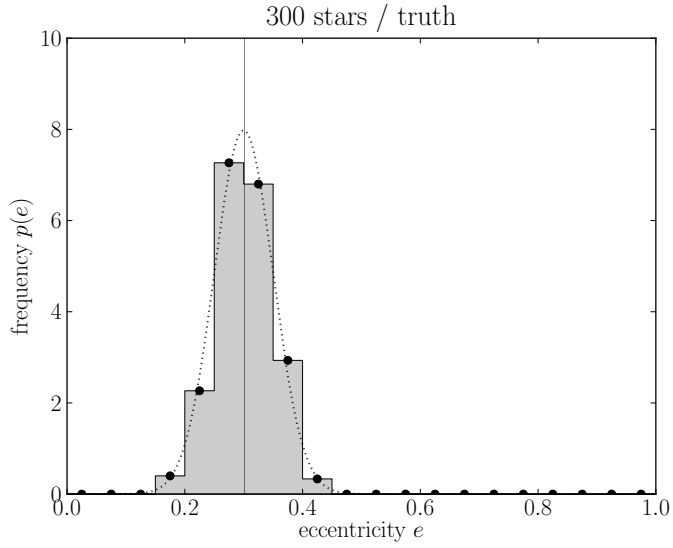


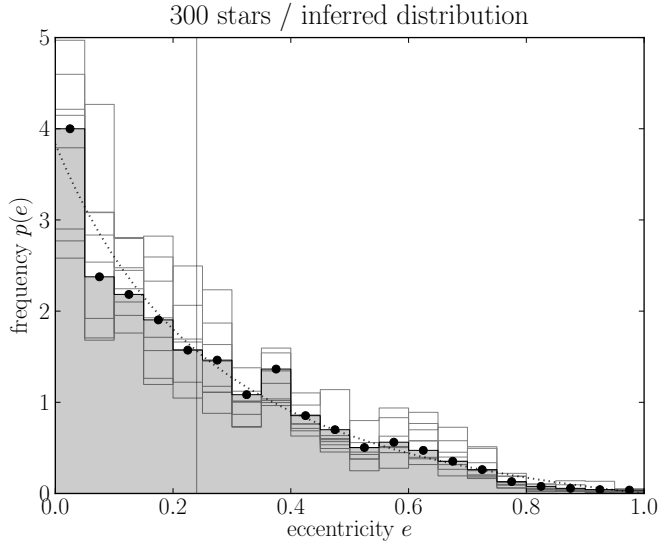


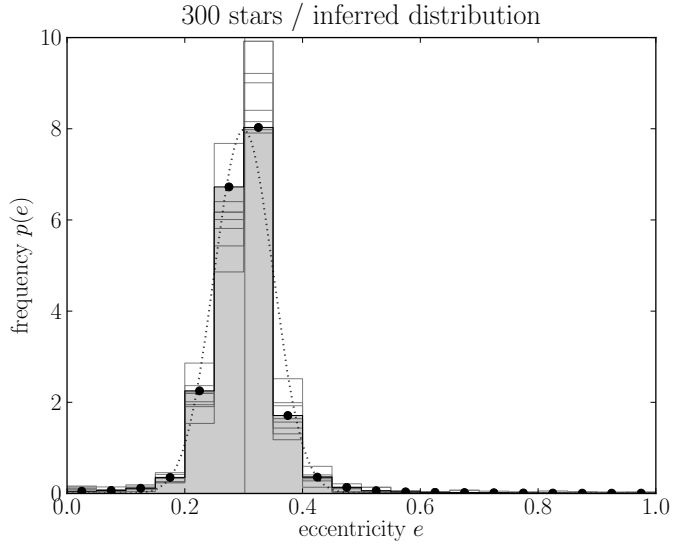


distribution inference demo: Hierarchical inference good!









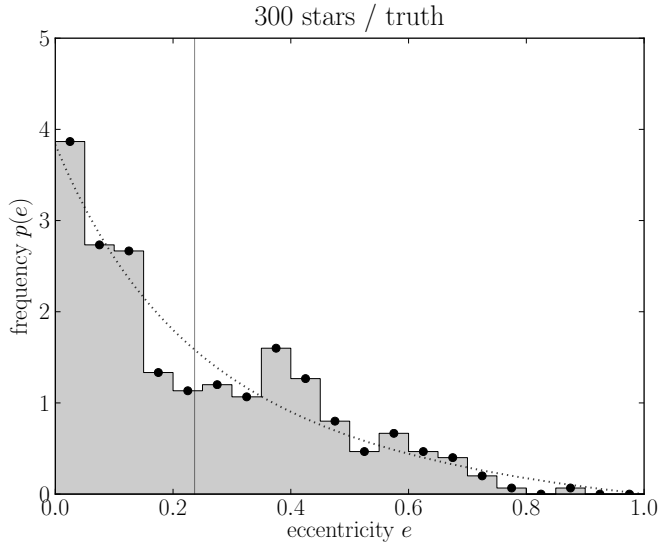
hierarchical population inference: Why does it work?

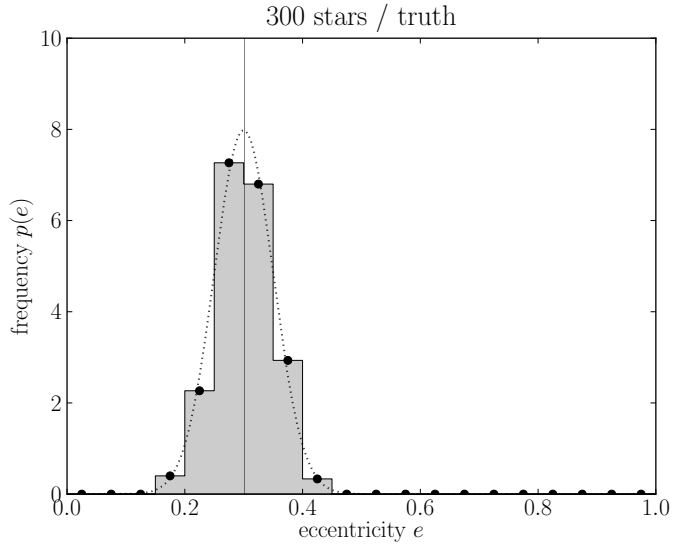
- The marginalized likelihood is large when there is high prior probability in locations where there is high likelihood.
- When likelihoods are broad, the best prior is the most concentrated prior that is "consistent with" **all** individual-object likelihood functions.
- The operation is a **heteroskedastic deconvolution**.
  - (in modern parlance, a "deconvolution" is always the result of fitting a generative or forward model)

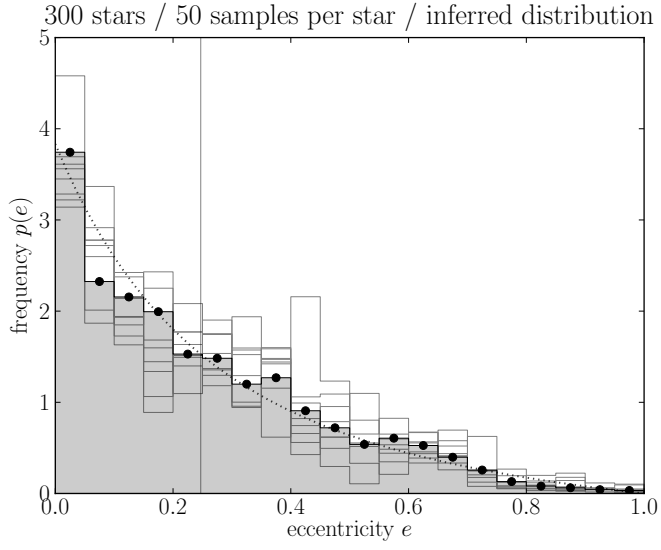
this is deconvolution

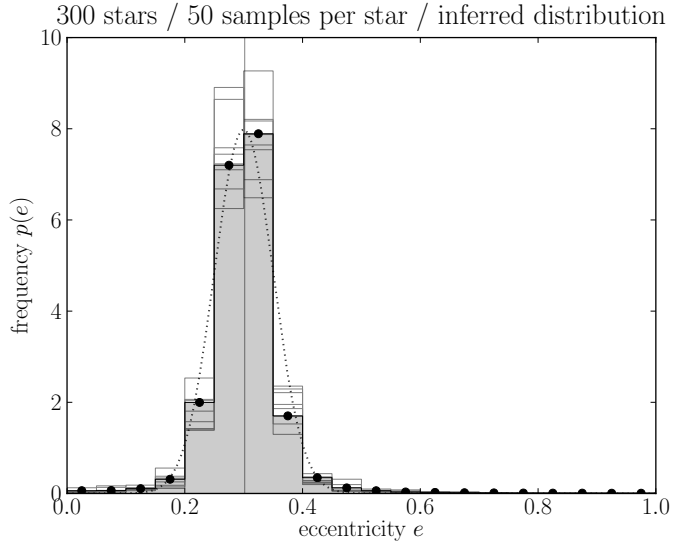
- We can infer the true distribution even with extremely noisy measurements.
- This is an extreme form of **deconvolution**.
  - (but not *Extreme Deconvolution (tm)*)
- Depends crucially on having full—and accurate—likelihood or posterior information.
- Performed by "forward modeling".

distribution inference demo: Small samplings

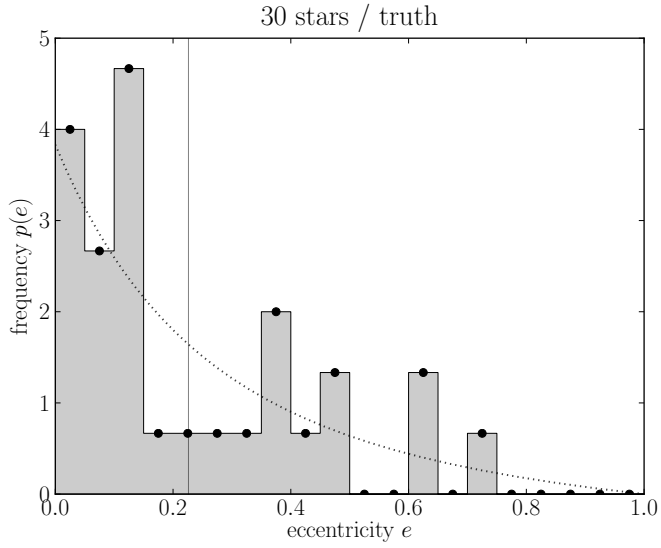


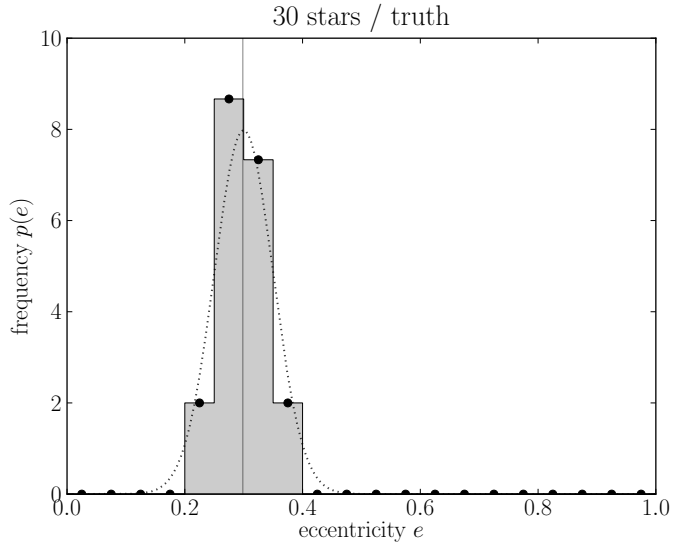


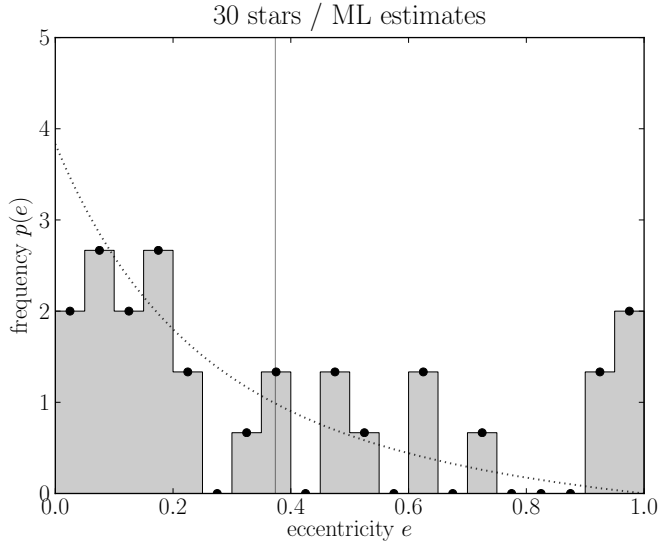


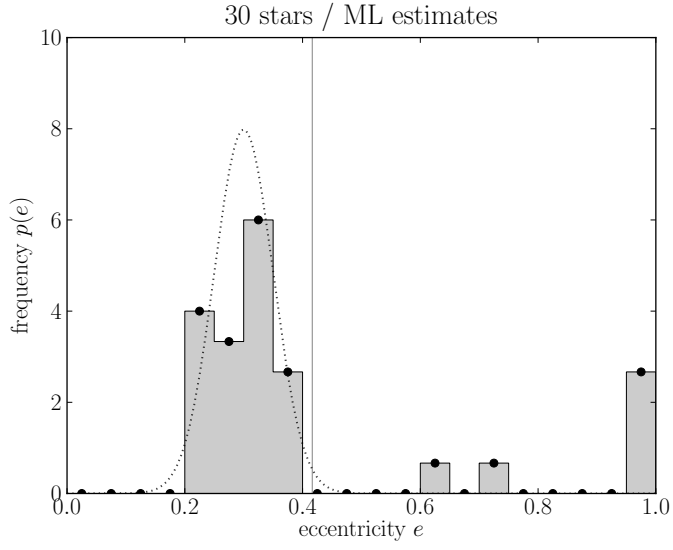


distribution inference demo: Small sample

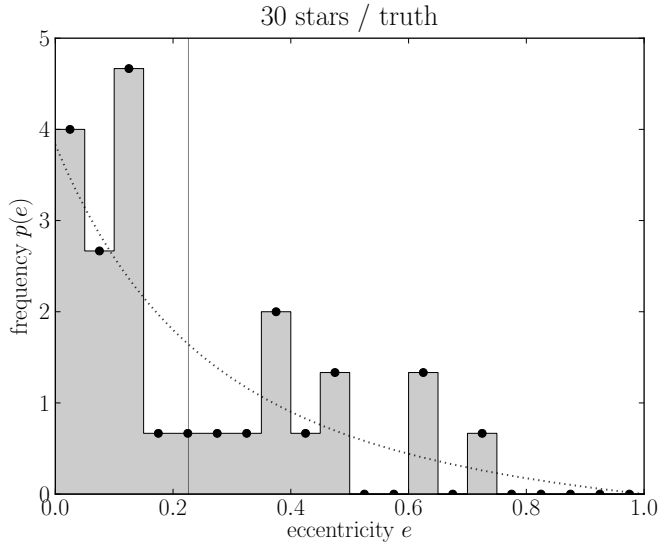


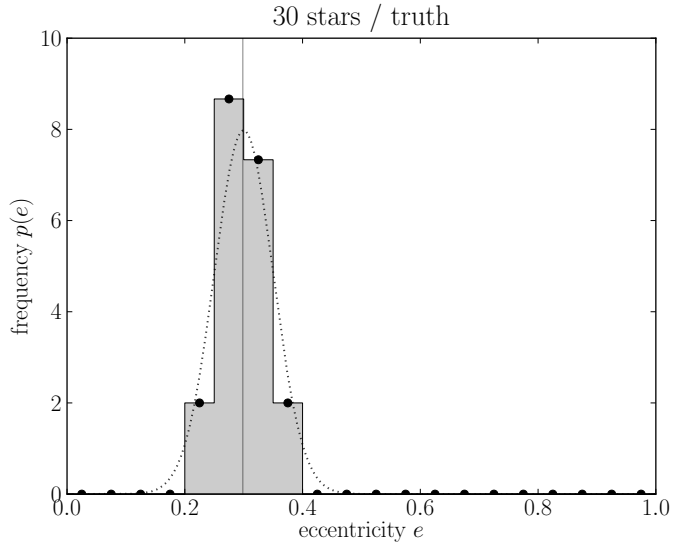


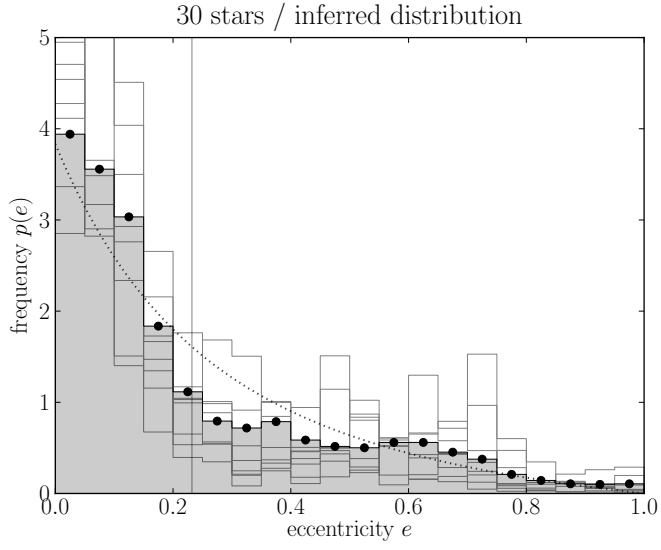


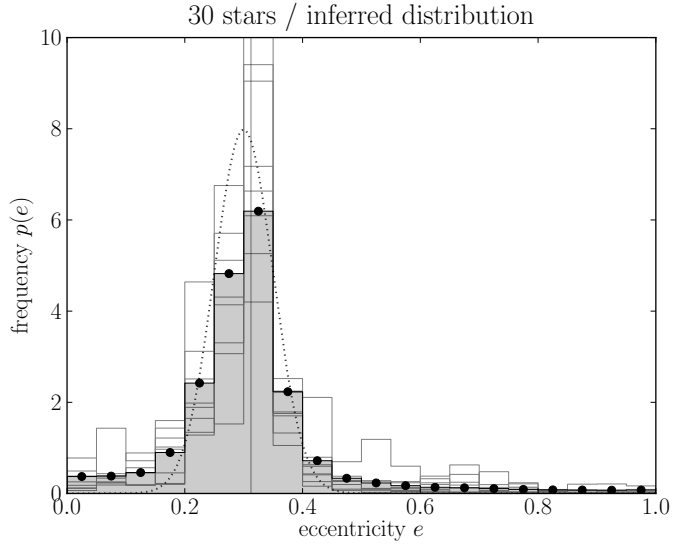


distribution inference demo: Still good!

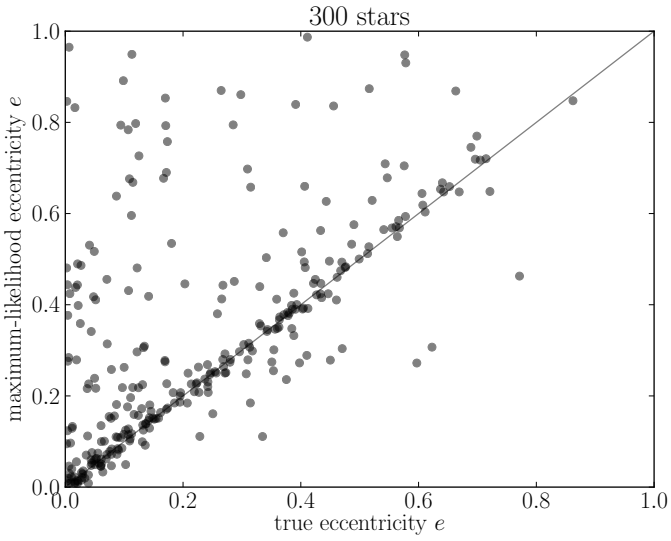


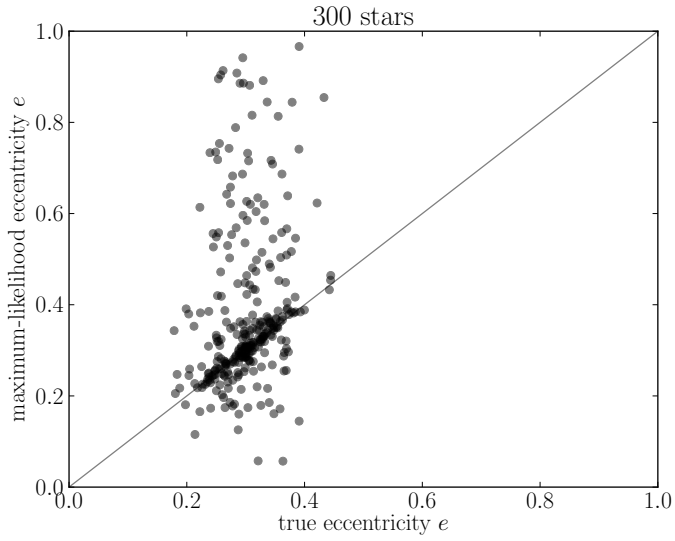


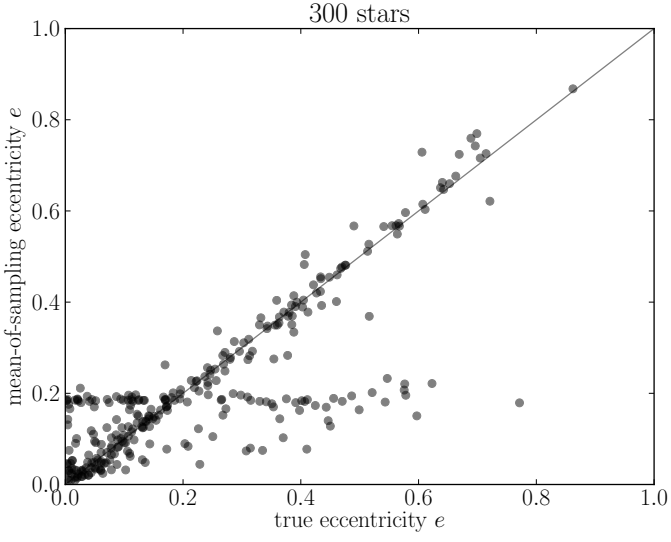


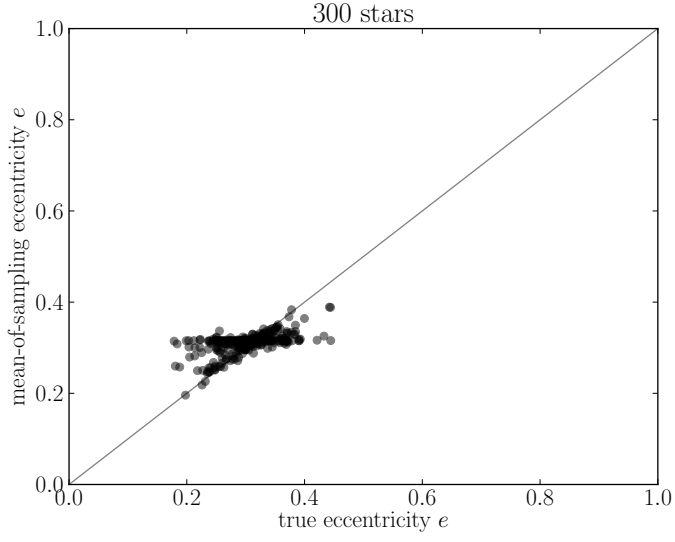


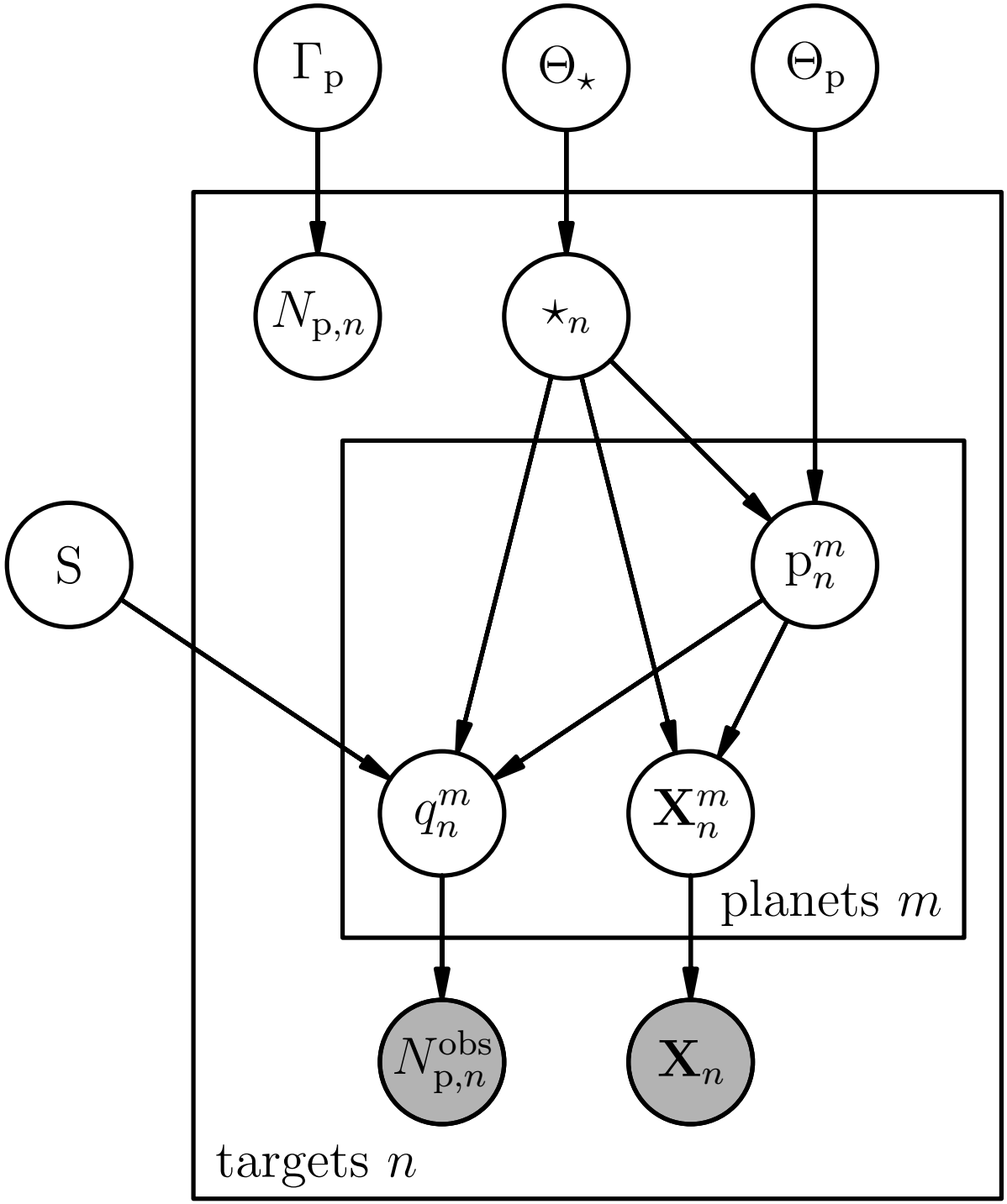
distribution inference demo: Truly hierarchical

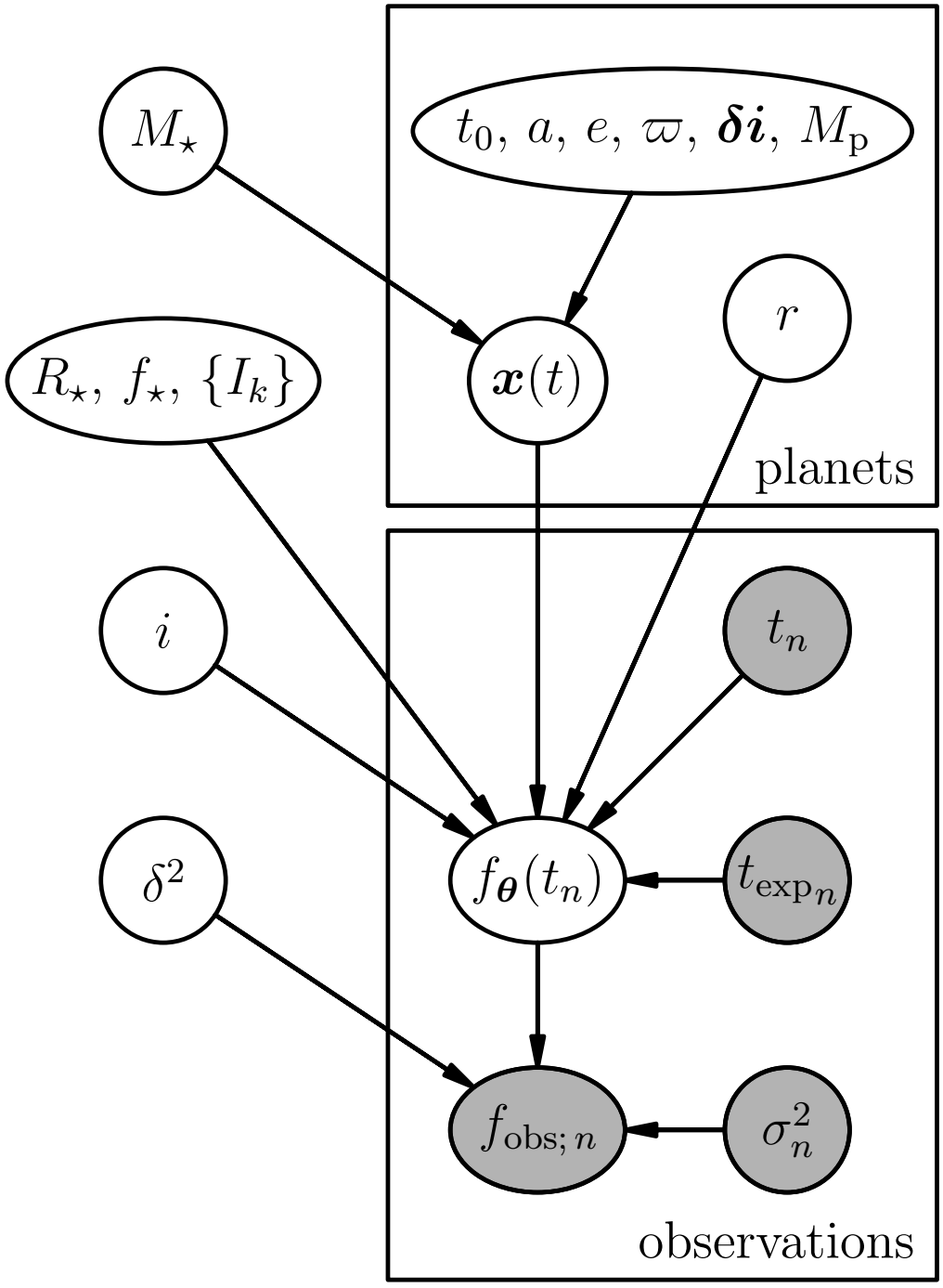


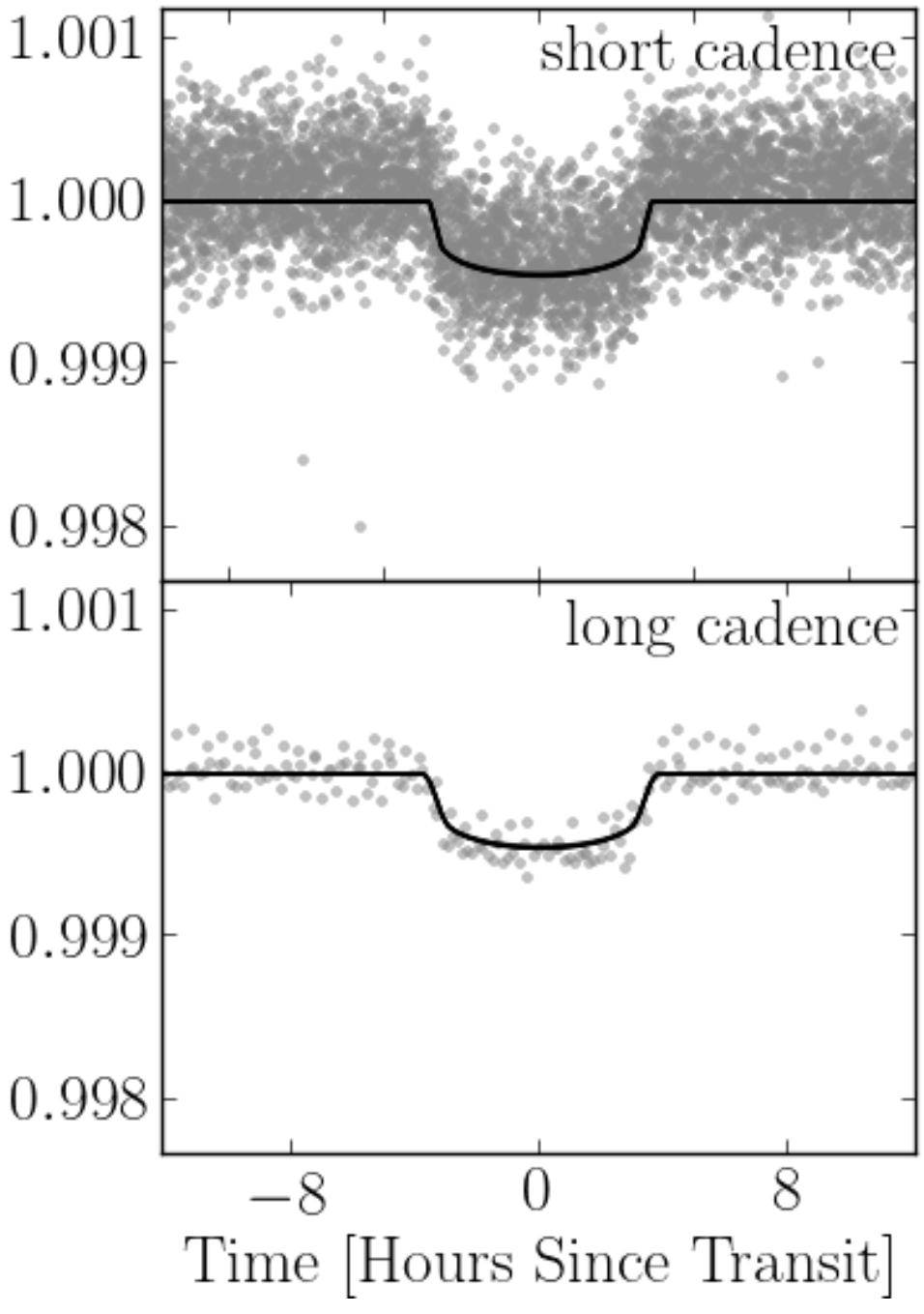


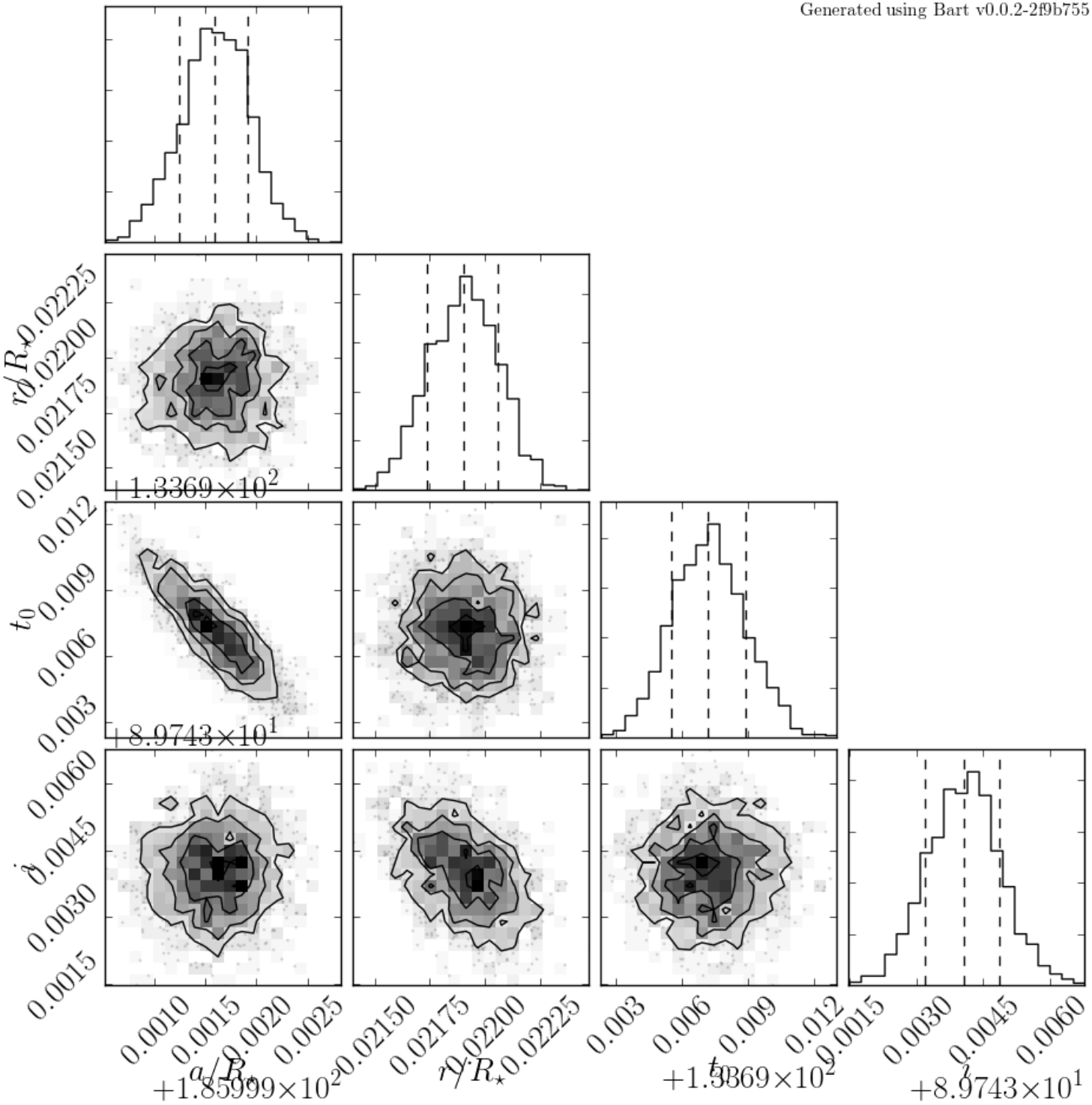





Γ_{p}
$\Theta_{\star}$
Θ_{p}
$N_{\mathrm{p},n}$
$\star_n$
S
p_n^m
q_n^m
$\mathbf{X}_n^m$
planets m
$N_{\mathrm{p},n}^{\mathrm{obs}}$
$\mathbf{X}_n$
targets n


$M_\star$
$t_0,\, a,\, e,\, \varpi,\, \boldsymbol{\delta i},\, M_\mathrm{p}$
r
$R_\star,\, f_\star,\, \{I_k\}$
$\boldsymbol{x}(t)$
planets
i
t_n
δ^2
$f_{\boldsymbol{\theta}}(t_n)$
$t_{\mathrm{exp}\,n}$
$f_{\mathrm{obs};\,n}$
σ_n^2
observations

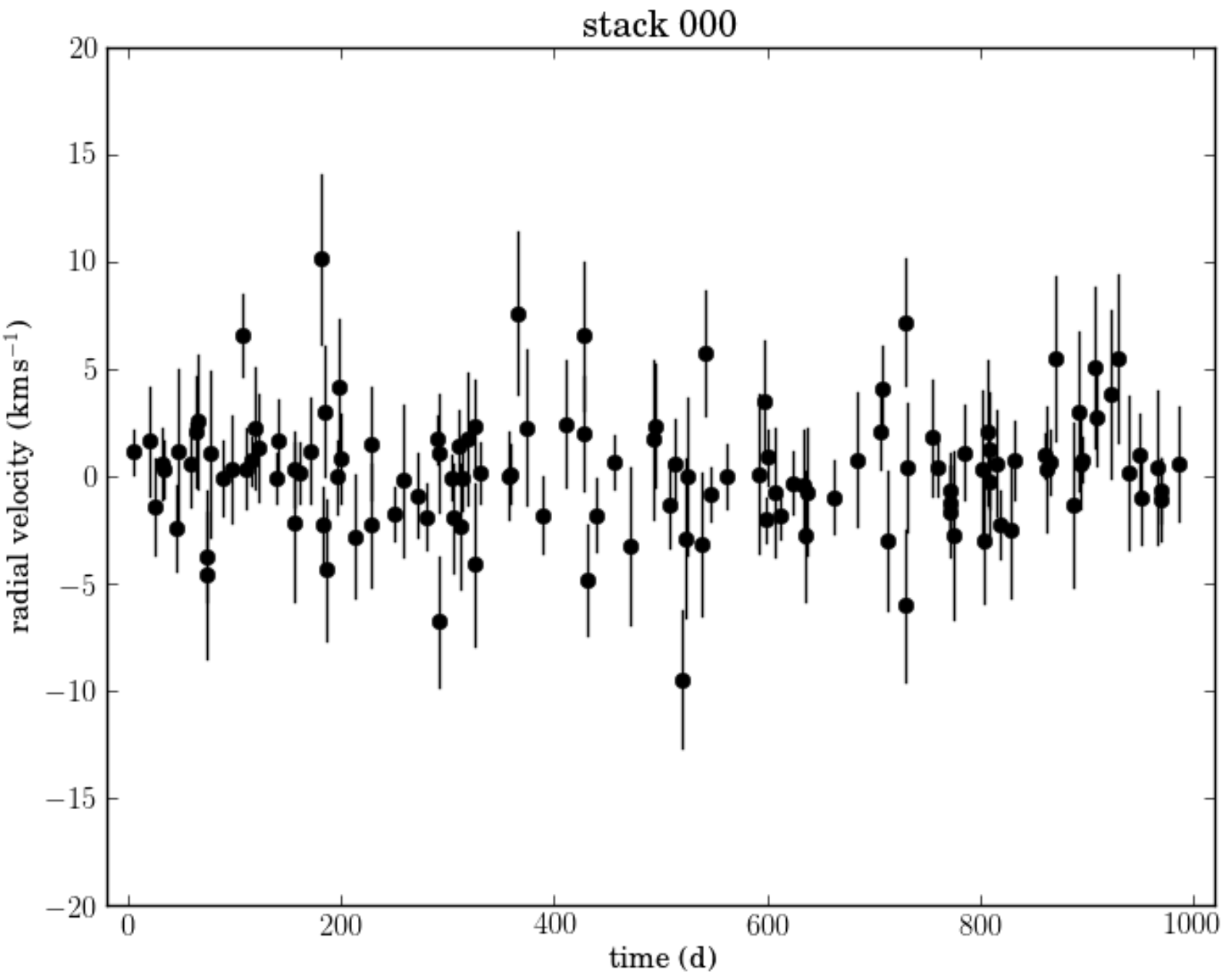

short cadence
long cadence
1.001
1.000
0.999
0.998
−8
0
8
Time [Hours Since Transit]

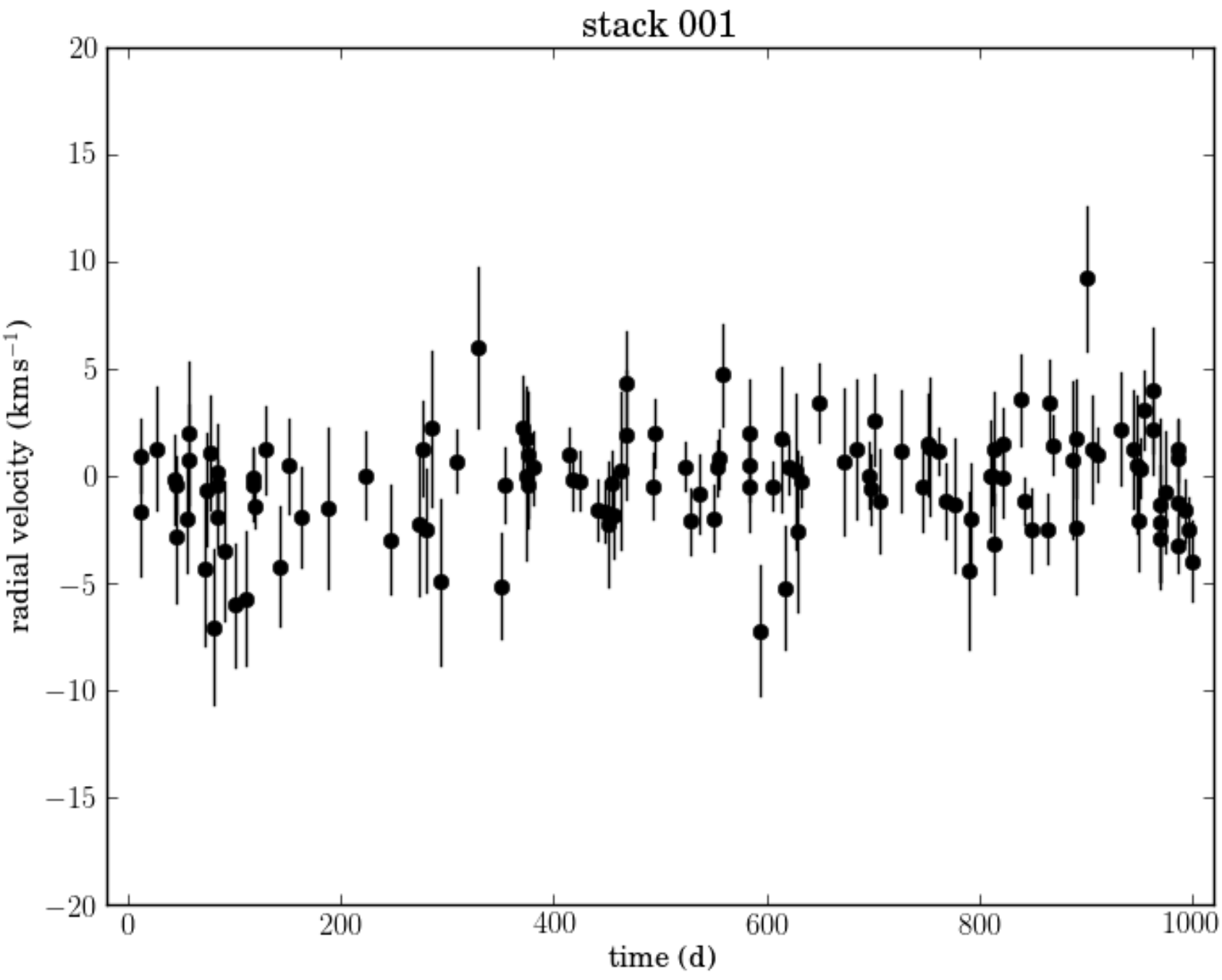

$a/R_\star$
$+1.85999\times10^{2}$
$r/R_\star$
t_0
$+1.3369\times10^{2}$
i
$+8.9743\times10^{1}$

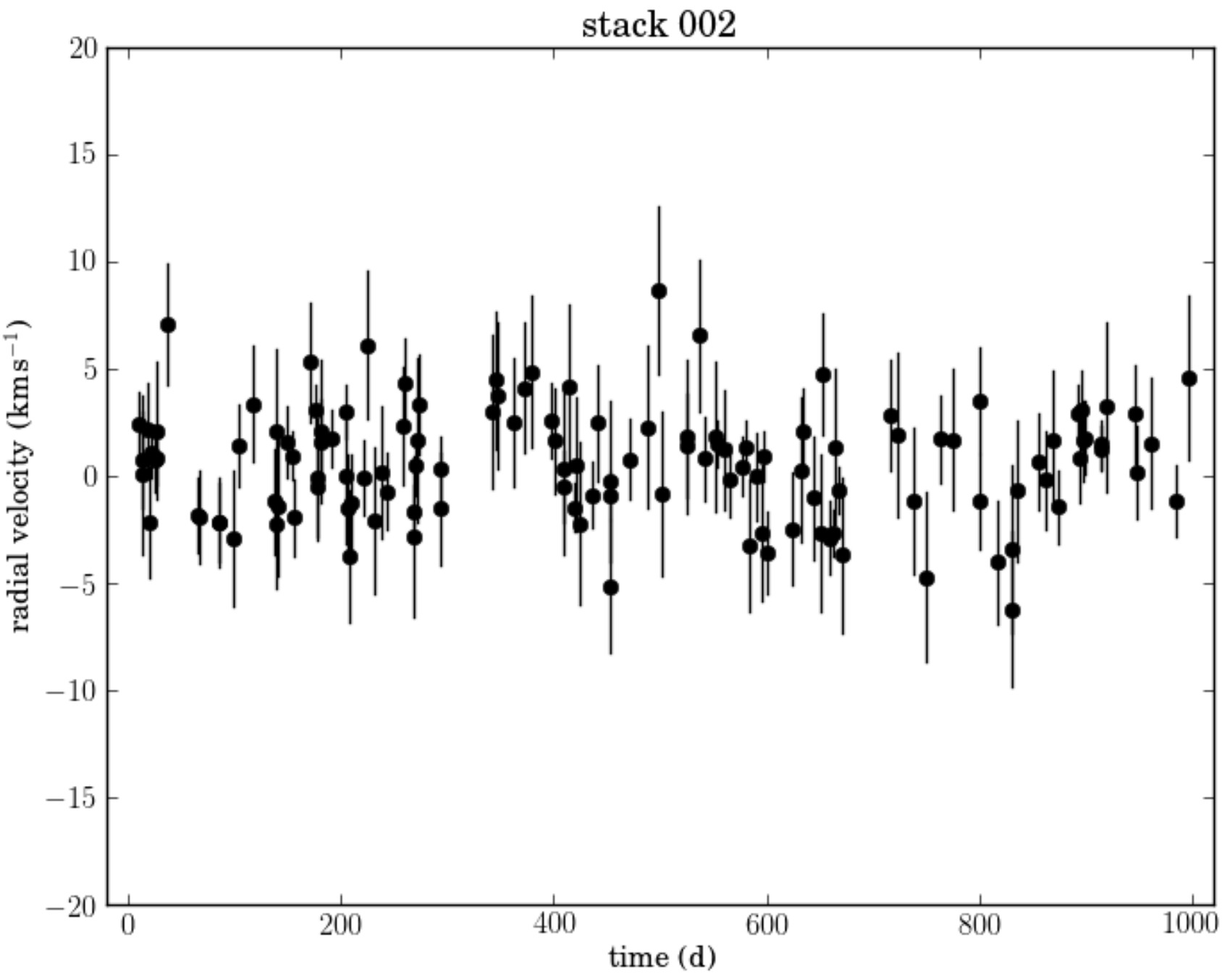
Bart (Foreman-Mackey *et al.*, forthcoming)

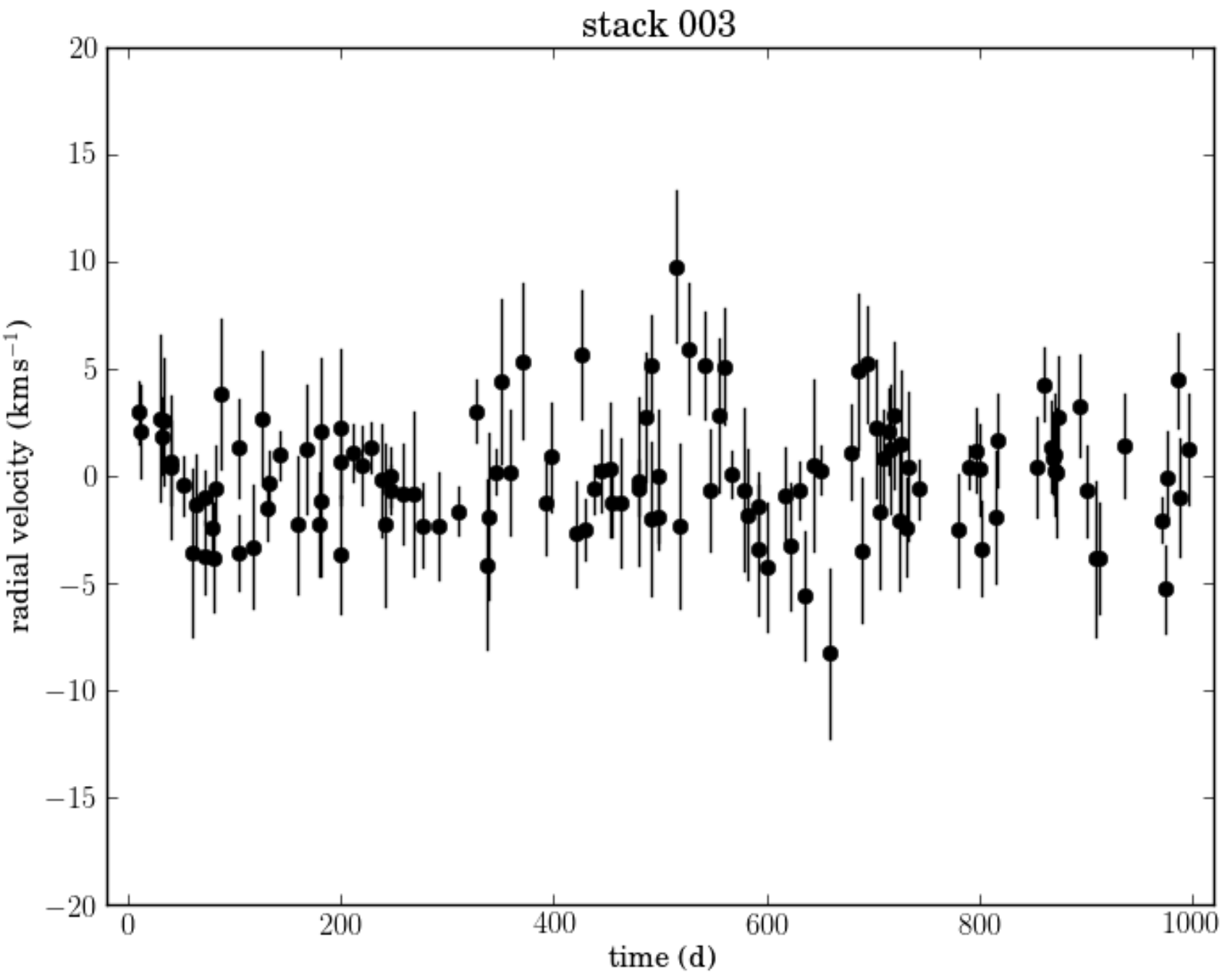
- built on very successful *emcee* package (Foreman-Mackey *et al.*, 1202.3665)
- designed for exoplanet measurement and discovery of false positives
- Gaussian Process models for stellar variability

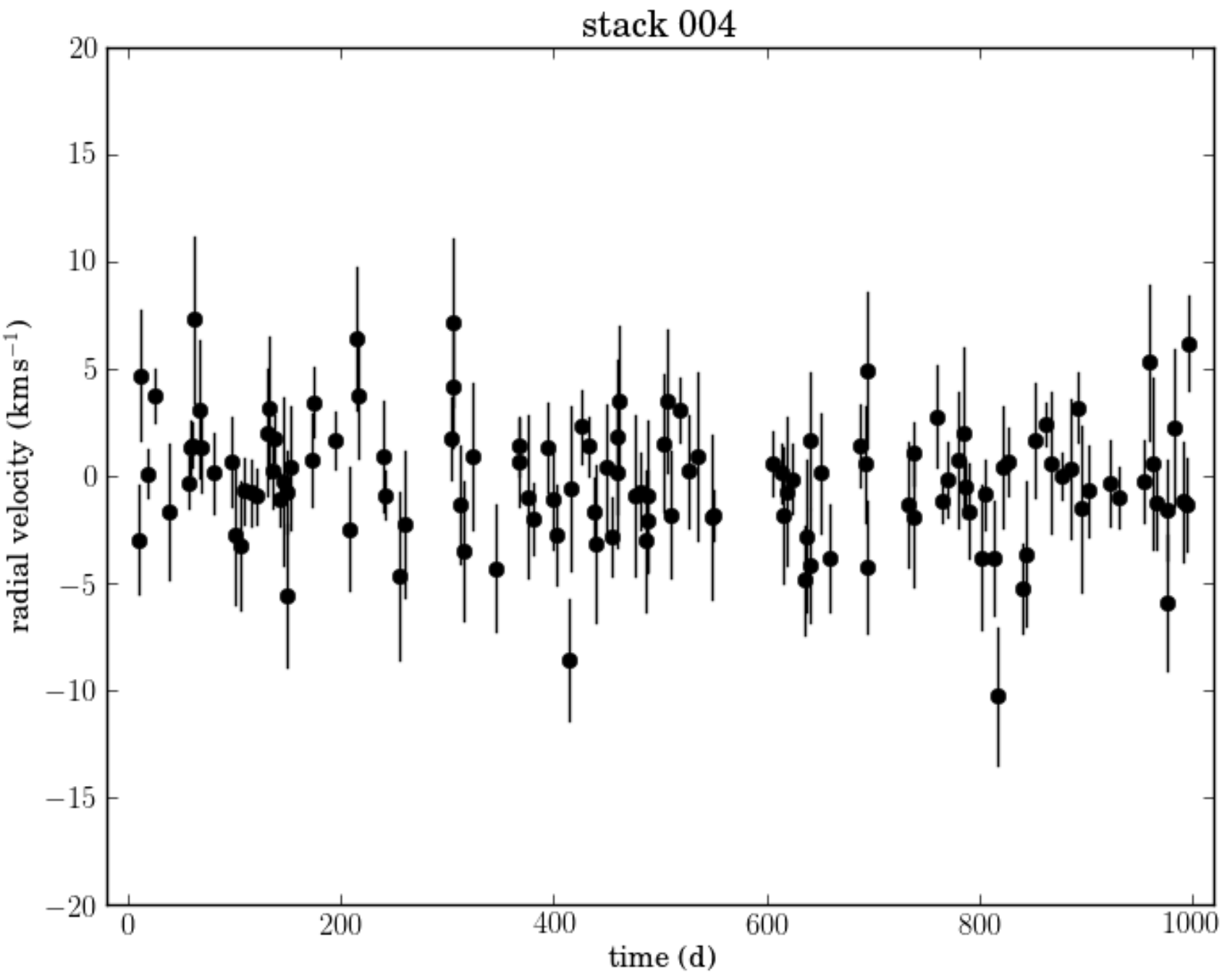
```
import bart

# Initialize a planet.
planet = bart.Planet(r=0.01, a=21.3, t0=3.85)
planet.parameters += [bart.parameters.Parameter(r"$r$", "r"),
                      bart.parameters.LogParameter(r"$a$", "a")]

# Initialize the star.
ldp = bart.kepler.fiducial_ldp(teff=6438, logg=4.28, feh=0.0)
star = bart.Star(mass=planet.get_mstar(12.4138), ldp=ldp)

# Set up the system.
system = bart.PlanetarySystem(star)
system.parameters.append(bart.parameters.CosParameter(r"$i$", "iobs"))
system.add_planet(planet)

# Add data and fit.
system.add_dataset(bart.KeplerDataset("path/to/kepler/data/lc.fits"))
system.fit(2000)
```

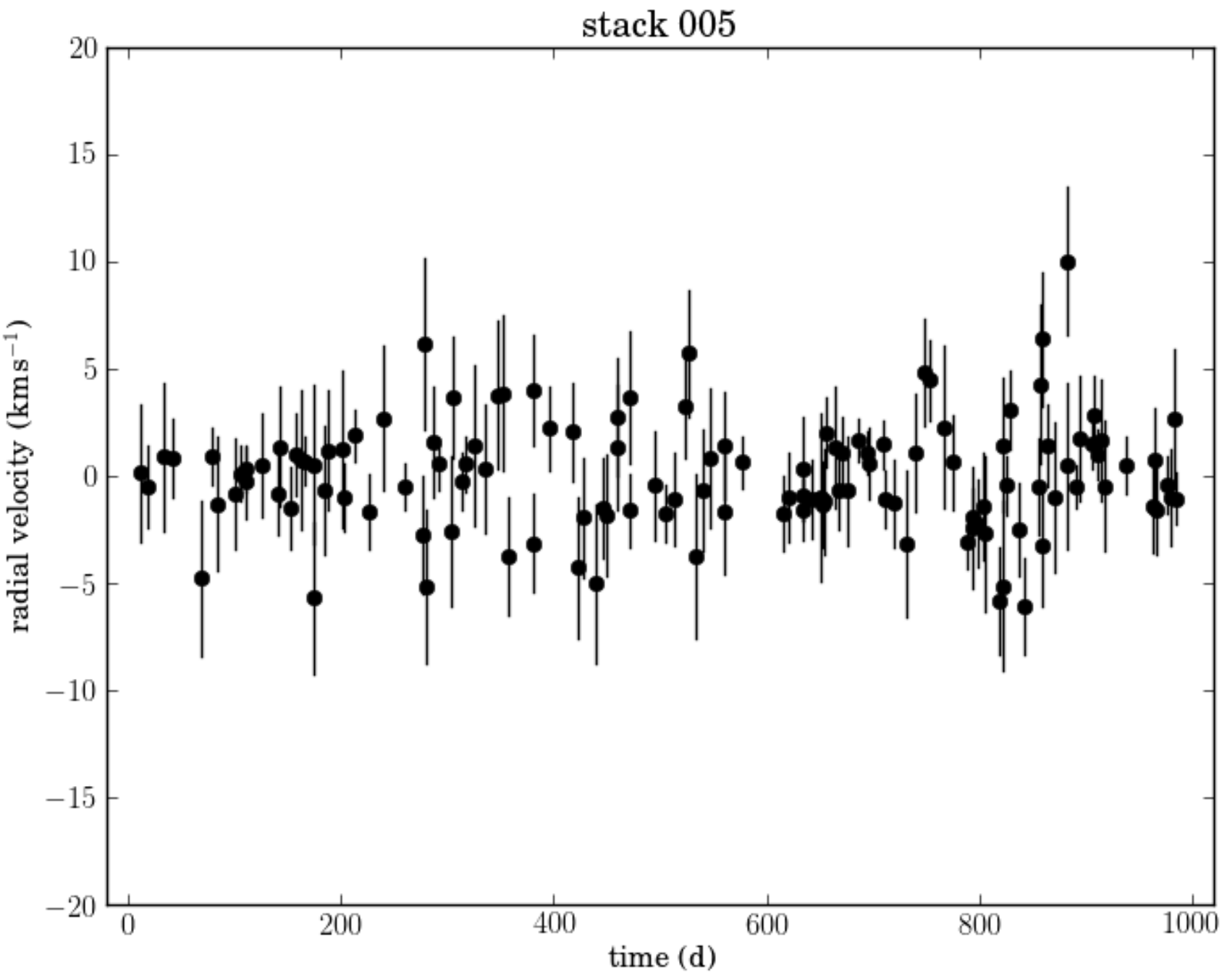

stack 000
radial velocity (kms^{-1})
time (d)
20
15
10
5
0
−5
−10
−15
−20
0
200
400
600
800
1000

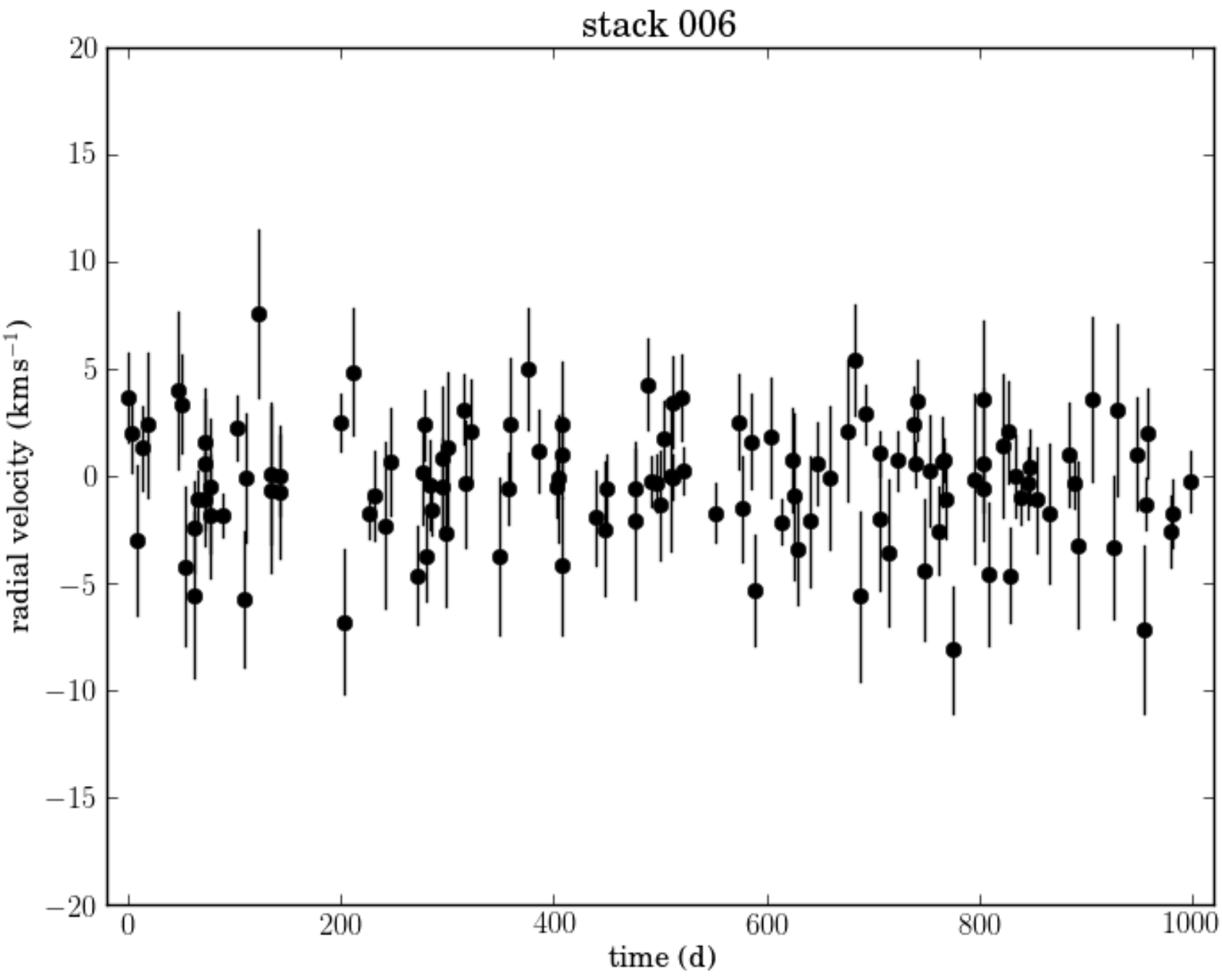

stack 001
20
15
10
5
0
−5
−10
−15
−20
0
200
400
600
800
1000
radial velocity (kms^{-1})
time (d)

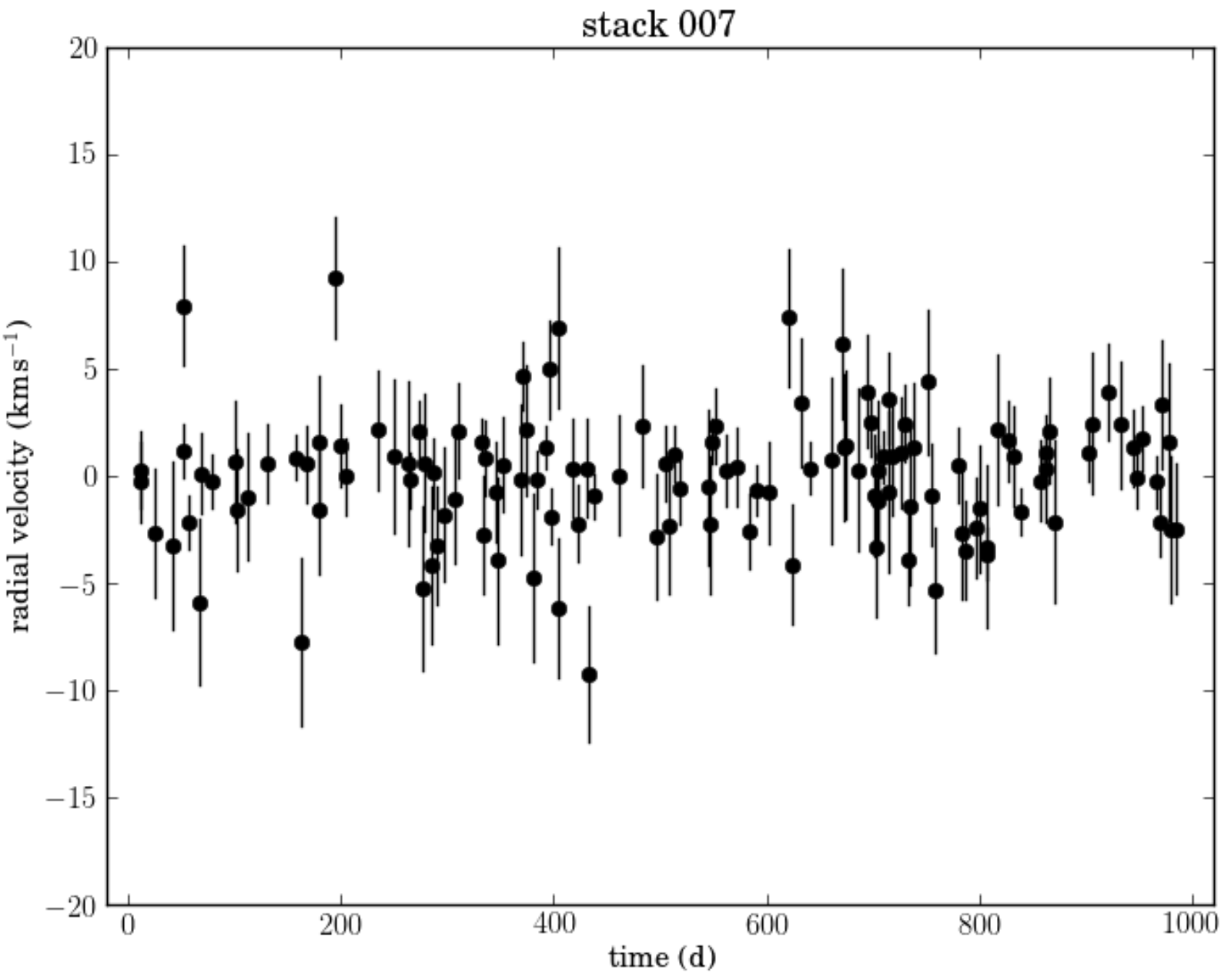

stack 002
radial velocity (kms^{-1})
time (d)
20
15
10
5
0
−5
−10
−15
−20
0
200
400
600
800
1000

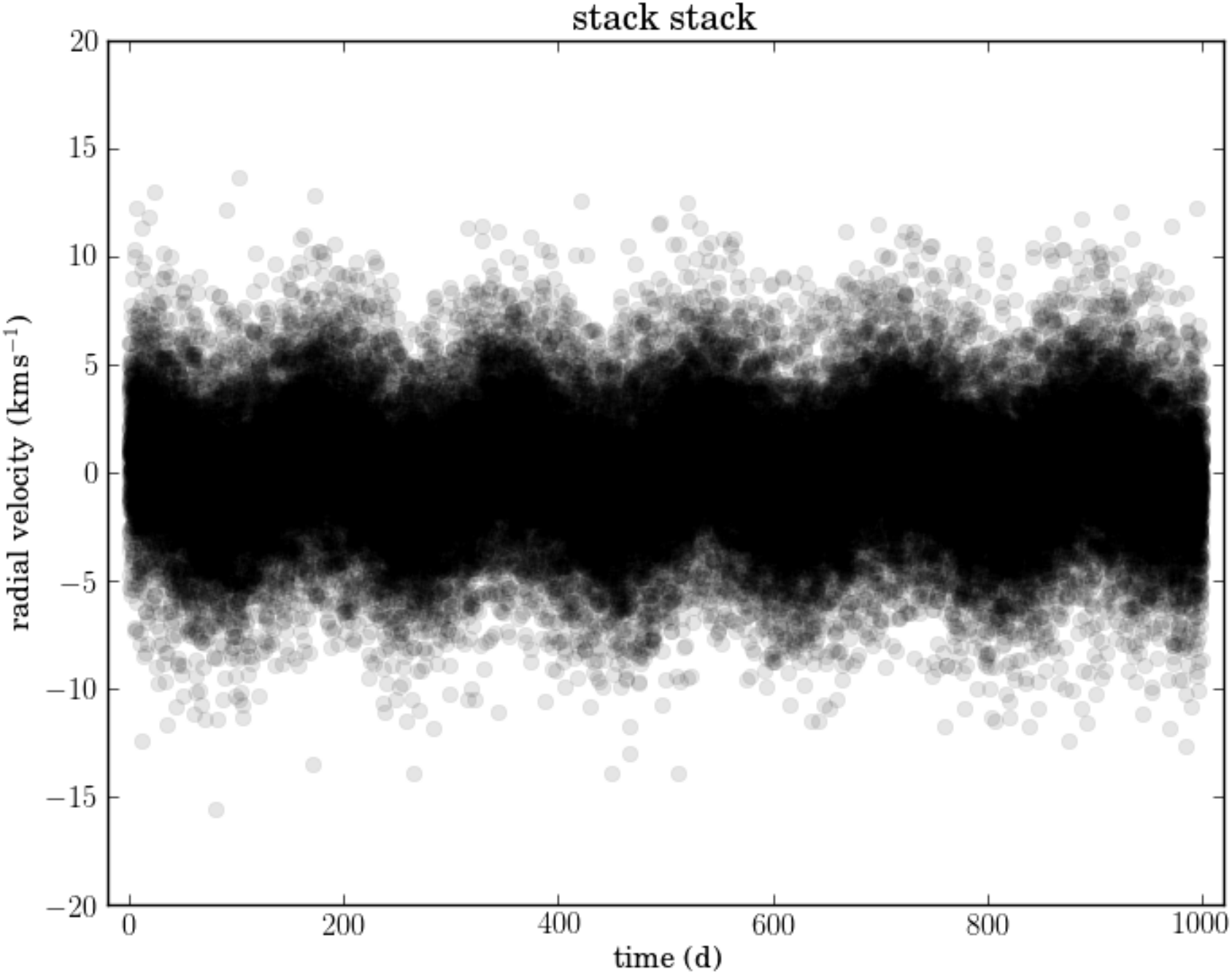

stack 003
20
15
10
5
0
−5
−10
−15
−20
0
200
400
600
800
1000
radial velocity (kms^{-1})
time (d)

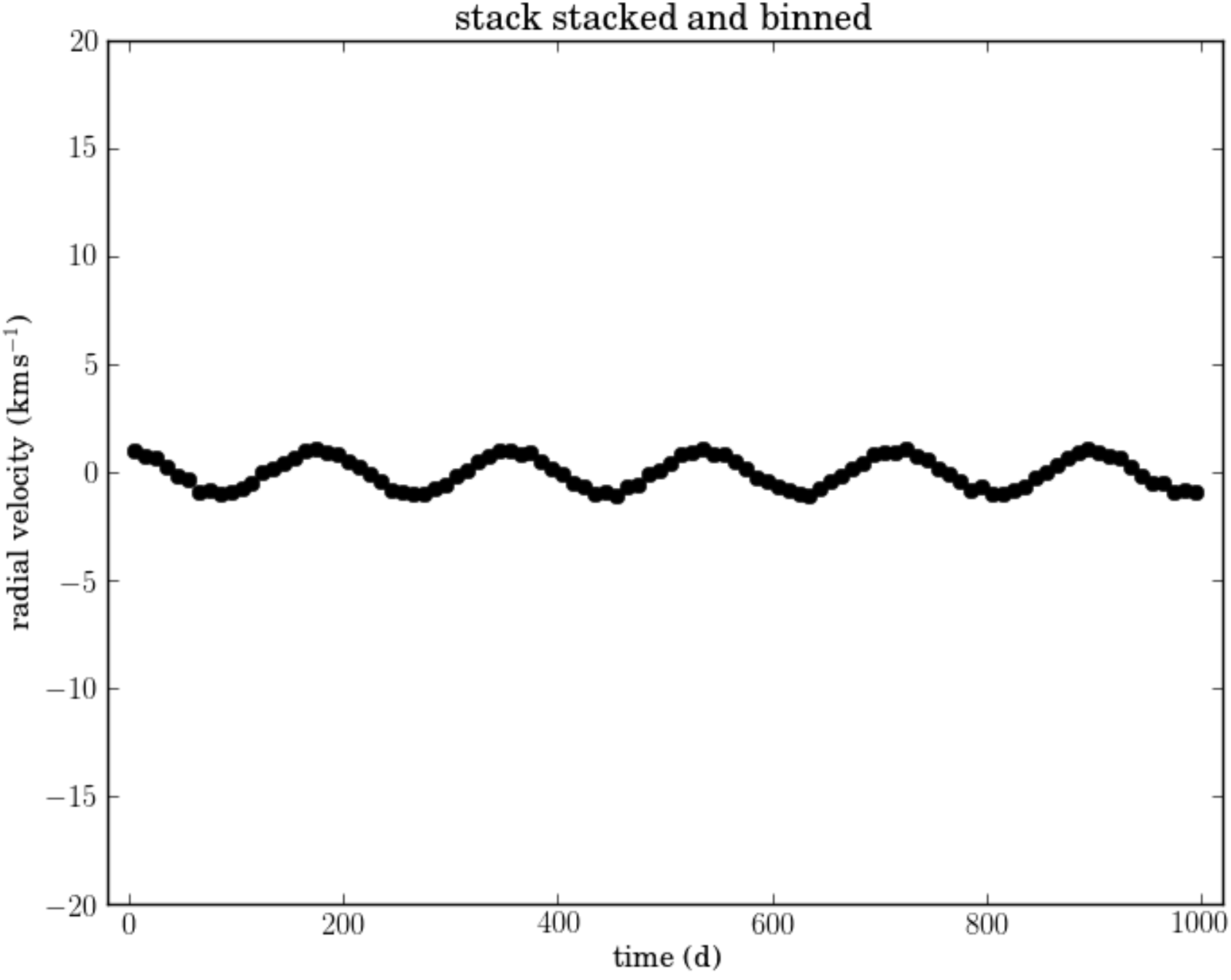

stack 004
radial velocity (kms^{-1})
time (d)
20
15
10
5
0
−5
−10
−15
−20
0
200
400
600
800
1000

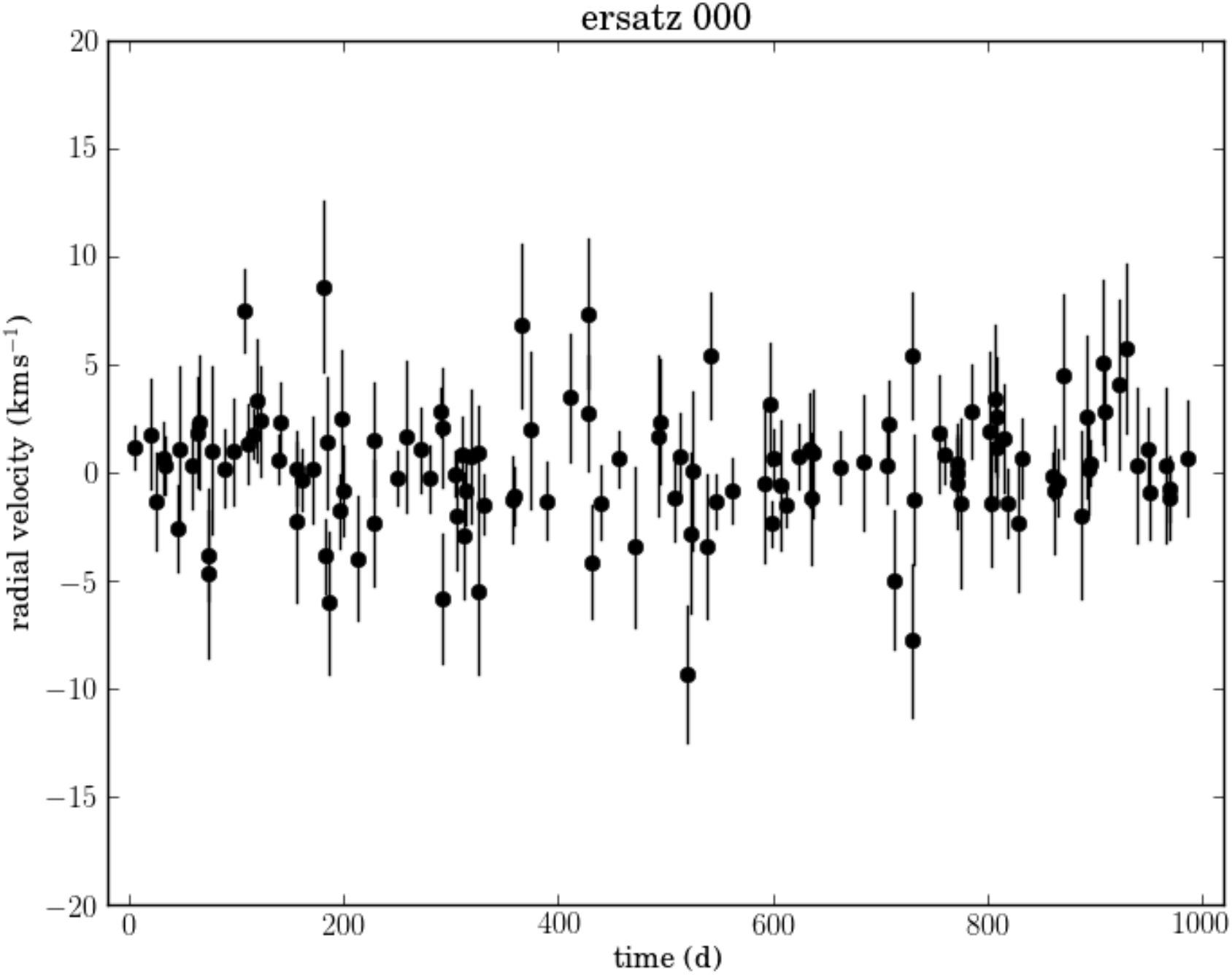

stack 005
20
15
10
5
0
−5
−10
−15
−20
0
200
400
600
800
1000
radial velocity (kms^{-1})
time (d)

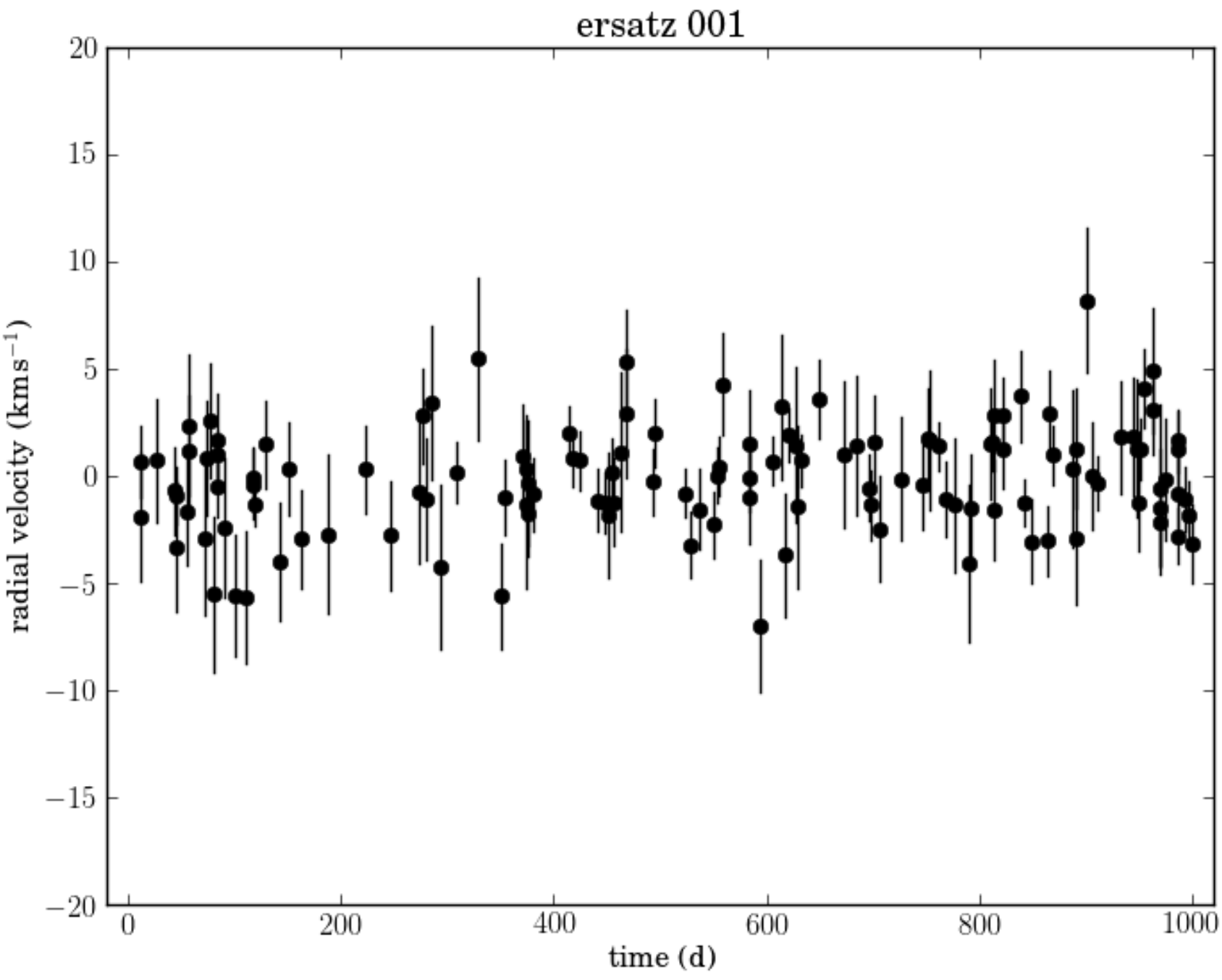

stack 006
radial velocity (kms^{-1})
time (d)
20
15
10
5
0
−5
−10
−15
−20
0
200
400
600
800
1000

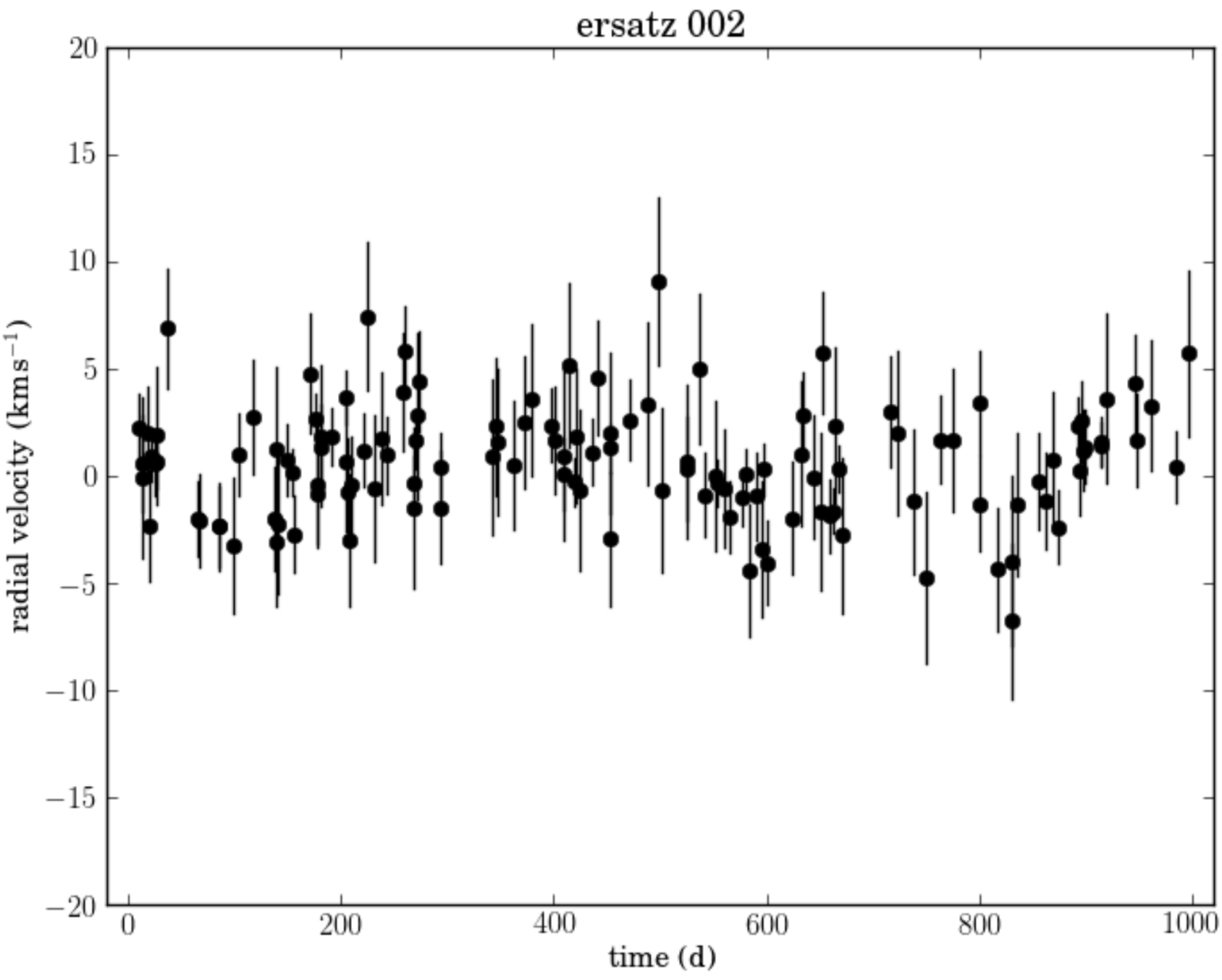

stack 007
radial velocity (kms^{-1})
time (d)
20
15
10
5
0
−5
−10
−15
−20
0
200
400
600
800
1000

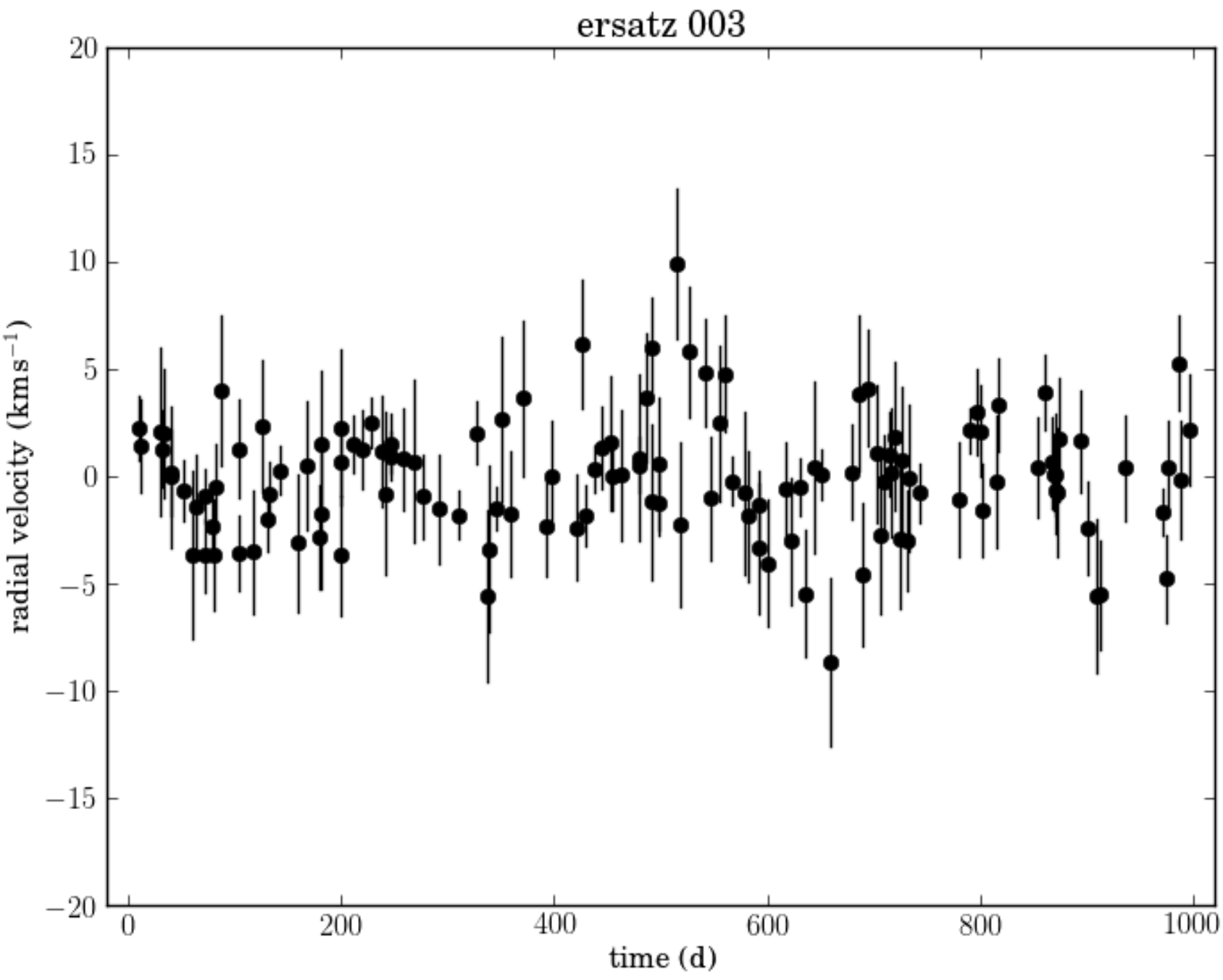

stack stack
20
15
10
5
0
−5
−10
−15
−20
radial velocity (kms^{-1})
0
200
400
600
800
1000
time (d)

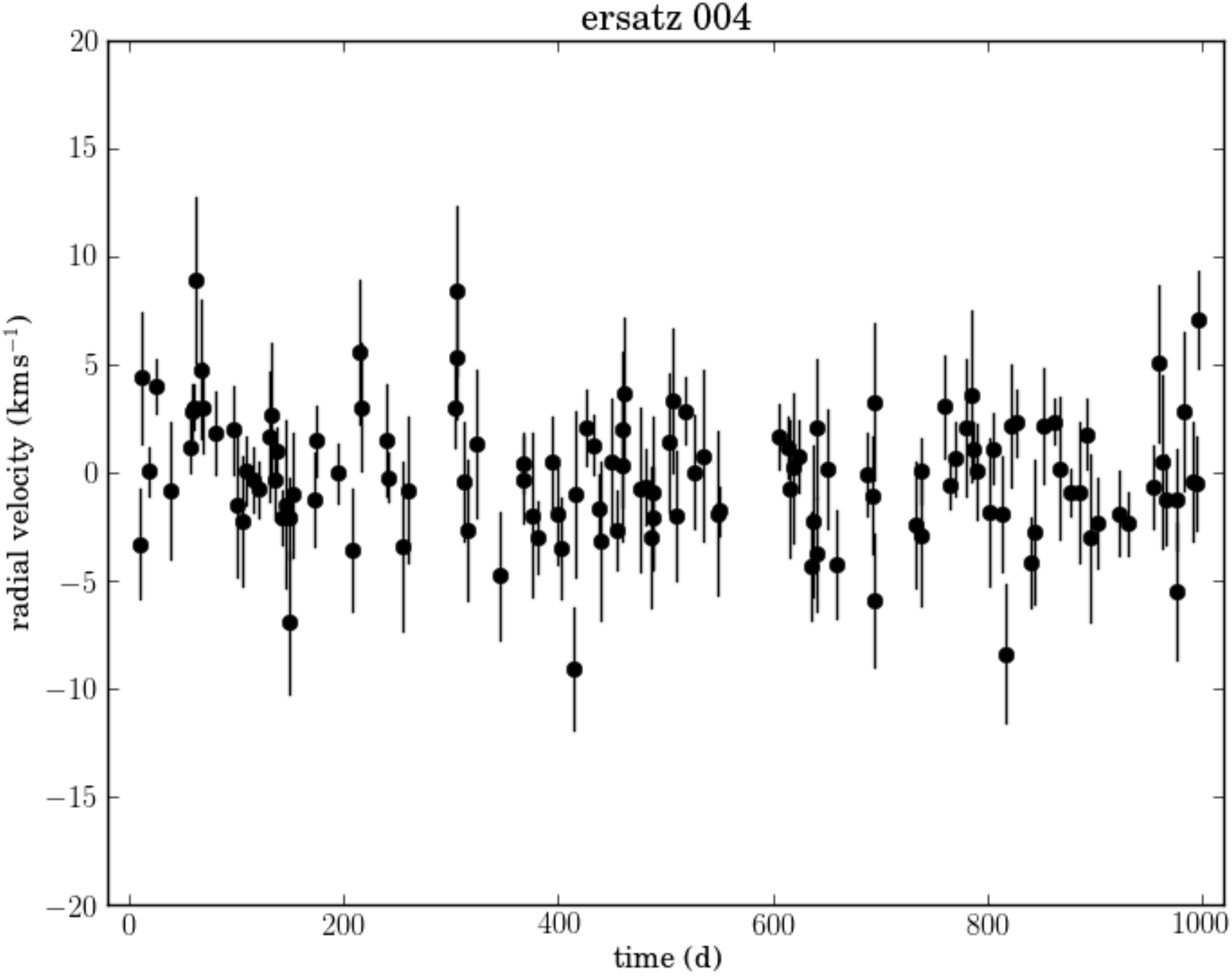

stack stacked and binned
radial velocity (kms^{-1})
time (d)
20
15
10
5
0
−5
−10
−15
−20
0
200
400
600
800
1000

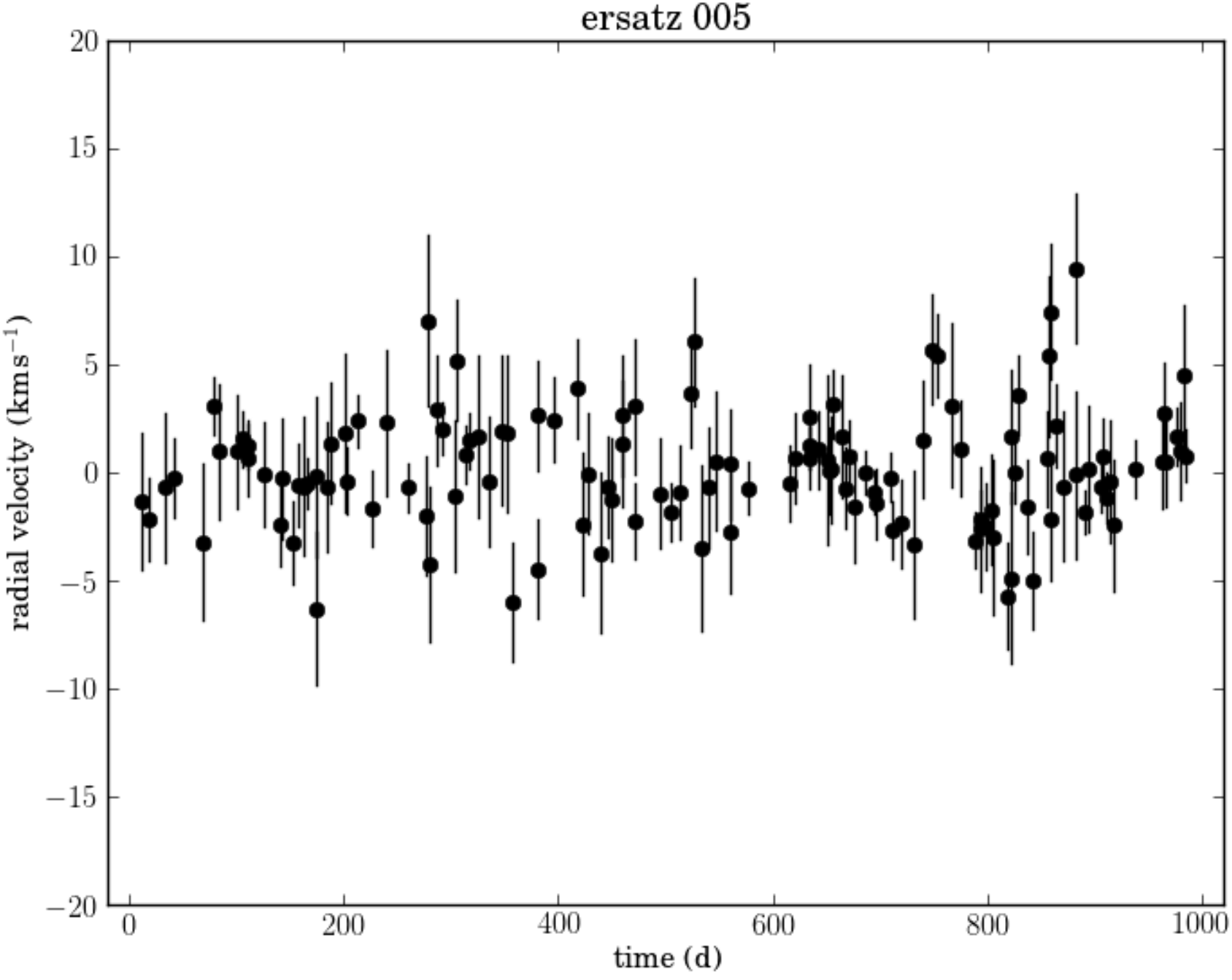

ersatz 000
radial velocity (kms^{-1})
time (d)
20
15
10
5
0
−5
−10
−15
−20
0
200
400
600
800
1000

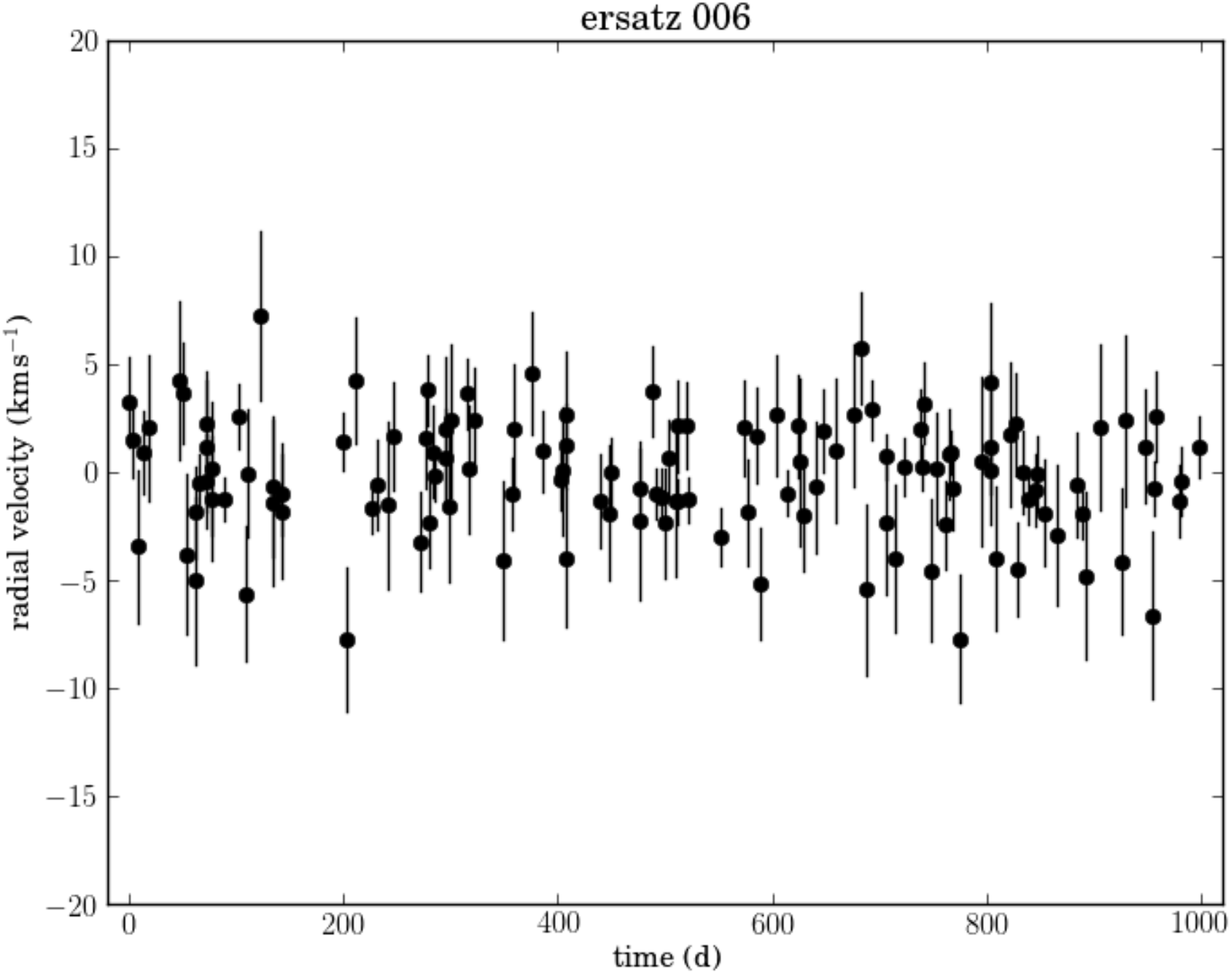

ersatz 001
radial velocity (kms^{-1})
time (d)
20
15
10
5
0
−5
−10
−15
−20
0
200
400
600
800
1000

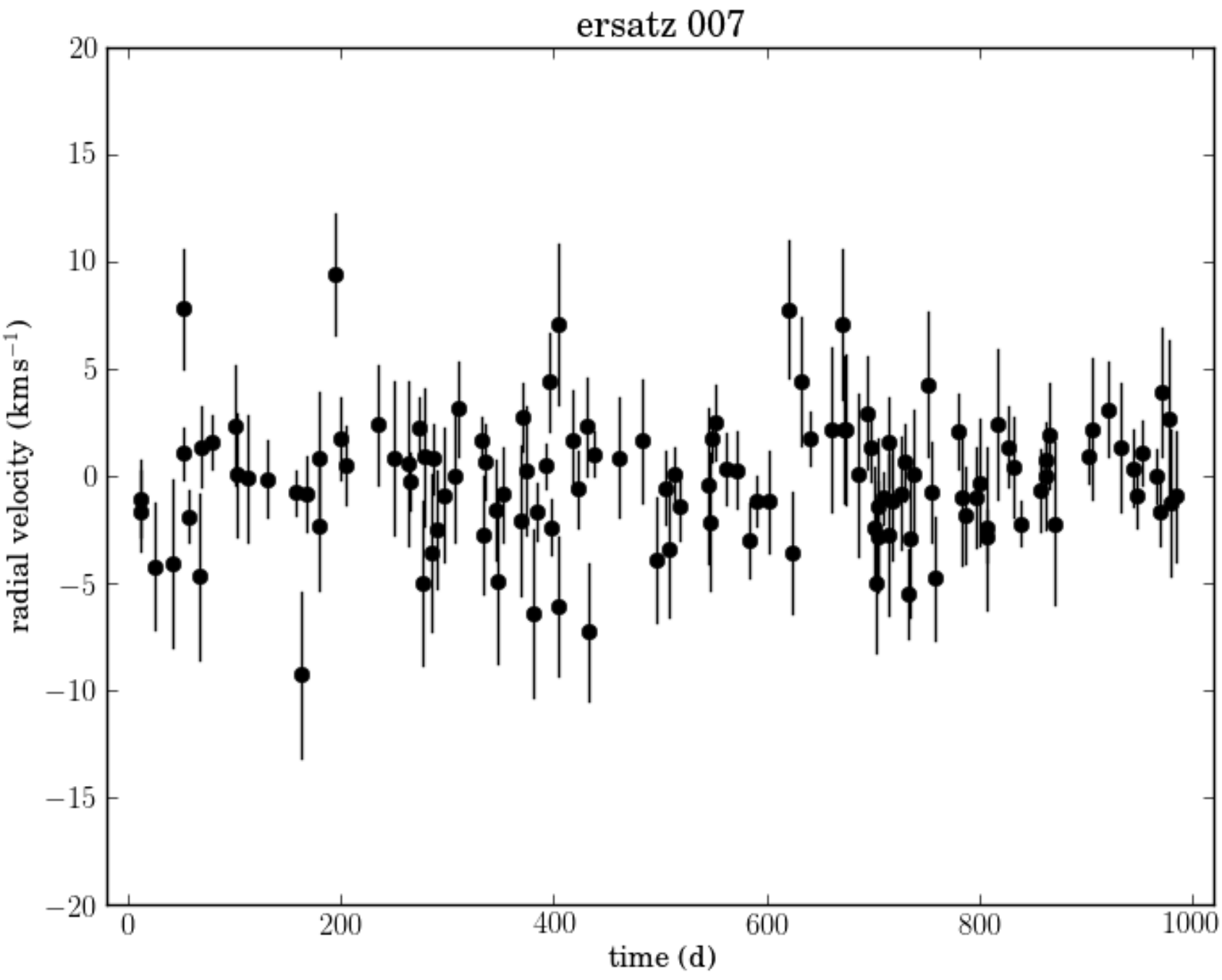

ersatz 002
radial velocity (kms^{-1})
time (d)
20
15
10
5
0
−5
−10
−15
−20
0
200
400
600
800
1000

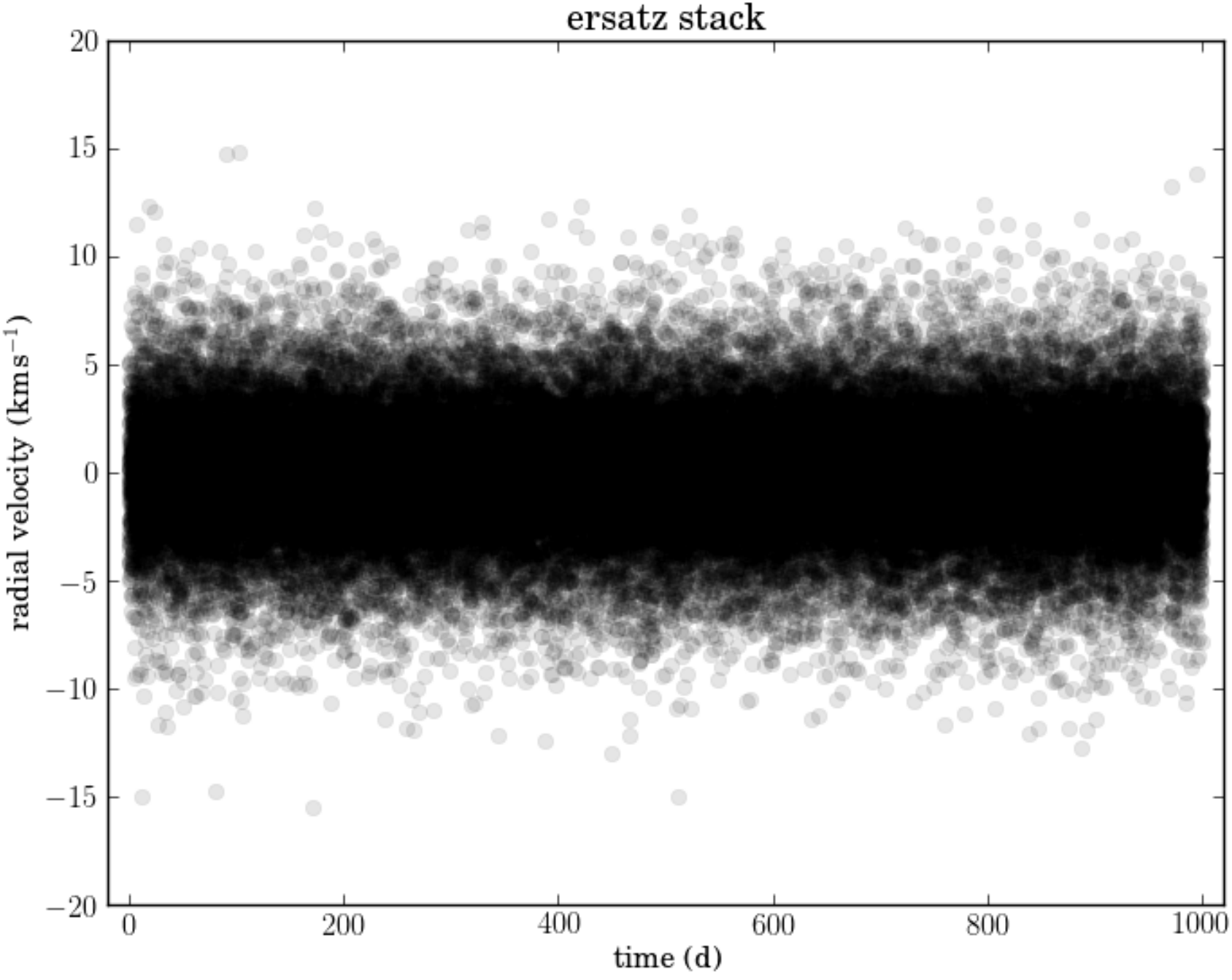

ersatz 003
radial velocity (kms^{-1})
time (d)
20
15
10
5
0
−5
−10
−15
−20
0
200
400
600
800
1000

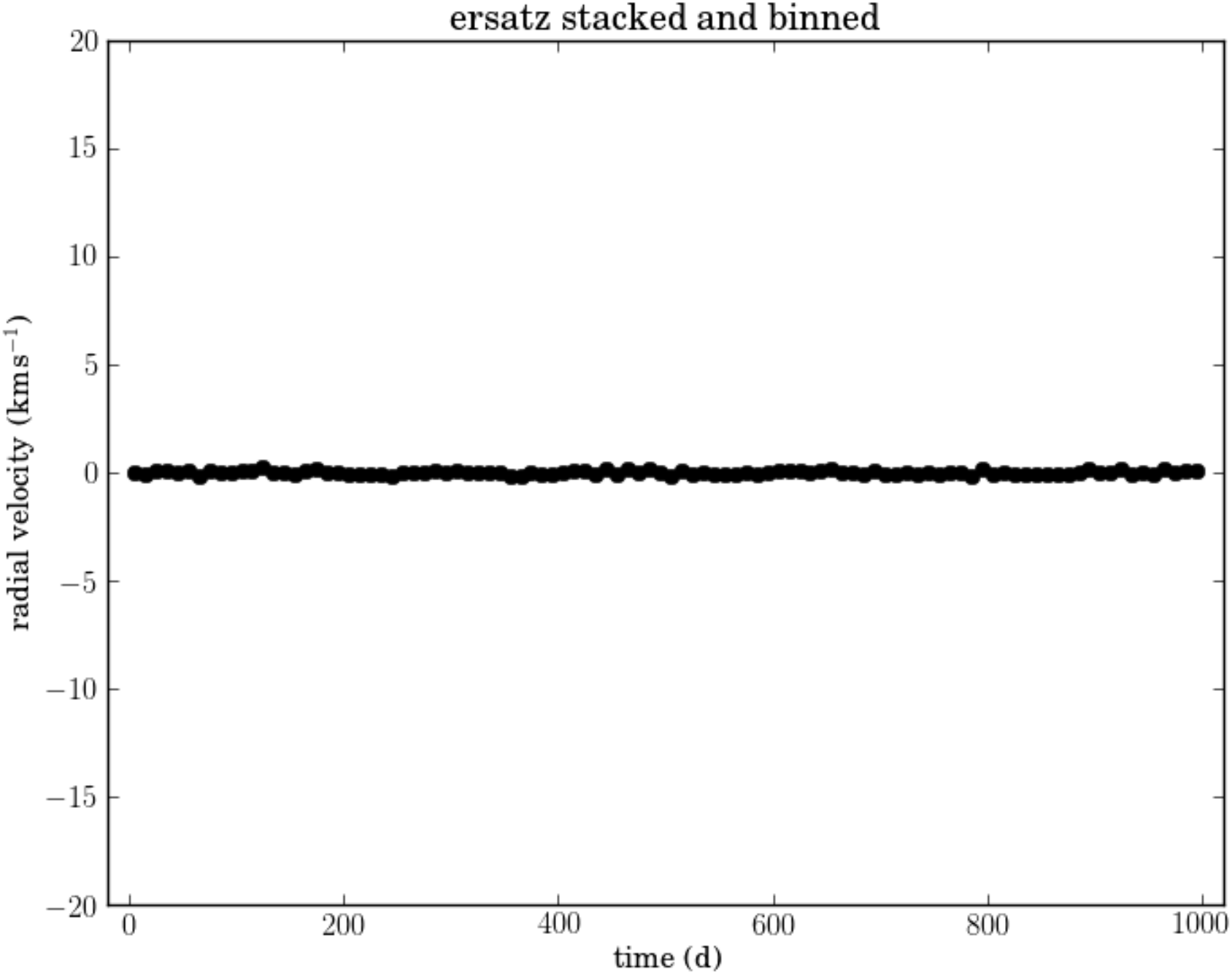

ersatz 004
radial velocity (kms^{-1})
time (d)
20
15
10
5
0
−5
−10
−15
−20
0
200
400
600
800
1000


ersatz 005
radial velocity (kms^{-1})
time (d)
20
15
10
5
0
−5
−10
−15
−20
0
200
400
600
800
1000


ersatz 006
radial velocity (kms^{-1})
time (d)
20
15
10
5
0
−5
−10
−15
−20
0
200
400
600
800
1000


ersatz 007
radial velocity (kms^{-1})
time (d)
20
15
10
5
0
−5
−10
−15
−20
0
200
400
600
800
1000


ersatz stack
radial velocity (kms^{-1})
time (d)
20
15
10
5
0
−5
−10
−15
−20
0
200
400
600
800
1000

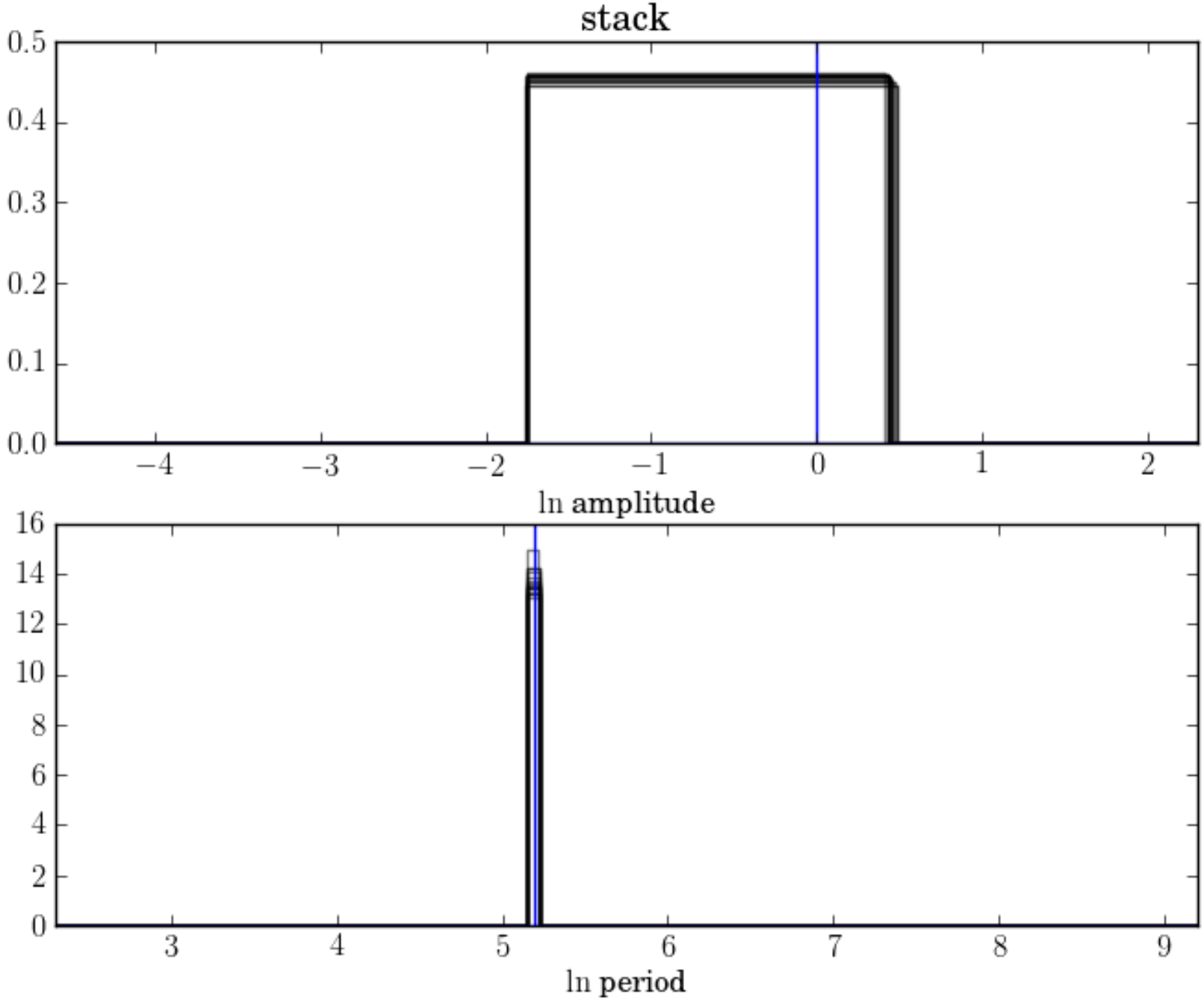

ersatz stacked and binned
radial velocity (kms^{-1})
time (d)
20
15
10
5
0
−5
−10
−15
−20
0
200
400
600
800
1000

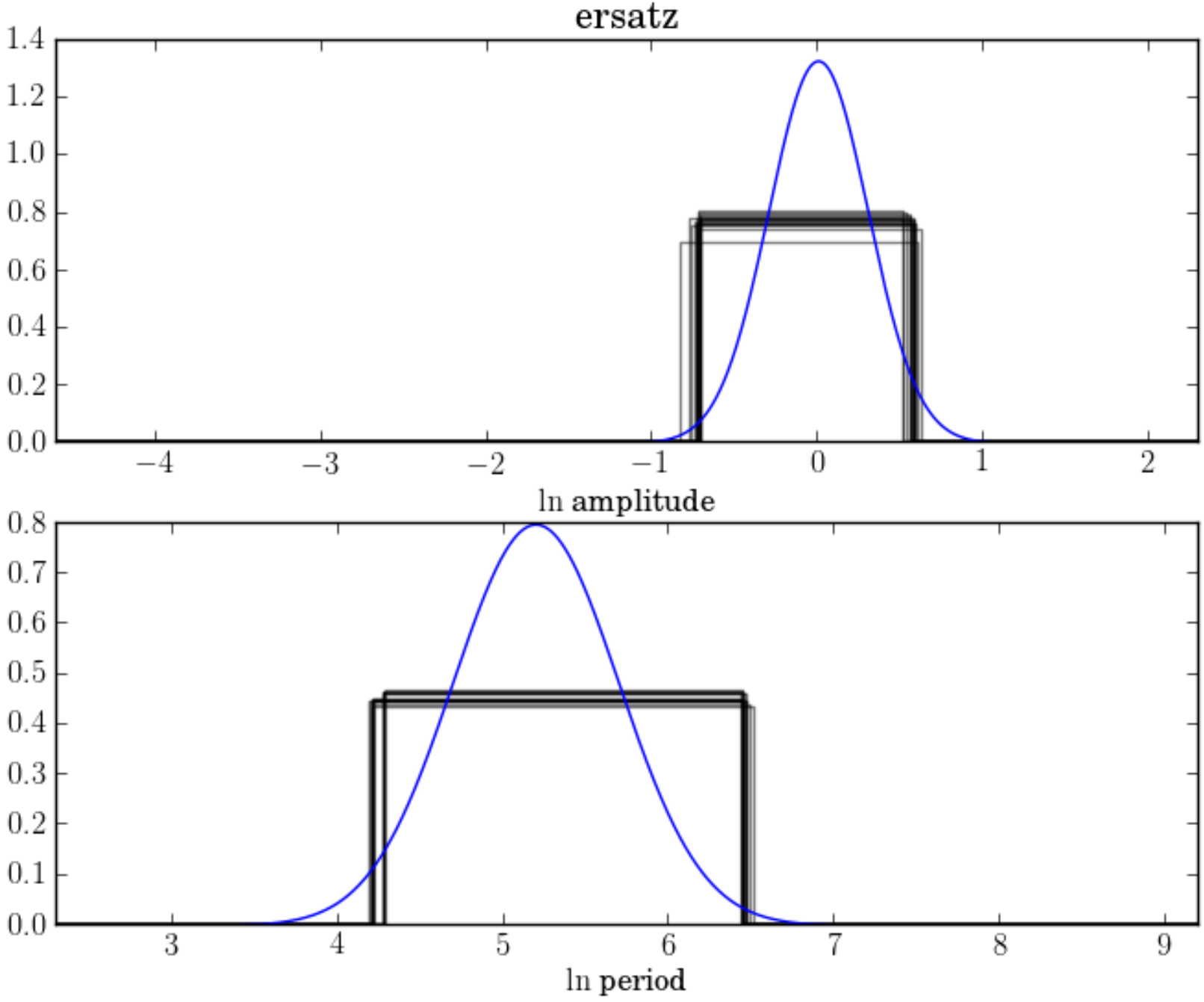
hierarchical population detection

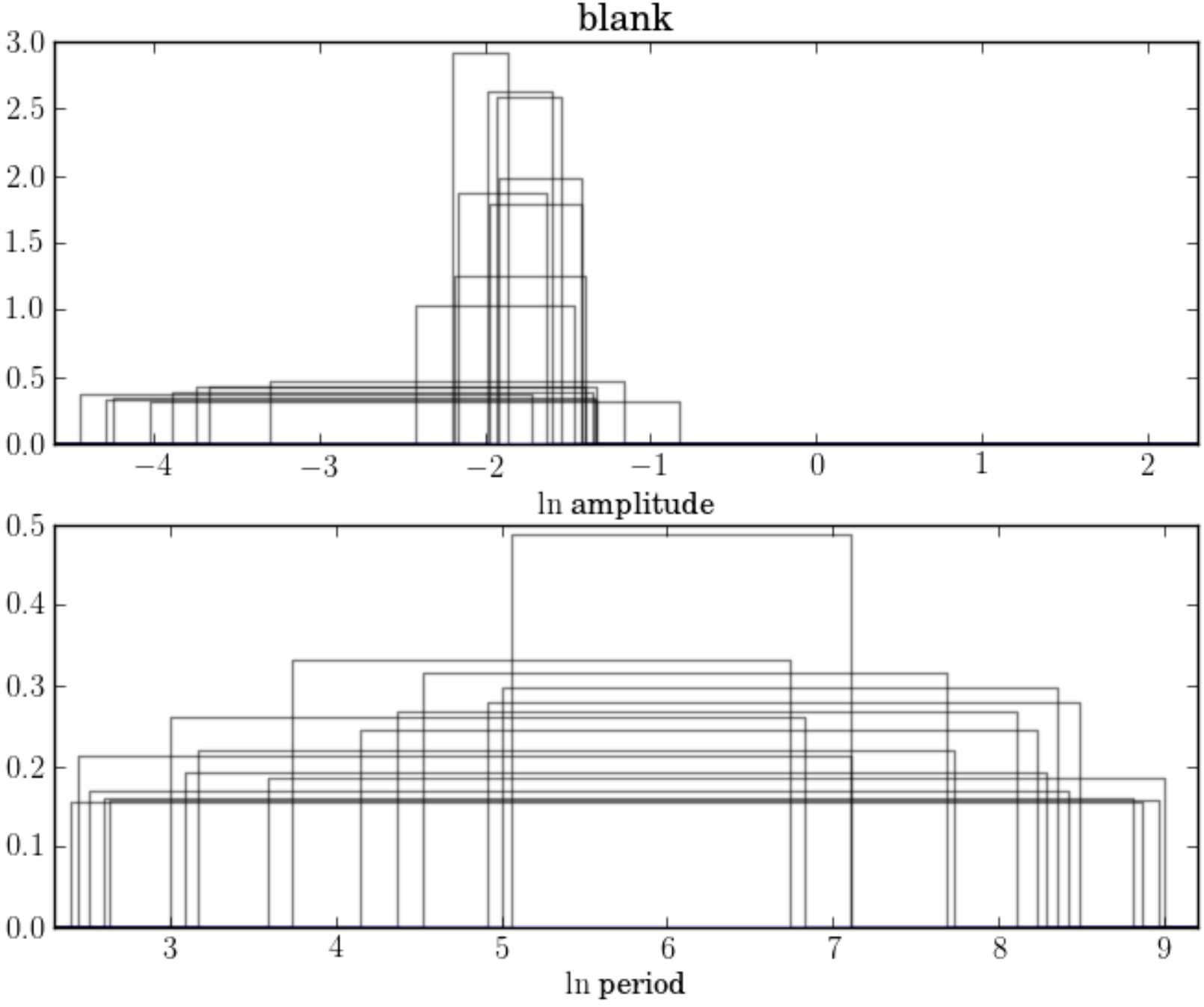
Once again, imagine you think there might be some family of priors $p(\boldsymbol{\omega}_n|\boldsymbol{\alpha})$, parameterized by some $\boldsymbol{\alpha}$, that describes the full population.

$$p(\{\mathbf{D}_n\}_{n=1}^{N}|\boldsymbol{\alpha}) \quad = \quad \prod_{n=1}^{N} \int \mathrm{d}\boldsymbol{\omega}_n \, p(\mathbf{D}_n|\boldsymbol{\omega}_n) \, p(\boldsymbol{\omega}_n|\boldsymbol{\alpha}) \quad .$$

The fact that each internal $p(\mathbf{D}_n|\boldsymbol{\omega}_n)$ contains no clear peak (no clear object detection at all) doesn't change anything!


stack
0.5
0.4
0.3
0.2
0.1
0.0
−4
−3
−2
−1
0
1
2
ln amplitude
16
14
12
10
8
6
4
2
0
3
4
5
6
7
8
9
ln period


ersatz
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−4
−3
−2
−1
0
1
2
ln amplitude
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
3
4
5
6
7
8
9
ln period


blank
3.0
2.5
2.0
1.5
1.0
0.5
0.0
−4
−3
−2
−1
0
1
2
ln amplitude
0.5
0.4
0.3
0.2
0.1
0.0
3
4
5
6
7
8
9
ln period

hierarchical population detection: Why does it work?

- The marginalized likelihood is large when there is high prior probability in locations where there is high likelihood.
- When likelihoods are broad, the best prior is the most concentrated prior that is "consistent with" **all** individual-object likelihood functions.
- The operation is a **heteroskedastic deconvolution**.
  - (in modern parlance, a "deconvolution" is always the result of fitting a generative or forward model)

hierarchical inference: What does it require?

- accurate likelihood functions
  - accurate noise models, or **parameterized** noise models
- fast inference
  - self-tuning MCMC (like *emcee*; Foreman-Mackey *et al.*, 1202.3665)
  - robustness to multimodal likelihood functions
- concept of self-calibration
  - calibration and noise parameters are not different from astrophysical parameters
- racks and racks of metal
  - (it can't be done in "map–reduce" framework)

astronomical source detection (Brewer *et al.*, 1211.5805)

- the usual story:
  - take images
  - use noise and calibration parameters to make a catalog of "detected sources"
  - perform scientific analyses on the catalog
- the hierarchical approach:
  - a scientific hypothesis implies a source distribution; this is a parameterized prior over sources
  - images are generated by sources, plus noise and calibration functions
  - infer high-level parameters by hierarchical inference
  - (involves sampling in "catalog space")

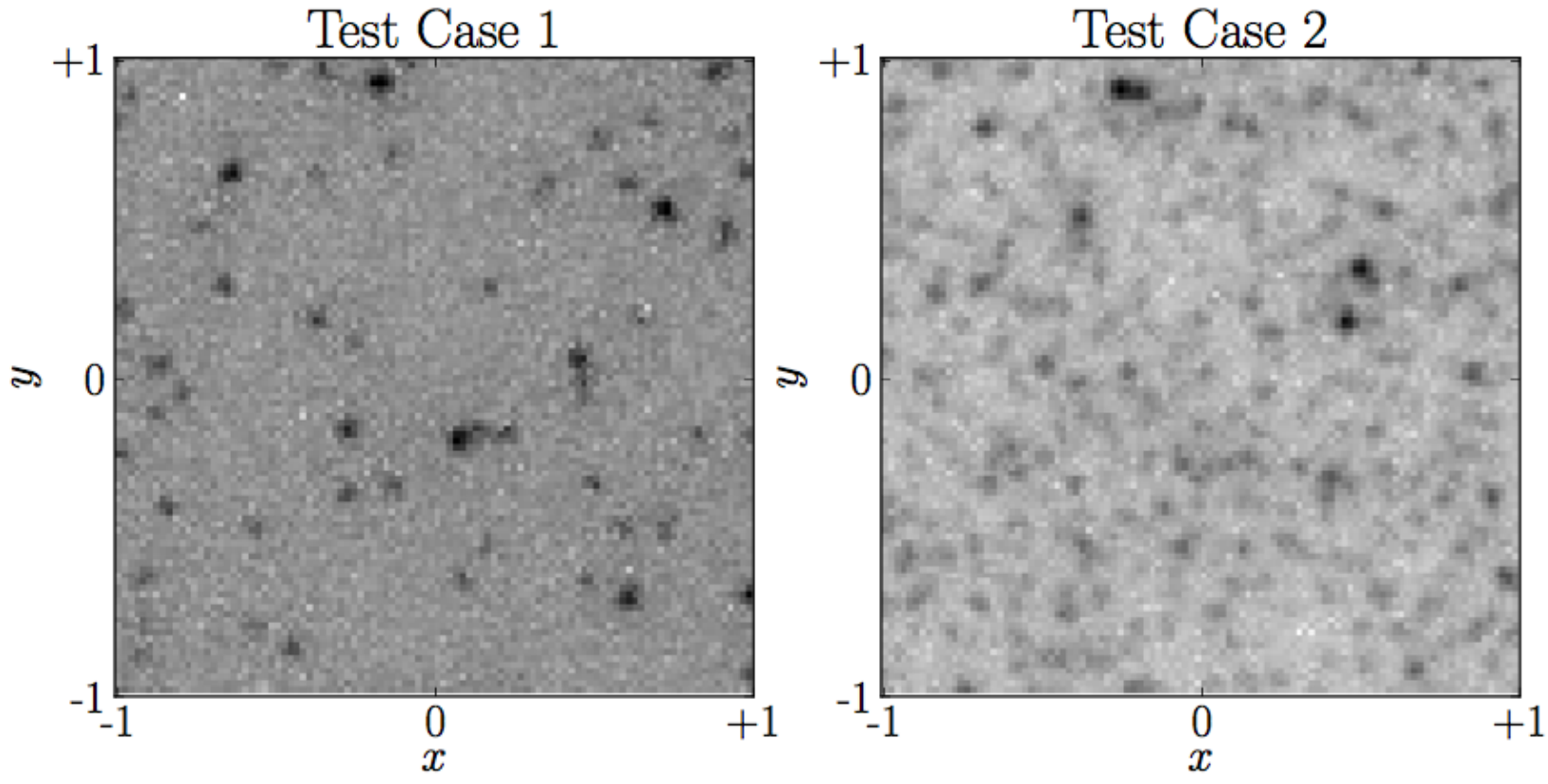
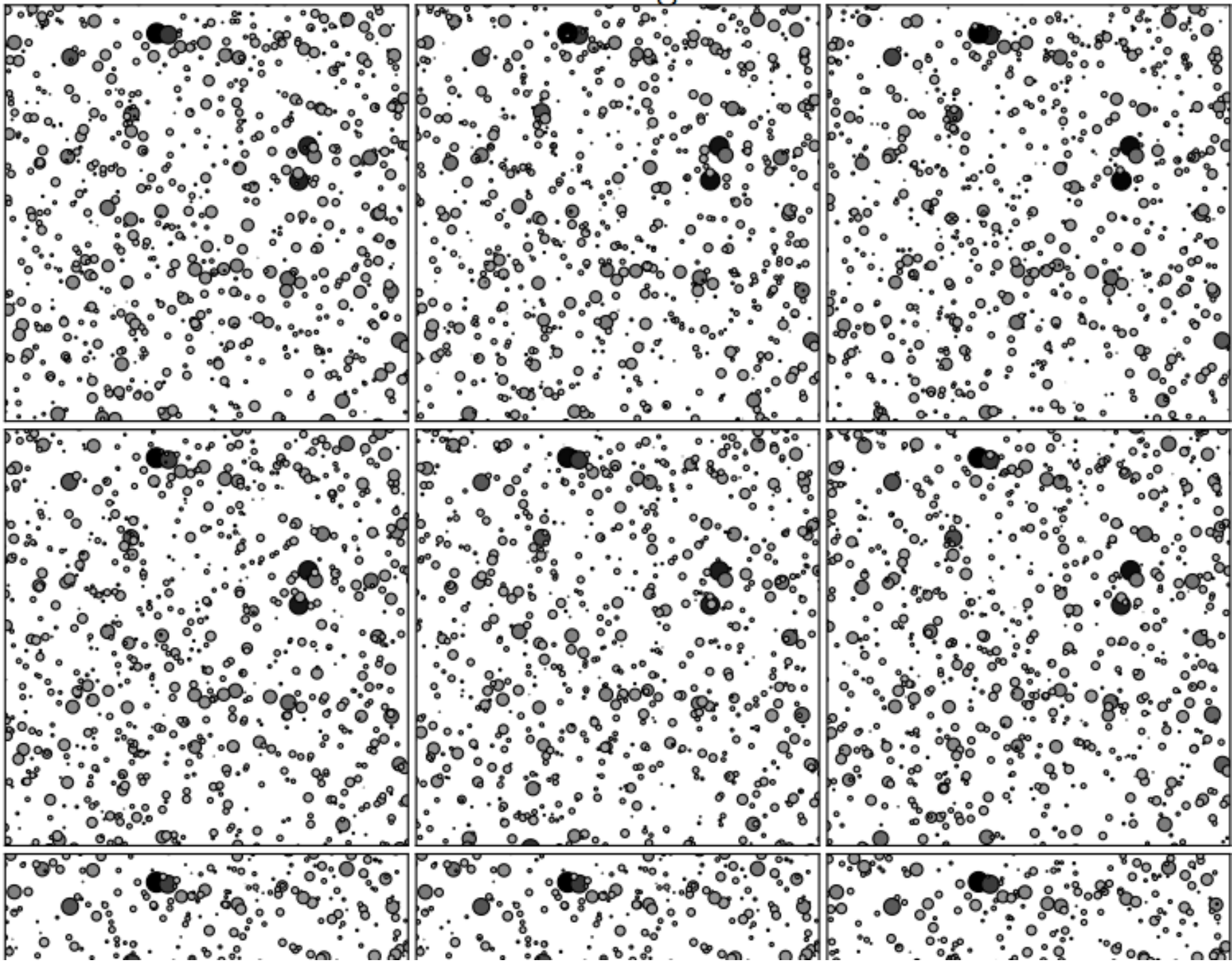


Fig. 1.— The two simulated images used to test our methodology. **Left**: An image containing ∼100 stars. **Right**: An image containing ∼1000 stars.

Catalogs



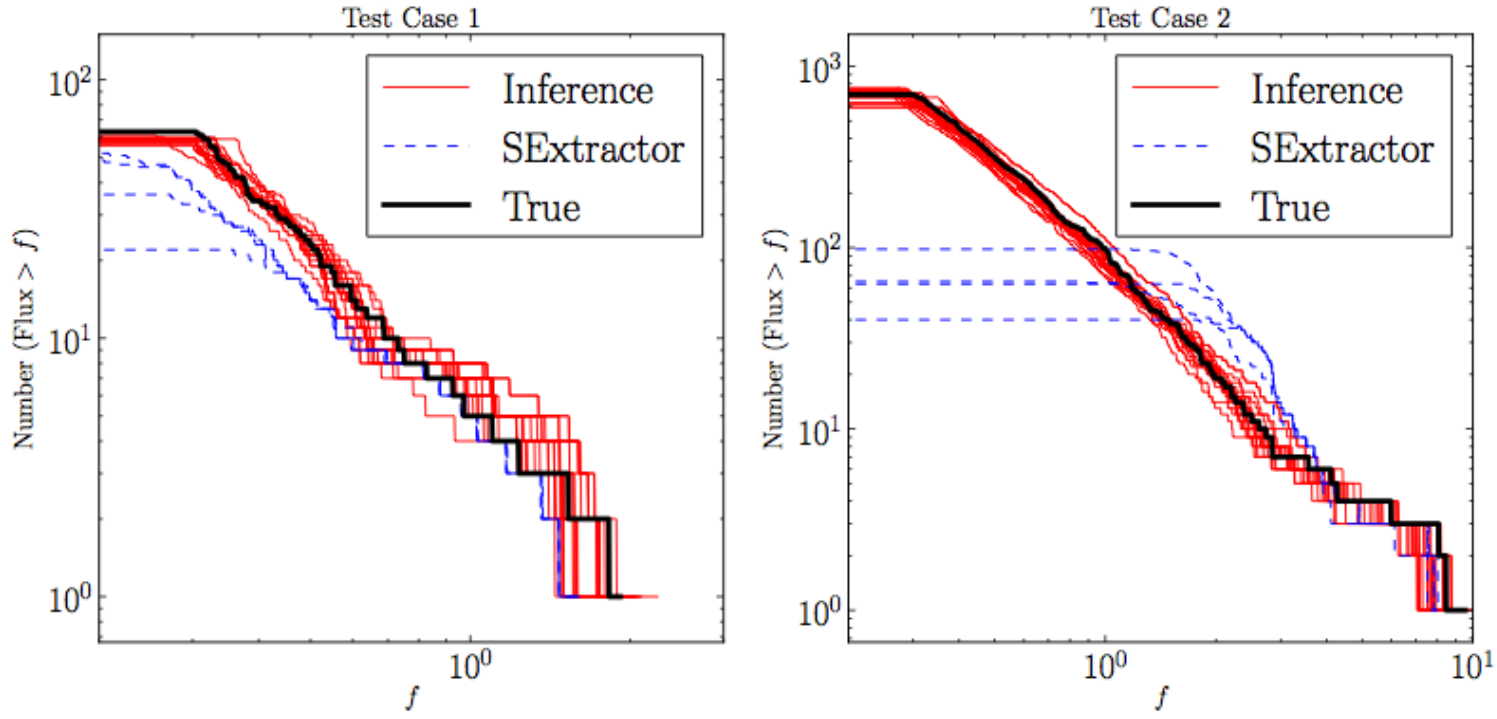


Fig. 6.— The cumulative luminosity functions (number of stars above a given flux, as a function of flux) produced by the Bayesian method (several posterior samples shown) and `SExtractor` (for various values of the threshold parameters), compared with the actual cumulative LF. Both methods correctly identify the fluxes at the bright end, with some uncertainty due to overlapping sources. However, at the lower end `SExtractor` is unable to detect all of the stars whereas the true CLF is typical of the posterior distribution.

The Tractor (Lang *et al.*, forthcoming)

- replacing the *SDSS* Catalog with
  - a maximum-likelihood model
  - a posterior sampling in catalog space
  - **a callable likelihood function**
    (see also Hogg & Lang, "Telescopes don't make catalogs!")
  - a hierarchical model

conclusions

- hierarchical inference permits
  - predictions of good data built entirely from **bad data**
  - learning of **noise-deconvolved** distribution functions
  - measurement of populations **"too faint to detect"**
  - improvements to inferences using **full-population information**
- hierarchical inference requires
  - enormous amounts of computation (in general)
  - extremely good understanding of the noise