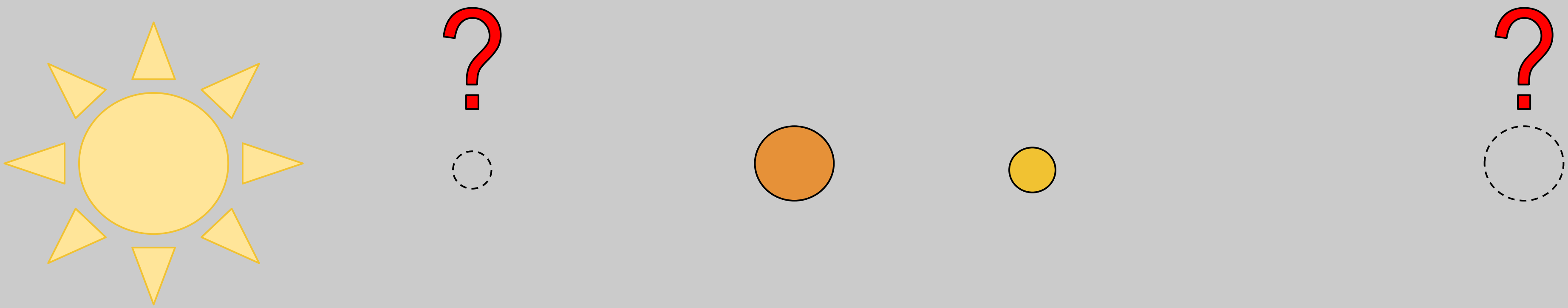
Bayesian model comparison for radial velocity: 1, 2, 3, or many planets?



Ben Nelson

Data Science Scholar at Northwestern University

@exobenelson

July 19, 2016

Sagan Summer Workshop

Outline

1. “Evidence”, Bayes factors, and decision making

2. How to efficiently compute “evidence”

3. Cross-validation as an alternative to BIC/BFs

Outline

1. “Evidence”, Bayes factors, and decision making

2. How to efficiently compute “evidence”

3. Cross-validation as an alternative to BIC/BFs

posterior prior likelihood

$$\mathtt{p}(\theta|\mathtt{d}) = \frac{\mathtt{p}(\theta)\mathtt{p}(\mathtt{d}|\theta)}{\mathtt{p}(\mathtt{d})}$$

$\mathtt{d}$: data

θ : parameters

posterior prior likelihood

$$\mathrm{p}(\theta|\mathrm{d},\mathcal{M}) = \frac{\mathrm{p}(\theta|\mathcal{M})\mathrm{p}(\mathrm{d}|\theta,\mathcal{M})}{\mathrm{p}(\mathrm{d}|\mathcal{M})}$$

d : data

θ : parameters

$\mathcal{M}$: model

$$\mathtt{p}(\theta|\mathtt{d},\mathcal{M}) = \frac{\mathtt{p}(\theta|\mathcal{M})\mathtt{p}(\mathtt{d}|\theta,\mathcal{M})}{\mathtt{p}(\mathtt{d}|\mathcal{M})}$$

$$\mathtt{p}(\mathtt{d}|\mathcal{M}) = \int \mathtt{p}(\theta|\mathcal{M})\mathtt{p}(\mathtt{d}|\theta,\mathcal{M})\mathtt{d}\theta$$

fully marginalized likelihood
or "evidence"

$\mathtt{d}$: data

θ : parameters

$\mathcal{M}$: model

Without something to compare to,
FML is not very useful...

$p(\theta$ … $\theta, \mathcal{M})$

$$p(\mathrm{d}|\mathcal{M}) = \int p(\theta|\mathcal{M})\,p(\mathrm{d}|\theta,\mathcal{M})\,\mathrm{d}\theta$$

fully marginalized likelihood
or "evidence"

d : data

θ : parameters

$\mathcal{M}$: model

Let's have two models compete!

$\mathcal{M}_{\mathtt{A}}$: model A

$\mathcal{M}_{\mathtt{B}}$: model B

Let's have two models compete!

$$\mathcal{M}_{\mathtt{A}} : \text{ model A}$$

$$\mathcal{M}_{\mathtt{B}} : \text{ model B}$$

$$\frac{\mathrm{p}(\mathrm{d}|\mathcal{M}_{\mathtt{A}})}{\mathrm{p}(\mathrm{d}|\mathcal{M}_{\mathtt{B}})} = \frac{\int \mathrm{p}(\theta_{\mathtt{A}}|\mathcal{M}_{\mathtt{A}})\mathrm{p}(\mathrm{d}|\theta_{\mathtt{A}},\mathcal{M}_{\mathtt{A}})\mathrm{d}\theta_{\mathtt{A}}}{\int \mathrm{p}(\theta_{\mathtt{B}}|\mathcal{M}_{\mathtt{B}})\mathrm{p}(\mathrm{d}|\theta_{\mathtt{B}},\mathcal{M}_{\mathtt{B}})\mathrm{d}\theta_{\mathtt{B}}}$$

Bayes Factor

Let's have two models compete!

$\mathcal{M}_{\mathtt{A}}$: model A

$\mathcal{M}_{\mathtt{B}}$: model B

$$\frac{\mathrm{p}(\mathrm{d}|\mathcal{M}_{\mathtt{A}})}{\mathrm{p}(\mathrm{d}|\mathcal{M}_{\mathtt{B}})} = \frac{\int \mathrm{p}(\theta_{\mathtt{A}}|\mathcal{M}_{\mathtt{A}})\mathrm{p}(\mathrm{d}|\theta_{\mathtt{A}},\mathcal{M}_{\mathtt{A}})\mathrm{d}\theta_{\mathtt{A}}}{\int \mathrm{p}(\theta_{\mathtt{B}}|\mathcal{M}_{\mathtt{B}})\mathrm{p}(\mathrm{d}|\theta_{\mathtt{B}},\mathcal{M}_{\mathtt{B}})\mathrm{d}\theta_{\mathtt{B}}}$$

Bayes Factor

Prior on models

Posterior Odds Ratio

$$\frac{\mathrm{p}(\mathrm{d}|\mathcal{M}_{\mathtt{A}})}{\mathrm{p}(\mathrm{d}|\mathcal{M}_{\mathtt{B}})} \times \frac{\mathrm{p}(\mathcal{M}_{\mathtt{A}})}{\mathrm{p}(\mathcal{M}_{\mathtt{B}})} = \frac{\mathrm{p}(\mathcal{M}_{\mathtt{A}}|\mathrm{d})}{\mathrm{p}(\mathcal{M}_{\mathtt{B}}|\mathrm{d})}$$

“So, what do I do with this?”

$$\frac{\mathrm{p}(\mathtt{d}|\mathcal{M}_{\mathtt{A}})}{\mathrm{p}(\mathtt{d}|\mathcal{M}_{\mathrm{B}})} \times \frac{\mathrm{p}(\mathcal{M}_{\mathtt{A}})}{\mathrm{p}(\mathcal{M}_{\mathrm{B}})} = \frac{\mathrm{p}(\mathcal{M}_{\mathtt{A}}|\mathtt{d})}{\mathrm{p}(\mathcal{M}_{\mathrm{B}}|\mathtt{d})}$$

“So, what do I do with this?”

$$\frac{\mathtt{p}(\mathtt{d}|\mathcal{M}_\mathtt{A})}{\mathtt{p}(\mathtt{d}|\mathcal{M}_\mathtt{B})} \times \frac{\mathtt{p}(\mathcal{M}_\mathtt{A})}{\mathtt{p}(\mathcal{M}_\mathtt{B})} = \frac{\mathtt{p}(\mathcal{M}_\mathtt{A}|\mathtt{d})}{\mathtt{p}(\mathcal{M}_\mathtt{B}|\mathtt{d})}$$

If BF/POR is really huge, favor $\mathcal{M}_\mathtt{A}$

If BF/POR is really small, favor $\mathcal{M}_\mathtt{B}$

Otherwise, it’s not very decisive.

"So, what do I do with this?"

$$\frac{p(d|\mathcal{M}_A)}{p(d|\mathcal{M}_B)} \times \frac{p(\mathcal{M}_A)}{p(\mathcal{M}_B)} = \frac{p(\mathcal{M}_A|d)}{p(\mathcal{M}_B|d)}$$

[redacted] favor $\mathcal{M}_A$

Takeaway #1:
model comparison ≠ model selection

If BF/POR is [redacted]

Otherwise, it's not very decisive.

Why model comparison ≠ model selection

Model comparison just gives probabilities.
Model selection is a decision based on other (outside) factors, i.e. a cost function/utility.

Why model comparison ≠ model selection

Model comparison just gives probabilities.
Model selection is a decision based on other (outside) factors, i.e. a cost function/utility.

Most rigorous thing to do is **average** all models, not select the most probable.

$$p(\theta|d) = \sum_{k=1}^{K} p(\theta|d, \mathcal{M}_k)\, p(\mathcal{M}_k|d)$$

Outline

1. "Evidence", Bayes factors, and decision making

2. How to efficiently compute "evidence"

3. Cross-validation as an alternative to BIC/BFs

Computing FML in practice

$$\mathtt{p}(\mathtt{d}|\mathcal{M}) = \int \mathtt{p}(\theta|\mathcal{M})\mathtt{p}(\mathtt{d}|\theta, \mathcal{M})\mathtt{d}\theta$$

Computing FML in practice

$$\mathrm{p}(\mathrm{d}|\mathcal{M}) = \int \mathrm{p}(\theta|\mathcal{M})\mathrm{p}(\mathrm{d}|\theta,\mathcal{M})\mathrm{d}\theta$$

Takeaway #2:
This integral is **HARD**

Computing FML in practice

$$\mathrm{p}(\mathrm{d}|\mathcal{M}) = \int \mathrm{p}(\theta|\mathcal{M})\mathrm{p}(\mathrm{d}|\theta,\mathcal{M})\mathrm{d}\theta$$

Takeaway #2:

This integral is **HARD***

* but there's an entire literature on how to compute this efficiently

Thermodynamic Integration (Theory)

1. Start with **parallel-tempering MCMC**.
   i.e. multiple MCMCs with likelihoods taken to different powers: $\mathtt{p}_\beta(\theta|\mathtt{d}) \propto \mathtt{p}(\theta)\mathtt{p}^\beta(\mathtt{d}|\theta)$

Earl & Deem 2010, Phys Chem Chem Phys

Thermodynamic Integration (Theory)

1. Start with **parallel-tempering MCMC**.
   i.e. multiple MCMCs with likelihoods taken to different powers: $\mathtt{p}_\beta(\theta|\mathtt{d}) \propto \mathtt{p}(\theta)\mathtt{p}^\beta(\mathtt{d}|\theta)$

2. FML at β is $\mathtt{p}_\beta(\mathtt{d}) = \int \mathtt{p}(\theta)\mathtt{p}^\beta(\mathtt{d}|\theta)\mathtt{d}\theta$

Earl & Deem 2010, Phys Chem Chem Phys

Thermodynamic Integration (Theory)

1. Start with **parallel-tempering MCMC**.
   i.e. multiple MCMCs with likelihoods taken to different powers: $\mathrm{p}_\beta(\theta|\mathrm{d}) \propto \mathrm{p}(\theta)\mathrm{p}^\beta(\mathrm{d}|\theta)$

2. FML at β is $\mathrm{p}_\beta(\mathrm{d}) = \int \mathrm{p}(\theta)\mathrm{p}^\beta(\mathrm{d}|\theta)\mathrm{d}\theta$

3. Ultimately, derive...

$$\mathbf{p}(\mathbf{d}) \approx \exp\left[\int_0^1 \mathbf{d}\beta \langle \log \mathbf{p}(\mathbf{d}|\theta)\rangle_\beta\right]$$

"average" log-likelihood at β

Earl & Deem 2010, Phys Chem Chem Phys

Thermodynamic Integration (Practice)

$$\mathtt{p}(\mathtt{d}) \approx \exp\left[\int_0^1 \mathtt{d}\beta \langle \log \mathtt{p}(\mathtt{d}|\theta)\rangle_\beta\right]$$

Advantages:

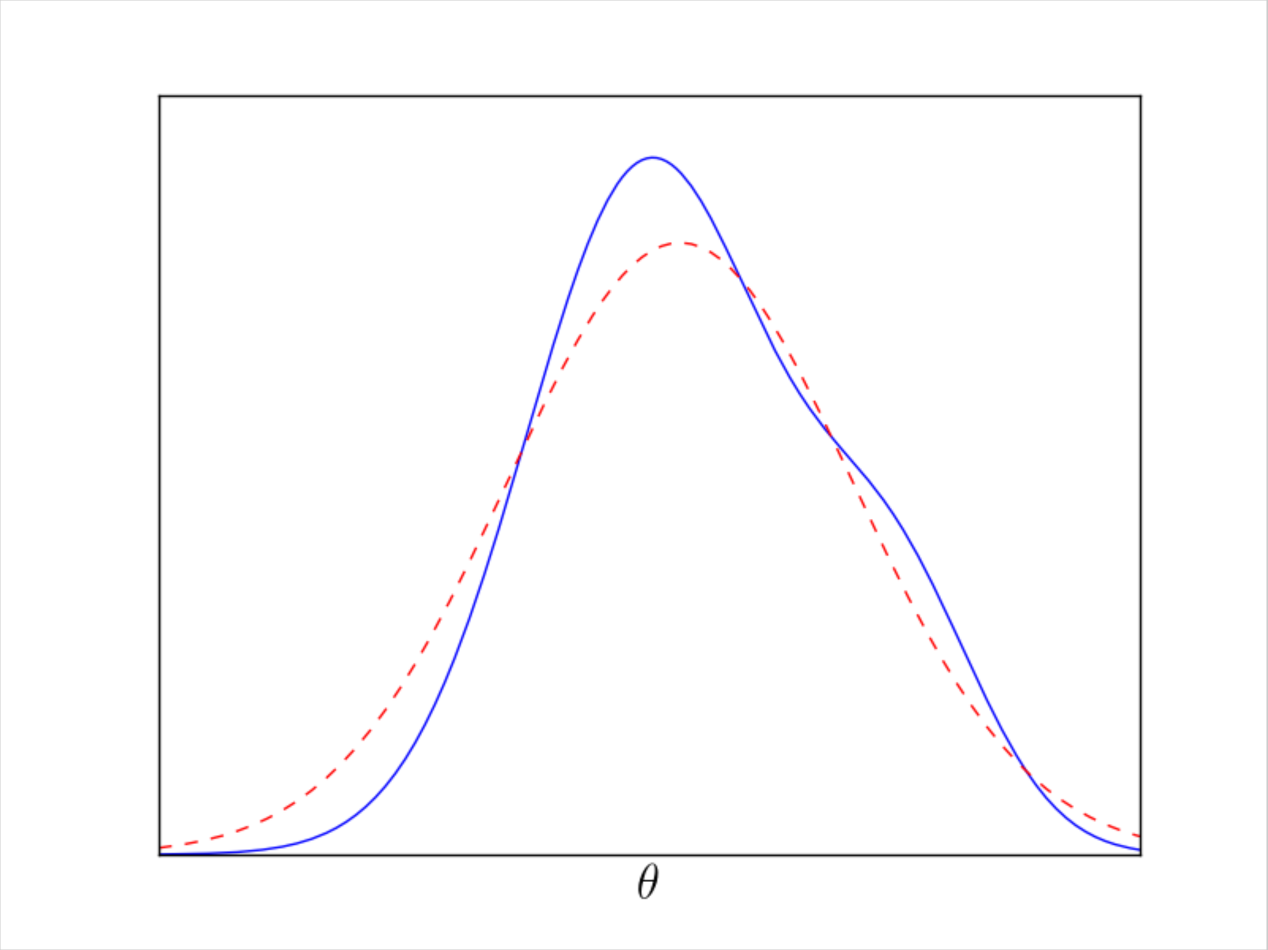
1. A nice side effect of performing PTMCMC
2. Already implemented in emcee*

Caveats:

Need a robust estimate of $\langle \log \mathtt{p}(\mathtt{d}|\theta)\rangle_\beta$ at every β

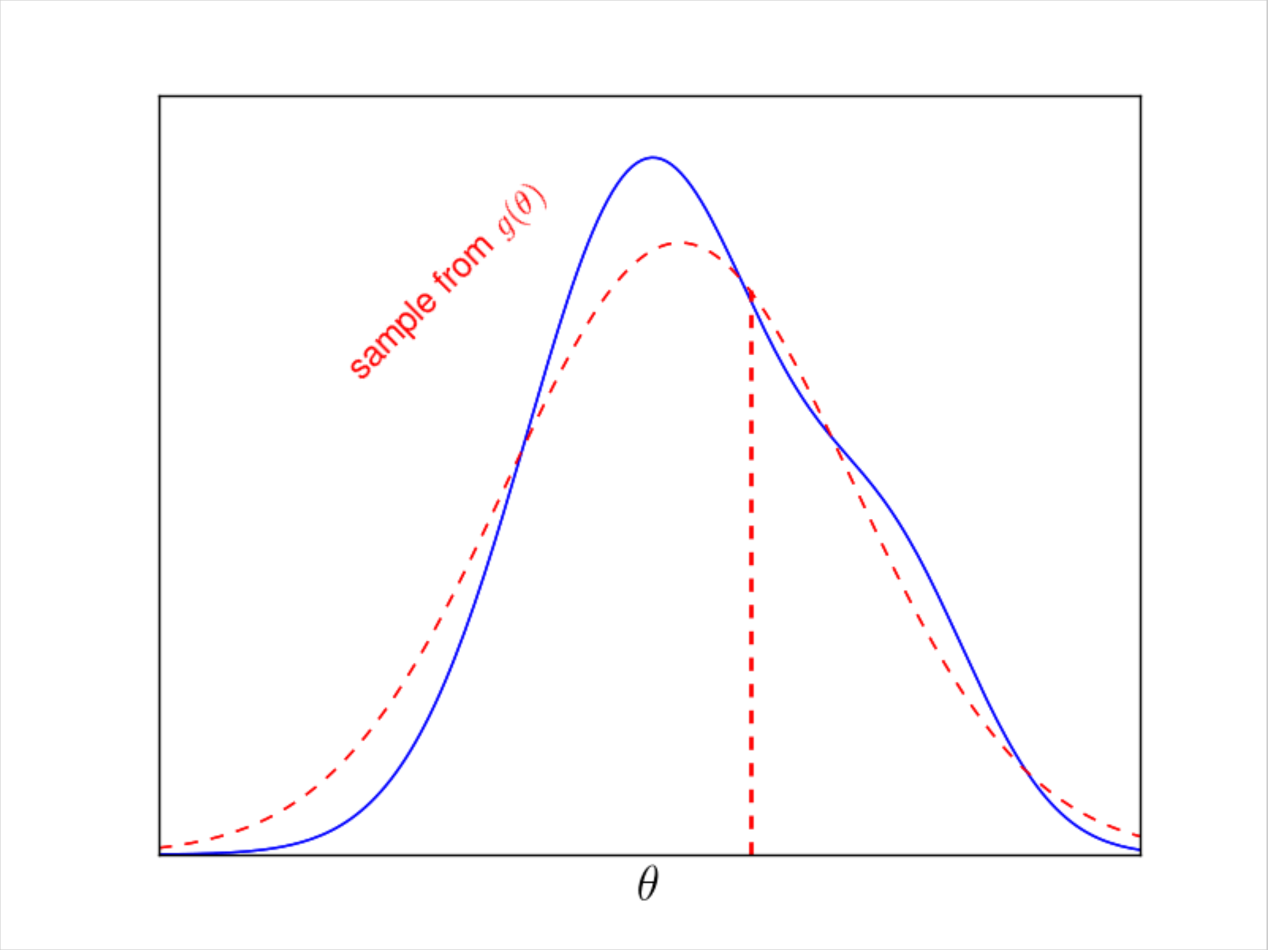
*dan.iel.fm/emcee/current/user/pt

Importance Sampling (Theory)



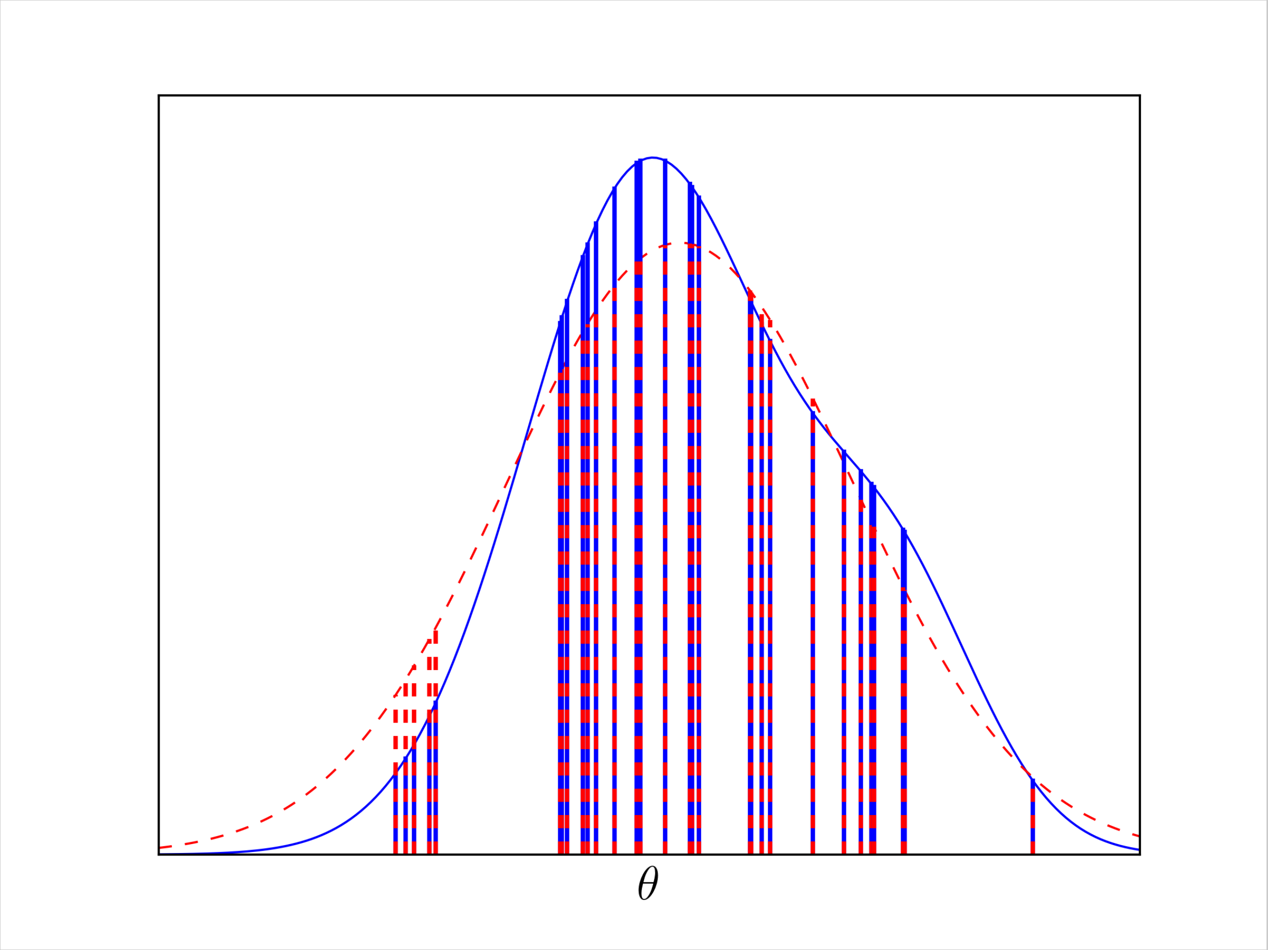
$$p(d) = \int \frac{p(\theta)p(d|\theta)}{g(\theta)} g(\theta) d\theta$$

Importance Sampling (Theory)



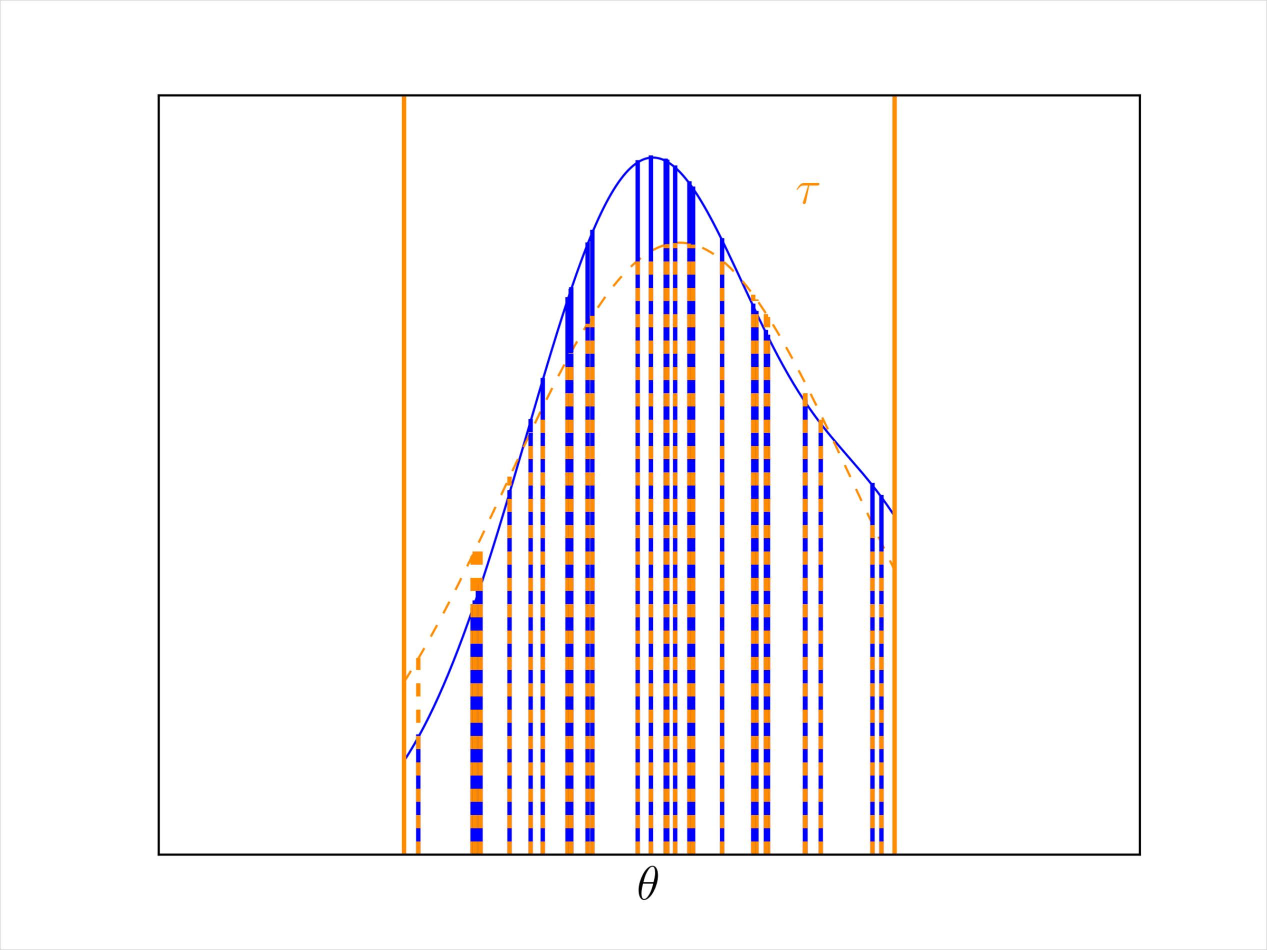
$$\mathrm{p(d)} = \int \frac{\mathrm{p}(\theta)\mathrm{p}(\mathrm{d}|\theta)}{\mathrm{g}(\theta)} \mathrm{g}(\theta) \mathrm{d}\theta$$

Importance Sampling (Theory)



$$\widehat{p(d)} \approx \frac{1}{N} \sum_{\theta_i \sim g(\theta)} \frac{p(\theta_i)p(d|\theta_i)}{g(\theta_i)}$$

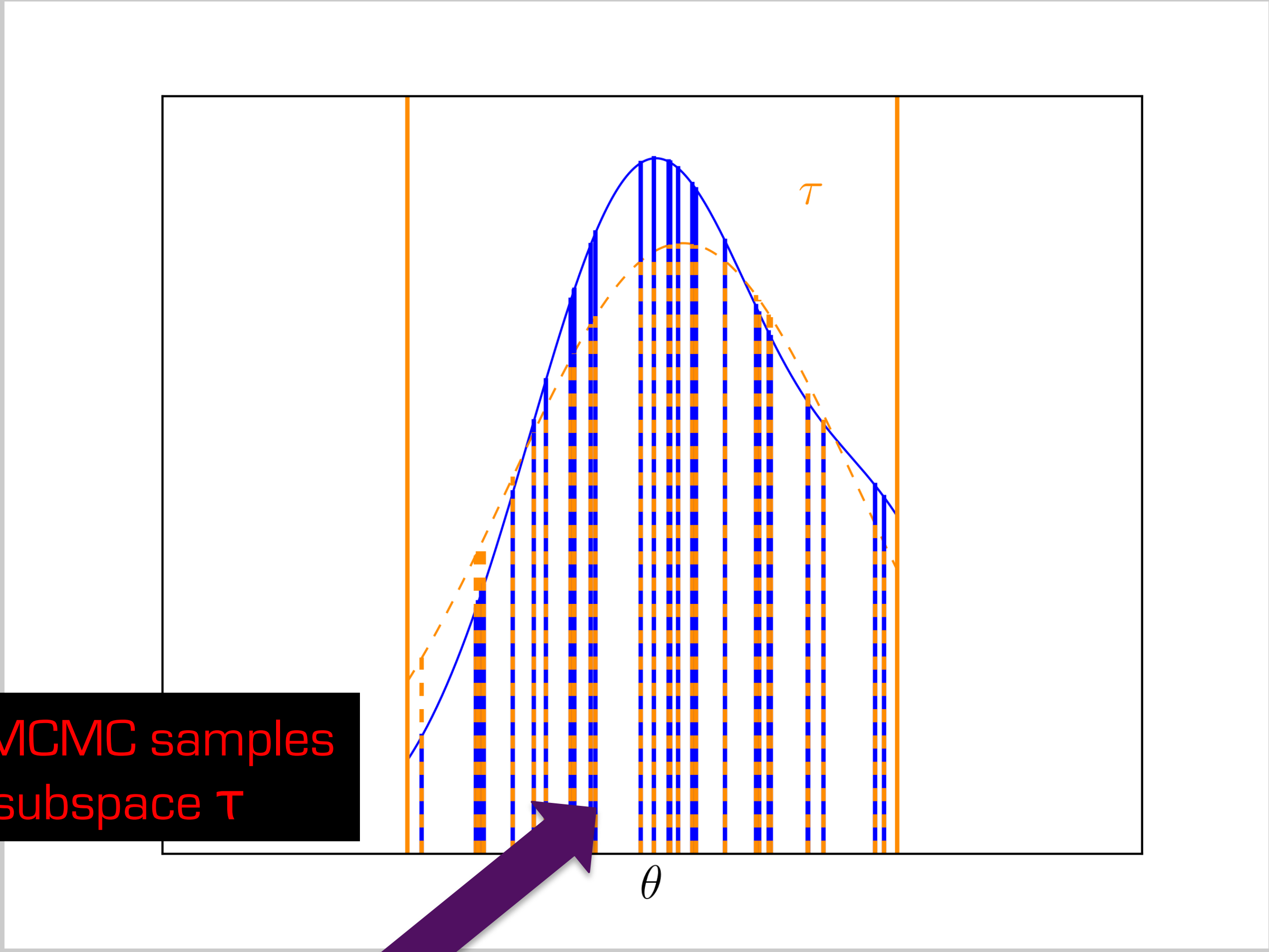
Importance Sampling (Theory)



$$\ldots = \frac{1}{N} \sum_{\theta_i \sim g_\tau(\theta)} \frac{p(\theta_i)p(d|\theta_i)}{g_\tau(\theta_i)}$$

PC Guo Thesis 2012
Weinberg+ 2013
Nelson+ 2016

Importance Sampling (Theory)

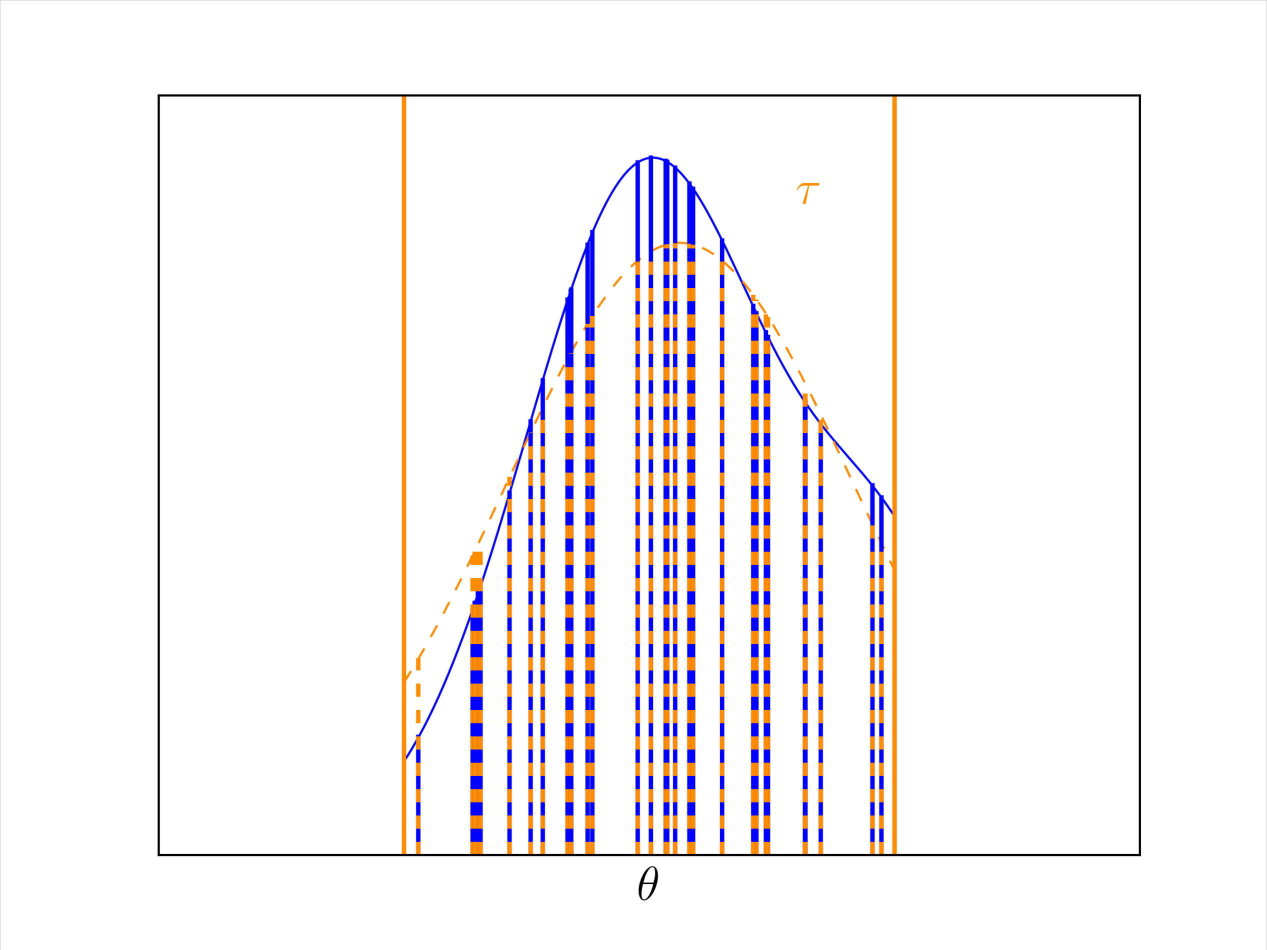


Fraction of MCMC samples to reside in subspace τ

$$f_{MCMC} \times \widehat{p(d)} \approx \frac{1}{N} \sum_{\theta_i \sim g_\tau(\theta)} \frac{p(\theta_i)p(d|\theta_i)}{g_\tau(\theta_i)}$$

PC Guo Thesis 2012
Weinberg+ 2013
Nelson+ 2016

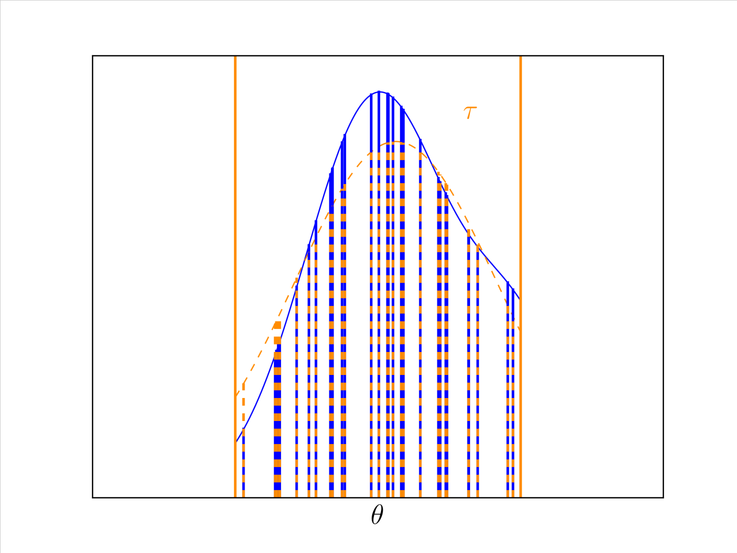
Importance Sampling (Theory)



$$\widehat{p(d)} \approx \frac{1}{N \times f_{MCMC}} \sum_{\theta_i \sim g_\tau(\theta)} \frac{p(\theta_i)p(d|\theta_i)}{g_\tau(\theta_i)}$$

PC Guo Thesis 2012
Weinberg+ 2013
Nelson+ 2016

Importance Sampling (Practice)



$$\widehat{p(d)} \approx \frac{1}{N} \sum_{\theta_i \sim g(\theta)} \frac{p(\theta_i)p(d|\theta_i)}{g(\theta_i)}$$

$$\widehat{p(d)} \approx \frac{1}{N \times f_{MCMC}} \sum_{\theta_i \sim g_\tau(\theta)} \frac{p(\theta_i)p(d|\theta_i)}{g_\tau(\theta_i)}$$

Advantages:

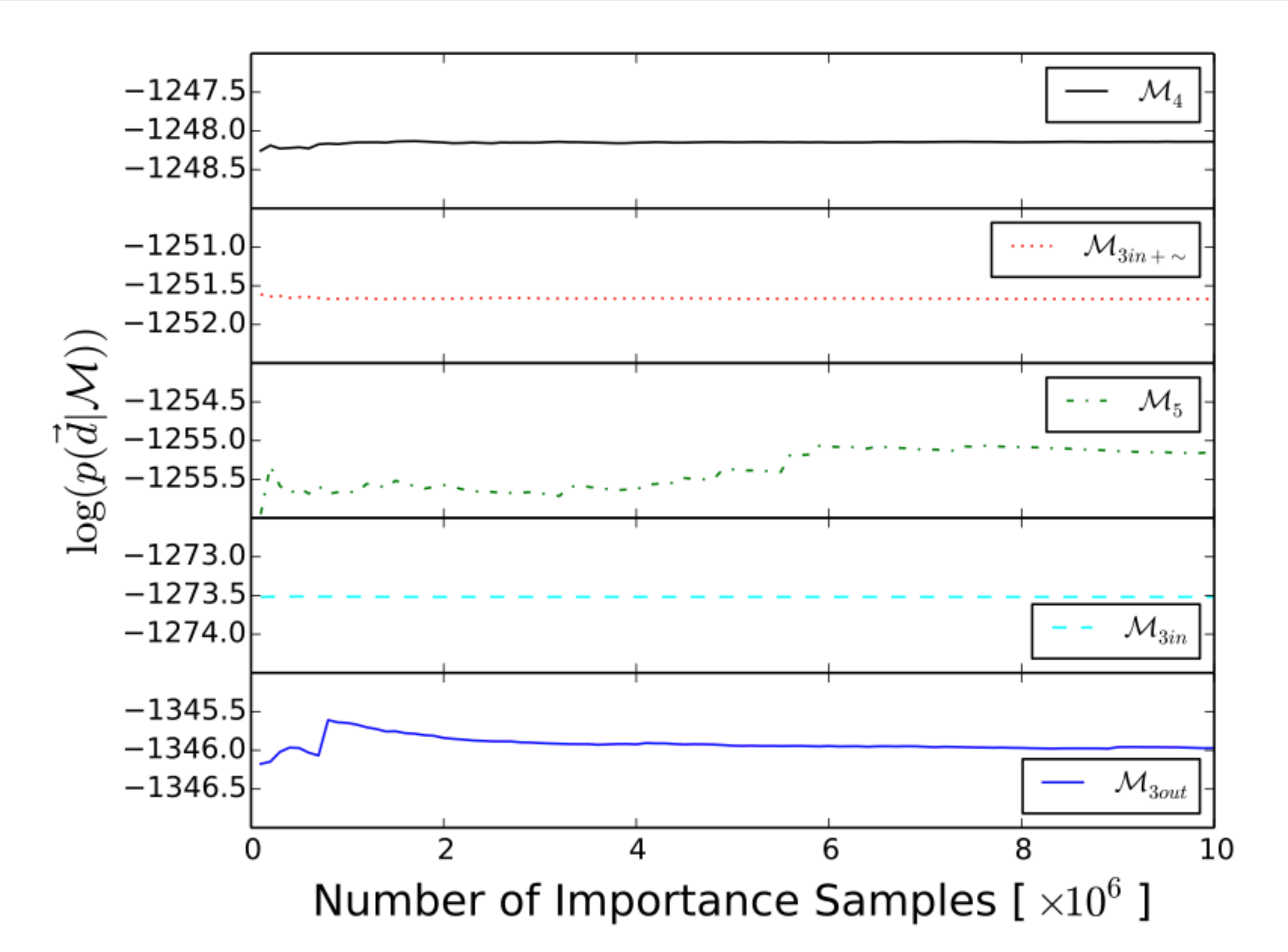
1. Embarrassingly parallel
2. Have a posterior sample? Already partway there!

Caveats:

1. Performance depends on chosen $g(\theta)$ or $g_\tau(\theta)$
2. Needs a robust value of f_{MCMC}

Importance Sampling (Gliese 876)

Seth Pritchard (undergrad at UT San Antonio)



Nelson+ 2016

Importance Sampling (Tutorial)

github.com/benelson/FML

Features:

- generate synthetic RVs of input planetary system
- MCMC with n-body model
- step-by-step importance sampling tutorial

Nelson+ 2016

Importance Sampling (Tutorial)

github.com/benelson/FML

Features:

- generate synthetic RVs of input planetary system
- MCMC with n-body model
- step-by-step importance sampling tutorial

Also check out John Boisvert's (UNLV) poster

Uncovering System Architectures Near 2:1 Resonance

More methods

Nested Sampling

Science: determining evidence for exomoons (Kipping+ 2013), functional form of eccentricity distribution (Kipping 2013), testing n-planets in RV observations (Brewer & Donovan 2015)

Publicly available code: Multinest (Feroz & Hobson 2008, Feroz 2009), DNest3/4 (Brewer+ 2010), Transdimensional MCMC (Brewer & Donovan 2015)

Geometric Path Monte Carlo

Science: testing n-planets in RV observations (Hou, Goodman, Hogg 2014)

Savage-Dickey Density Ratio

Specializes in comparing nested models with 1-2 parameter difference

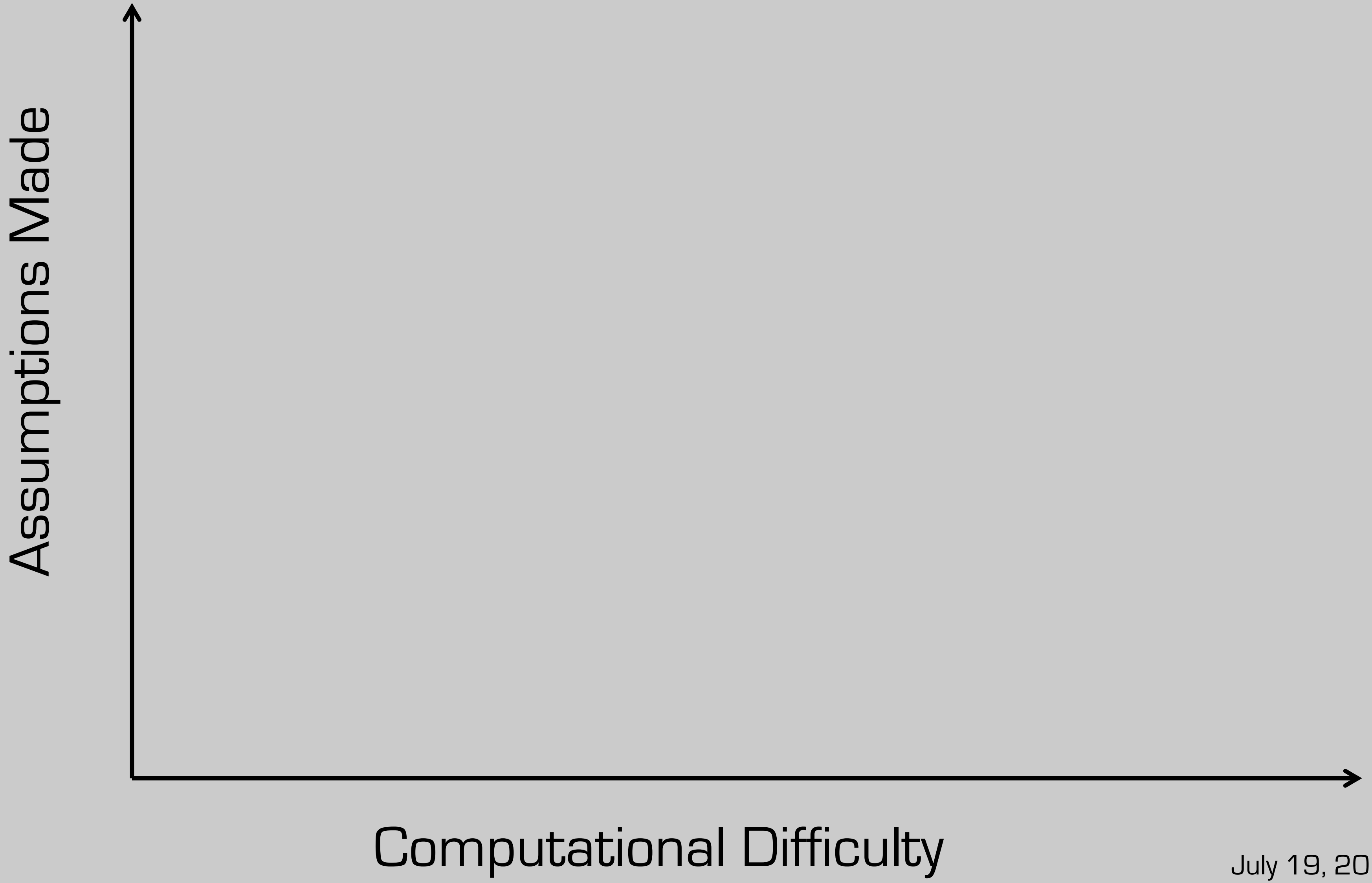
Science: Mass of Mars-sized Kepler-138b (Jontof-Hutter+ 2016)

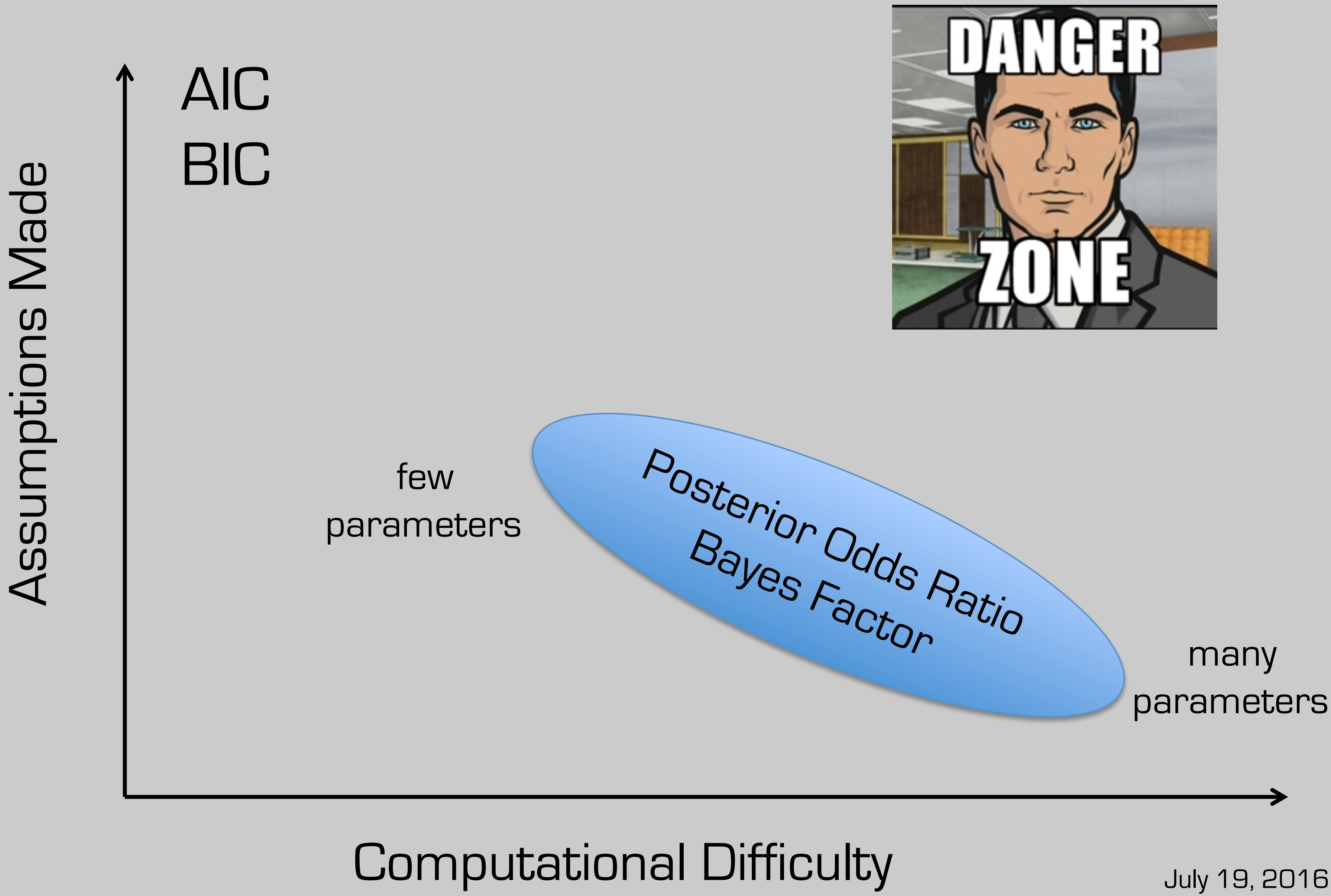
Outline

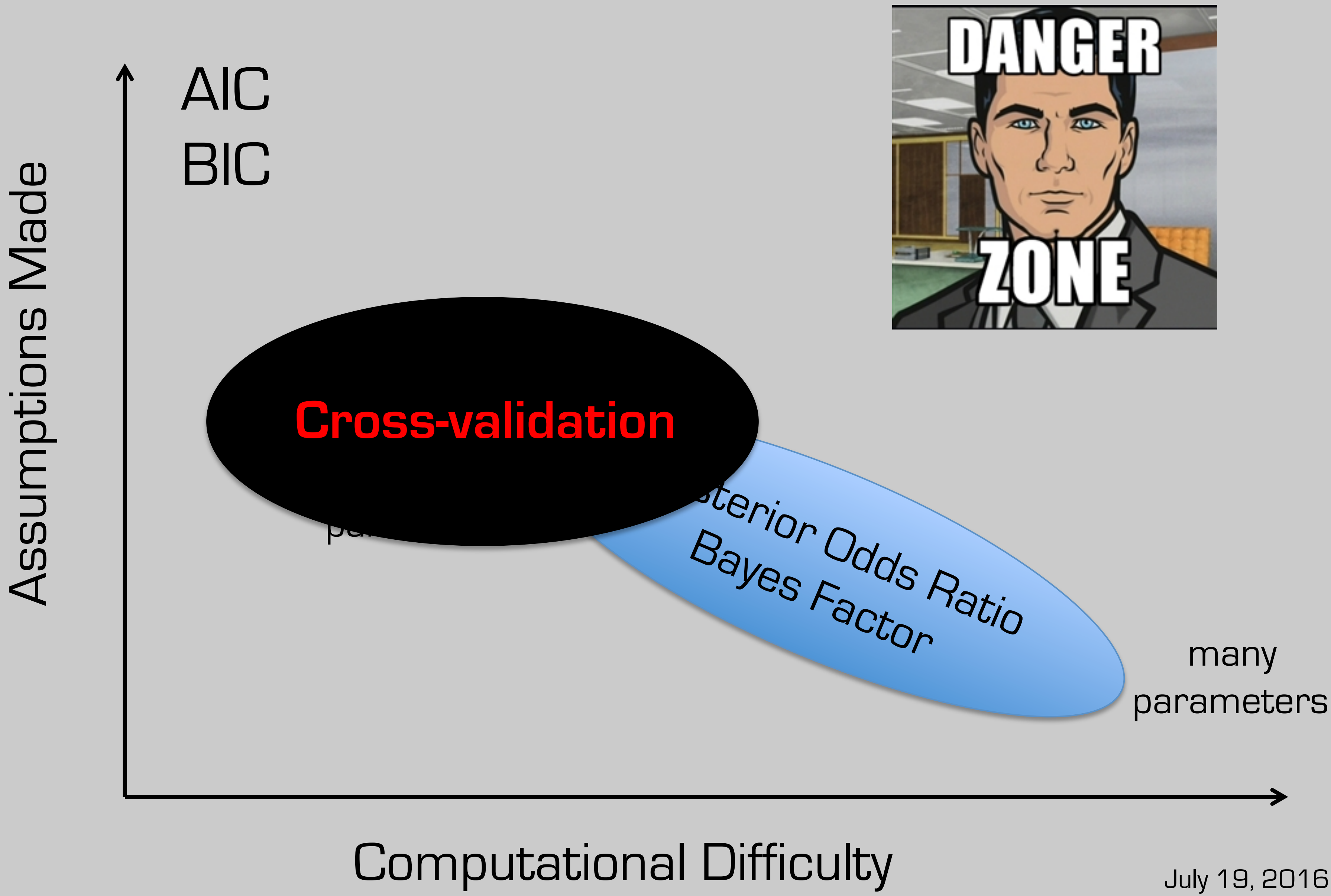
1. “Evidence”, Bayes factors, and decision making

2. How to efficiently compute “evidence”

3. Cross-validation as an alternative to BIC/BFs


Assumptions Made
Computational Difficulty


AIC
BIC
DANGER
ZONE
Assumptions Made
few parameters
Posterior Odds Ratio
Bayes Factor
many parameters
Computational Difficulty


DANGER
ZONE
AIC
BIC
Assumptions Made
Cross-validation
sterior Odds Ratio
Bayes Factor
many
parameters
Computational Difficulty

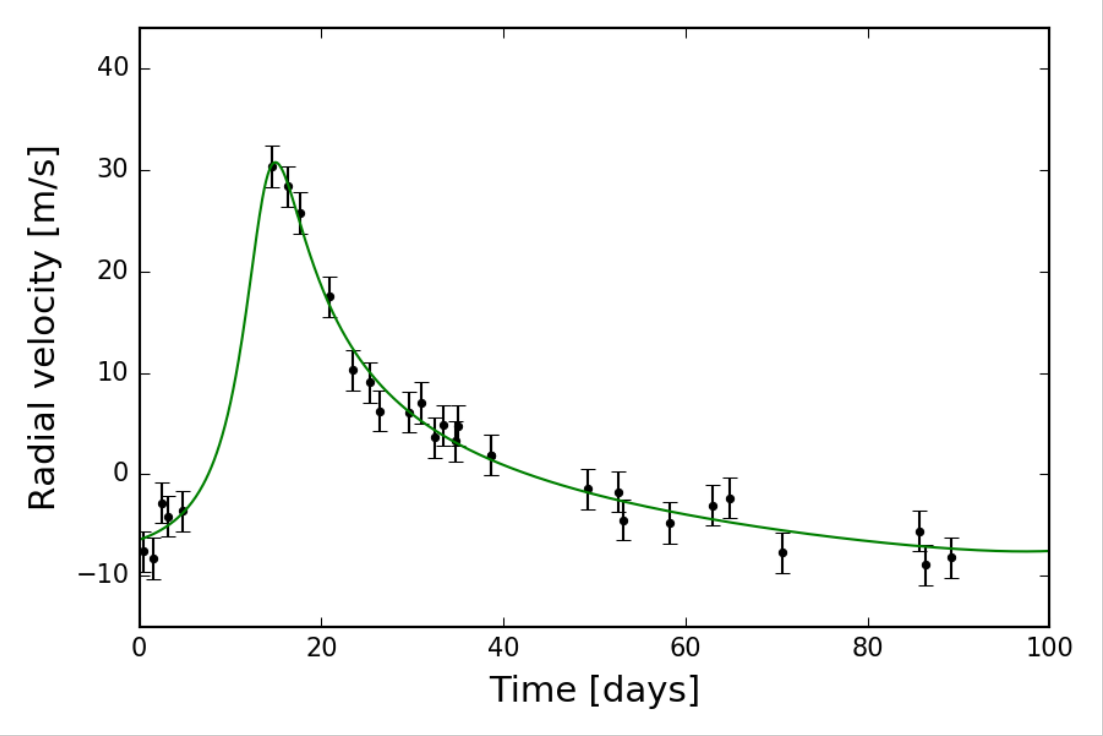
Cross-validation

cvl = 1.;

for (d in data){

get parameters $\theta_{(\mathtt{d})}$ that optimize on data WITHOUT d;

cvl $*= \mathtt{p}\left(\mathtt{d} \middle| \theta_{(\mathtt{d})}, \mathcal{M}\right)$; }



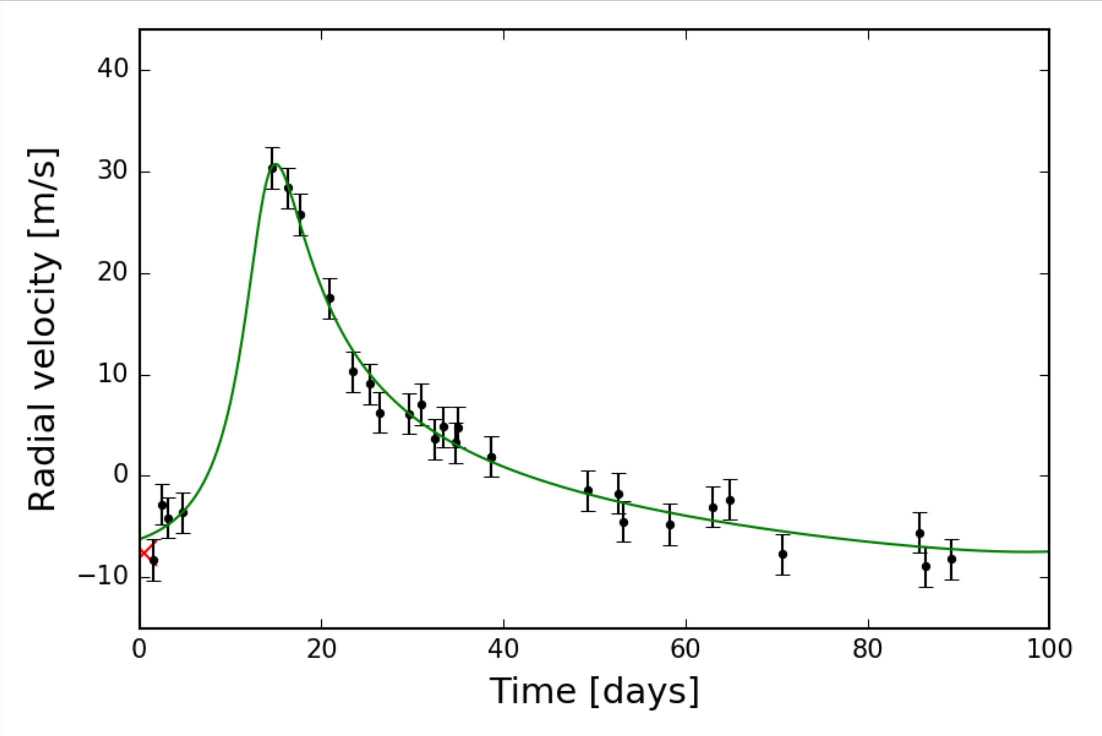
Cross-validation

cvl = 1.;

for (d in data){

get parameters $\theta_{(\mathtt{d})}$ that optimize on data WITHOUT d;

cvl *= $\mathrm{p}(\mathtt{d}|\theta_{(\mathtt{d})},\mathcal{M})$; }



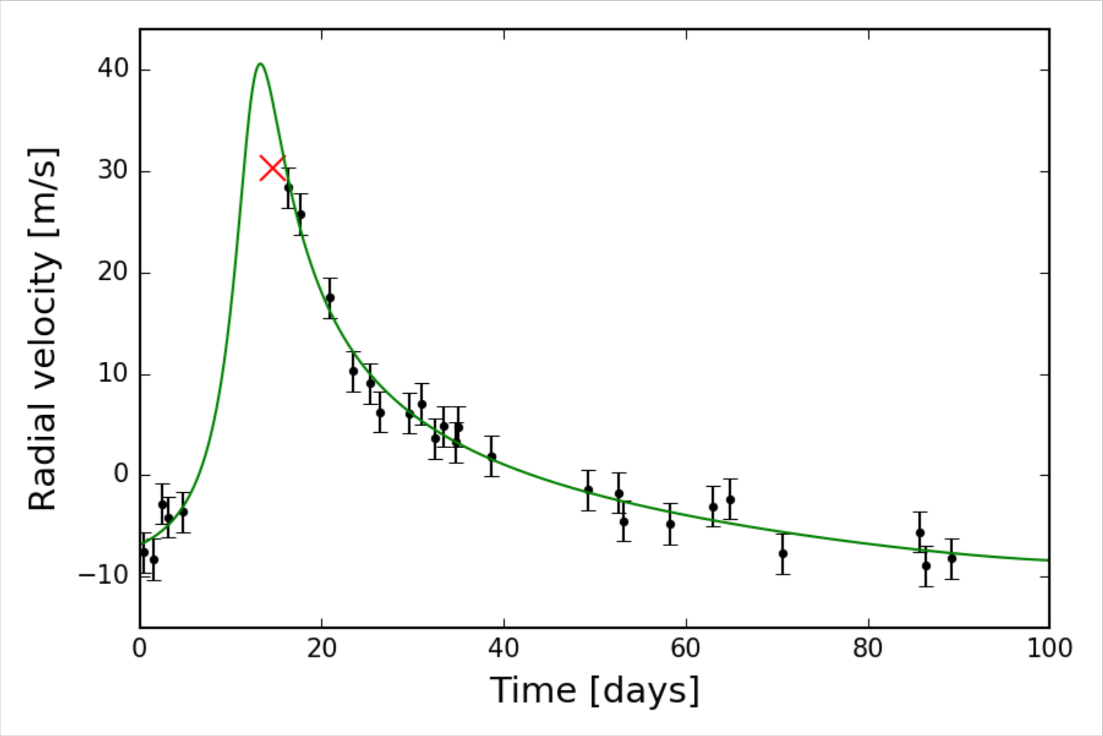
Cross-validation

cvl = 1.;

for (d in data){

get parameters $\theta_{(\mathtt{d})}$ that optimize on data WITHOUT d;

cvl $*= \mathrm{p}(\mathtt{d}|\theta_{(\mathtt{d})}, \mathcal{M})$; }



Cross-validation

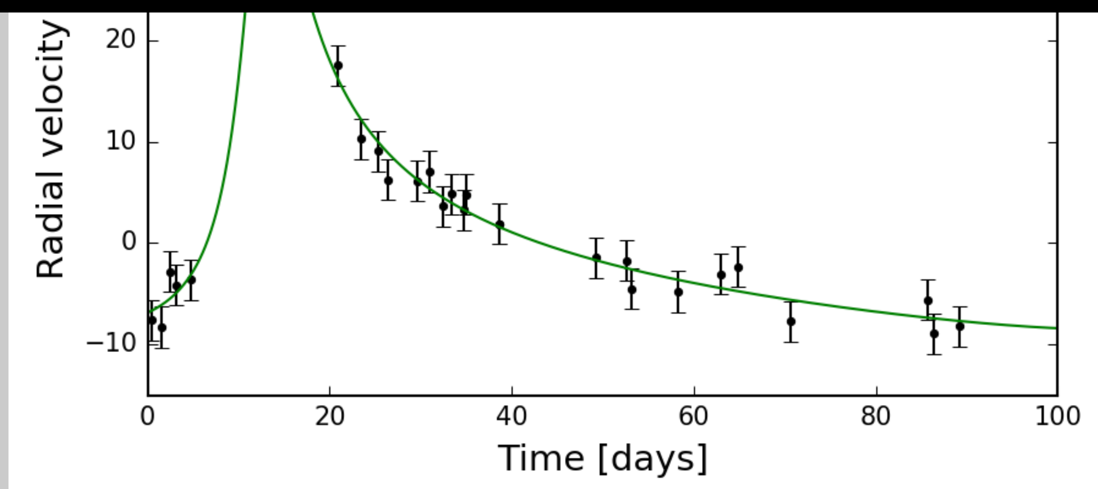
cvl = 1.;

for (d in data){

get parameters $\theta_{(d)}$ that optimize on data WITHOUT d;

cvl *=

The model with the largest cross-validation likelihood (cvl) is preferred.



Takeaway #3: General Recommendations

1. Do you need to make many rough decisions quickly (i.e. milliseconds)?
   AIC/BIC

2. Do you have decent computational resources and really understand your priors/utility?
   Bayes factor/posterior odds ratio

3. Is your problem somewhat in between?
   Cross-validation

Conclusions

model comparison ≠ model selection: how to decide depends on your utility

For 3+ parameter models, computing FML is hard. But it's an active problem in exoplanet research.

For tutorial on using importance sampling to compute FMLs: **github.com/benelson/FML**