Bayesian search for other Earths

Low-mass planets orbiting nearby M dwarfs

Mikko Tuomi
University of Hertfordshire, Centre for Astrophysics Research
Email: mikko.tuomi@utu.fi

Presentation, 19.4.2013

The Bayes' rule

$$\pi(\theta|m) = \frac{l(m|\theta)\pi(\theta)}{P(m)}, P(m) > 0 \tag{1}$$

- $\pi(\theta|m)$ is the posterior density of the parameter given the measurements.
- $\pi(\theta)$ is the prior density, the information on θ before the measurement m was made.
- $l(m|\theta)$ is the likelihood function of the measurement − the statistical model.
- $P(m)$ is constant, usually called the marginal integral, that makes sure that the posterior is a proper probability density.

$$P(m) = \int_{\theta\in\Theta} l(m|\theta)\pi(\theta)d\theta. \tag{2}$$

The Bayes' rule

The posterior density of parameters is actually always conditioned on the chosen model (i.e. the likelihood is defined according to some statistical model $\mathcal{M}$). Hence, the Bayes' rule becomes

$$\pi(\theta|m,\mathcal{M}) = \frac{l(m|\theta,\mathcal{M})\pi(\theta|\mathcal{M})}{P(m,\mathcal{M})} \tag{3}$$

Similarly, the marginal integral can be written as

$$P(m|\mathcal{M}) = \int_{\theta\in\Theta} l(m|\theta,\mathcal{M})\pi(\theta,\mathcal{M})d\theta. \tag{4}$$

Because it is actually the probability of getting the measurements given that the model $\mathcal{M}$ is the "correct one", it can be called the likelihood of the model. Sometimes it is also called the Bayesian evidence, though that might be slightly misleading.

The Bayes' rule

The Bayes' rule works the same way regardless of the number of measurements (or sets of measurements) available. For independent measurements $m_i, i = 1, ..., N$,

$$\pi(\theta|m_1, ..., m_N) = \frac{l(m_1, ..., m_N|\theta)\pi(\theta)}{P(m_1, ..., m_N)} = \frac{\pi(\theta)\prod_{i=1}^{N} l_i(m_i|\theta)}{P(m_1, ..., m_N)} \tag{5}$$

The best part about the above equation is that:

- The likelihood model(s) can be anything that can be expressed mathematically.
- The measurements can be anything (from different sources: RV, transit, astrometry, etc.).
- No assumptions are required about the nature of the probability densities of model parameters θ.
- Also, if N is large (etc.), the prior $\pi(\theta)$ can also be pretty much anything.

Bayesian model selection

The relative posterior probability of model $\mathcal{M}_i$ given measurement m can be written as

$$P(\mathcal{M}_i|m) = \frac{P(m|\mathcal{M}_i)P(\mathcal{M}_i)}{\sum_{j=1}^{M} P(m|\mathcal{M}_j)P(\mathcal{M}_j)}, \tag{6}$$

where $P(\mathcal{M}_i)$ is the prior probability of the ith model.

The probability of (the event of getting) the measurement given the ith model $P(m|\mathcal{M}_i)$ is in fact the marginal integral. With the model in the notation, it can be written as

$$P(m|\mathcal{M}_i) = \int_{\theta_i \in \Theta_i} l(m|\theta_i, \mathcal{M}_i)\pi(\theta_i|\mathcal{M}_i)d\theta_i \tag{7}$$

To compare the different models, all that is needed is the ability to calculate the integral in Eq. (7).

Likelihood functions

The common choice for a likelihood, arising from the famous central limit theorem, is the Gaussian density:

$$l(m|\theta) = l(m|\phi, \Sigma) = (2\pi)^{-N/2} \big|\Sigma\big|^{-1/2} \exp \left\{ -\frac{1}{2} \big[g(\phi) - m\big]^T \Sigma^{-1} \big[g(\phi) - m\big] \right\} \quad (8)$$

The matrix Σ is unknown too (one of the nuisance parameters), though usually it is assumed that $\Sigma = \sigma^2 I$, where $\sigma^2 \in \mathbb{R}_+$.

Function $g(\phi)$ is commonly called the model, though it is the whole function in Eq. (8) that actually is the model. The Gaussianity of the likelihood is simply a very common assumption in practice (perhaps too much so).

Priors

Priors are the only subjective part of Bayesian analyses – the rest consists of mindless repetitive tasks (i.e. computing).

- Prior probability densities (prior models): something has to be assumed always, a flat prior is still a prior (one that all frequentists using likelihoods assume).

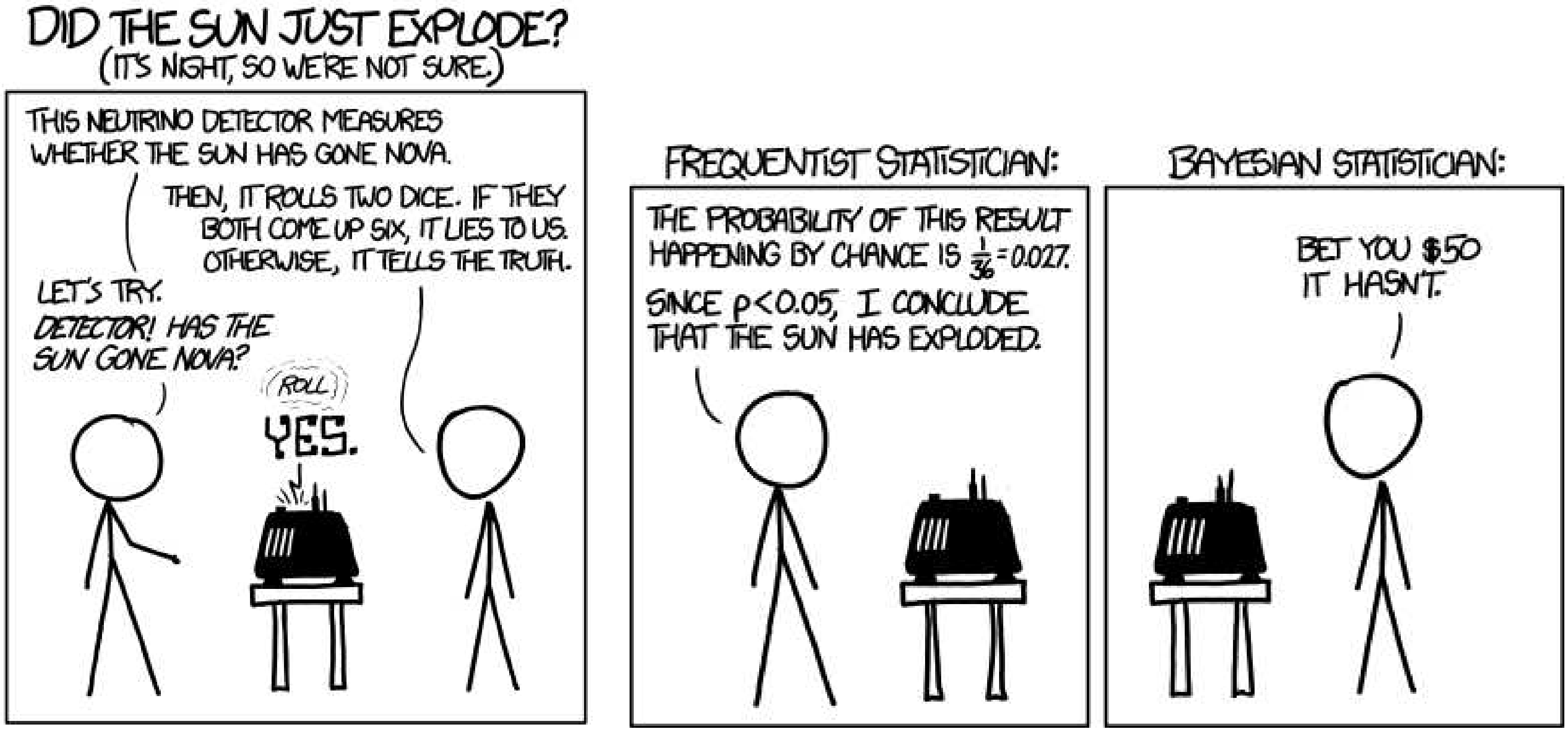
- Fixing parameters (e.g. fixing $e = 0$) corresponds to a delta-function prior.

- Flat priors on different parameterisations, e.g. using parameters (e, ω) vs. $(e \sin \omega, e \cos \omega)$, result in different results.

- Prior probabilities for models do not have to be equal.

- The collection of candidate models is also selected *a priori* – comparison of these models might be indistinguishable from comparison of different prior models.

Priors


DID THE SUN JUST EXPLODE?
(IT'S NIGHT, SO WE'RE NOT SURE.)
THIS NEUTRINO DETECTOR MEASURES WHETHER THE SUN HAS GONE NOVA.
THEN, IT ROLLS TWO DICE. IF THEY BOTH COME UP SIX, IT LIES TO US. OTHERWISE, IT TELLS THE TRUTH.
LET'S TRY. DETECTOR! HAS THE SUN GONE NOVA?
(ROLL)
YES.
FREQUENTIST STATISTICIAN:
THE PROBABILITY OF THIS RESULT HAPPENING BY CHANCE IS 1/36 = 0.027. SINCE p<0.05, I CONCLUDE THAT THE SUN HAS EXPLODED.
BAYESIAN STATISTICIAN:
BET YOU $50 IT HASN'T.

Bartlett's paradox

Assume that for model $\mathcal{M}_1$, θ is defined such that $\theta \in \Theta$.

Assume that the prior is of the form: $\pi(\theta) = ch(\theta)$ for all $\theta \in \Theta$ (zero otherwise).

Given any m, it follows that

$$\begin{aligned} P(\mathcal{M}_1|m) &\propto P(m|\mathcal{M}_1)P(\mathcal{M}_1) \\ &= cP(\mathcal{M}_1)\int_{\theta\in\Theta} l(m|\theta, \mathcal{M}_1)h(\theta)d\theta \end{aligned} \tag{9}$$

If model $\mathcal{M}_0$ is the "null hypothesis such that e.g. $\theta = 0$, and we assume that $h(\theta) = 1$ (because it can be chosen to be anything anyway) the model probability $P(\mathcal{M}_1|m) \propto c$ and $P(\mathcal{M}_0|m) \propto c^{-1}$, which means that given a sufficiently small c, the null hypothesis can never be rejected regardless of the measured m!

A paradox?

Prior range

It is always possible to define the model parameters in such a way that the prior range, i.e. the space Θ integrates to unity.

This corresponds to a linear transformation (or a chance of unit system):

$$\theta \rightarrow \theta', \text{ s.t. } \theta' = a\theta + b. \tag{10}$$

For instance, if the radial velocity amplitude $K \in [0, 1000]$ ms^{-1}, we can choose $K' = 10^{-3}K$ and the parameter space of K' indeed integrates to unity.

This does not affect the analyses because we can use K or K' in the likelihood mean $g(\theta)$ to receive the exact same results. It does, however, change the units of dK nicer.

Generally, any linear transformation of the parameters does not (cannot) affect the results. What about non-linear transformations?

Non-linear transformation

Assume a change of variables: $\theta \rightarrow \theta'$. This changes the prior density according to

$$\pi(\theta') = \pi(\theta) \left| \frac{d\theta}{d\theta'} \right|. \tag{11}$$

For instance, when performing periodogram analyses, is it implicitly assumed that the prior density of the period parameter is flat – in the frequency space.

A transformation $P^{-1} \rightarrow P$ changes the priors such that a flat prior on P^{-1} is equivalent to $\pi(P) \propto P^{-2}$ on P.

Therefore, periodogram analyses underestimate the prior probabilities of signals at larger periods *a priori*.

Priors indeed are everywhere.

Posterior sampling

Monte Carlo methods can be used efficiently to estimate the posterior densities and the marginal integral in Eq. (4). First, however, it is necessary to draw a sample from the (unknown) posterior density of the model parameters given the measurements: Markov chain Monte Carlo.

Several posterior sampling (MCMC) methods exist:

- Gibbs sampling algorithm.
- Metropolis-Hastings algorithm.
- Adaptive Metropolis algorithm.
- ...

Out of these, the adaptive Metropolis algorithm (Haario et al. 2001) is reasonably reliable: converges rapidly to the posterior density and does not usually care about the initially selected proposal density nor the initial state.

Importance sampling

The marginal integral can be estimated using the importance sampling method. Estimate $\widehat{P}(m)$ can be calculated using

$$\widehat{P}(m) = \frac{\sum_{i=1}^{N} w_i l(m|\theta_i)}{\sum_{i=1}^{N} w_i}, \tag{12}$$

where $w_i = \pi(\theta_i)\pi^{\star}(\theta_i)^{-1}$ and $\pi^{\star}(\theta_i)$ is called the importance sampling function that can be selected rather freely. N is the size of the sample drawn from $\pi^{\star}$.

A simple choise would be the posterior density. Setting $\pi^{\star}(\theta_i) = \pi(\theta_i|m)$ leads to the harmonic mean estimate $\widehat{P}_{HM}$ defined as

$$\widehat{P}_{HM}(m) = N \left[\sum_{i=1}^{N} l(m|\theta_i)^{-1} \right]^{-1}. \tag{13}$$

However, this estimate has poor convergence properties. Other more complex estimates should be preferred.

Importance sampling

Another simple choise for the importance sampling function is a truncated posterior, where

$$\pi^{\star}(\theta_i) = (1-\lambda)\pi(\theta_i|m) + \lambda\pi(\theta_{i-h}|m). \tag{14}$$

Parameter h is some small integer such that θ_{i-h} is independent of θ_i and λ is a small positive number.

The resulting estimate, the “truncated posterior mixture“ (TPM) estimate is

$$\hat{P}_{TPM}(m) = \Bigg[\sum_{i=1}^{N} \frac{l_i\pi_i}{(1-\lambda)l_i\pi_i + \lambda l_{i-h}\pi_{i-h}}\Bigg]\Bigg[\sum_{i=1}^{N} \frac{\pi_i}{(1-\lambda)l_i\pi_i + \lambda l_{i-h}\pi_{i-h}}\Bigg]^{-1}. \tag{15}$$

In practice, it appears to converge rapidly but more work is needed to assess its properties. Especially, what would be the best choises of h and λ.

Is the best model good enough?

With the aforementioned tools, the best model out of the M different models is available. But how do we know that this “best” model describes all the information in the measurement?

The “Bayesian model inadequacy criterion” (BMIC) can help answering this question.

- Assume that the best model $\mathcal{M}$ has been found.
- Assume that there are several independent measurements or sets of measurements $m_i, i = 1, ..., N$.
- Model $\mathcal{M}$ describes the measurement m_i using parameter θ.
- Assume that another model $\mathcal{N}$ describes the measurement m_i with parameter θ_i that differ from one another for $i = 1, ..., N$. But model $\mathcal{N}$ has the same likelihood function as $\mathcal{M}$.

Bayesian model inadequacy criterion (BMIC)

With the above assumptions, it follows that:

$$P(m_1, ..., m_N|\mathcal{N}) = \prod_{i=1}^{N} P(m_i|\mathcal{M}) \tag{16}$$

Now, we compare the models $\mathcal{M}$ and $\mathcal{N}$ and say that the model $\mathcal{M}$ is an inadequate description of the measurements if its probability is less than some threshold probability s. Therefore, the model $\mathcal{M}$ is an inadequate description of the measurements if

$$B(m_1, ..., m_N) = \frac{P(m_1, ..., m_N)}{\prod_{i=1}^{N} P(m_i)} < \frac{s}{1-s}. \tag{17}$$

If the condition in Eq. (17) is satisfied, the model is an inadequate description of the measurements with a probability of $1 - s$.

Interpretation of the model inadequacy criterion

The Kullback-Leibler divergence is a measure of "difference" between two probability density functions $f(\theta)$ and $g(\theta)$ of random variable θ and is defined as

$$D_{KL}\{f(\theta)||g(\theta)\} = \int_{\theta\in\Theta} f(\theta) \log \frac{f(\theta)}{g(\theta)} d\theta. \tag{18}$$

It is not symmetric, and generally $D_{KL}\{f(\theta)||g(\theta)\} \neq D_{KL}\{g(\theta)||f(\theta)\}$.

When using the prior and posterior densities, it can be interpreted as the "information gain" of moving from the prior to the posterior or "information loss" of moving from the posterior back to the prior. With this interpretation, the information loss is simply

$$D_{KL}\{\pi(\theta)||\pi(\theta|m)\} = \int_{\theta\in\Theta} \pi(\theta) \log \frac{\pi(\theta)}{\pi(\theta|m)} d\theta. \tag{19}$$

Interpretation of the BMIC

In terms of information loss, the Bayes factor $B(m_1, ..., m_N)$ can be written as

$$\log B(m_1, ..., m_N) = D_{KL}\{\pi(\theta)||\pi(\theta|m_1, ..., m_N)\}$$
$$- \sum_{i=1}^{N} D_{KL}\{\pi(\theta)||\pi(\theta|m_i)\}. \tag{20}$$

In other words, the information loss of not using any of the measurements minus the information losses of each measurement is equal to the logarithm of the "model inadequacy".

Therefore, the Bayesian model inadequacy criterion is naturally related to the information in the data in the above manner.

Interpretation of the BMIC

If it holds that $B(m_1, ..., m_N) \geq 1$, we cannot say that there is any model inadequacy. However, this implies that

$$D_{KL}\{\pi(\theta)||\pi(\theta|m_1, ..., m_N)\} \geq \sum_{i=1}^{N} D_{KL}\{\pi(\theta)||\pi(\theta|m_i)\}. \tag{21}$$

This can be interpreted as saying that there is more information in the combined set of data than there are in the individual measurements.

But it is only the case if the model is good enough in the sense of Eq. (17) with $s = 1/2$.

BMIC for nested models

If each measurements m_i can be described the best using some model $\mathcal{M}_i$ nested in the full model $\mathcal{M}$, it follows that

$$B(m_1, ..., m_N|\mathcal{M}) \geq \frac{P(m_1, ..., m_N|\mathcal{M})}{\prod_{i=1}^{N} P(m_i|\mathcal{M}_i)} \geq \frac{s}{1-s}. \tag{22}$$

If the last inequality holds for the nested “submodels”, the full model cannot be found inadequate either.

A simple application of this result would be that the full model describes the combined radial velocities of some target well but some nested submodels are better for data from certain instruments because of their greater noise levels or shorter baselines (a very common situation in practice).

Also, it holds that selecting $s = 1/2$,

$$B(m_i, m_j) \geq 1, \text{ forall } i, j \Rightarrow B(m_1, ..., m_N) \geq 1 \tag{23}$$

Bayes' rule and dynamical information

In case of detections of exoplanet systems, there is a very useful additional source of information – the Newtonian (or post-Newtonian if necessary) mechanics.

Because we cannot expect to detect an unstable planetary system, we can say that the prior probability of detecting something unstable is zero (at least negligible). Hence:

$$\pi(\theta|m, \mathcal{S}) = \frac{l(m, \mathcal{S}|\theta)\pi(\theta)}{P(m, \mathcal{S})} = \frac{l(m|\mathcal{S}, \theta)l(\mathcal{S}|\theta)\pi(\theta)}{P(m, \mathcal{S})} \tag{24}$$

Because the laws of gravity do not depend on what we measured but the results we obtain depend on them. We call $\mathcal{S}$ the "dynamical information".

But what is the likelihood function of dynamical information, $l(\mathcal{S}|\theta)$?

Bayes' rule and dynamical information

The approximated Lagrange stability criterion for two subsequent planets (Barnes & Greenberg, 2006) is defined as

$$\alpha^{-3}\left(\mu_1 - \frac{\mu_2}{\delta^2}\right)\left(\mu_1\gamma_1 + \mu_2\gamma_2\delta\right)^2 > 1 + \mu_1\mu_2\left(\frac{3}{\alpha}\right)^{4/3}, \tag{25}$$

where $\mu_i = m_i M^{-1}$, $\alpha = \mu_1 + \mu_2$, $\gamma_i = \sqrt{1 - e_i^2}$, $\delta = \sqrt{a_2/a_1}$, $M = m_\star + m_1 + m_2$, e_i is the eccentricity, a_i is the semimajor axis, m_i is the planetary mass, and $m_\star$ is stellar mass.

We simply set $l(\mathcal{S}|\theta) = c$ if the criterion is satisfied and $l(\mathcal{S}|\theta) = 0$ otherwise (we call the set of stable orbits in the parameter space $B \subset \Theta$).

Alternatively, we could use a simpler form that only prevents orbital crossings of the planets. Note that the stellar mass is one of the parameters.

The above does not take e.g. resonances into account, and is only a rough approximation. Can we do better?

Posterior sampling with dynamics

A posterior sample from a MCMC analysis: $\exists \theta_i \sim \pi(\theta|m), i = 1, ..., K$.

Each θ_i as an initial state of orbital integration: K chains of N vectors with $\theta_i^j, j = 1, ..., N$, and $\theta_i^j = \theta_i(t_j)$ and t_j is some moment between $t_0 = 0$ and the duration of the integrations $t_N = T$.

Hence, we can approximate the posterior probability of finding $\theta \in I_l \subset \Theta$ of dynamical information and data for each n-interval I_l as

$$\begin{aligned} P(\theta \in I_l|\mathcal{S}, d) \qquad &\approx \frac{1}{K}\sum_{i=1}^{K} P(\theta \in I_l|\mathcal{S}, \theta_i) \\ &\approx \frac{1}{KN}\sum_{i=1}^{K}\sum_{j=1}^{N} \mathbf{1}_l(\theta_i^j)\mathbf{1}(\theta_1^j), \end{aligned} \tag{26}$$

where

$$\mathbf{1}_l(\theta) = \begin{cases} 1 \text{ if } \theta \in I_l \\ 0 \text{ otherwise} \end{cases} \text{ and } \mathbf{1}(\theta) = \begin{cases} 1 \text{ if } \theta \in B \\ 0 \text{ otherwise} \end{cases}$$

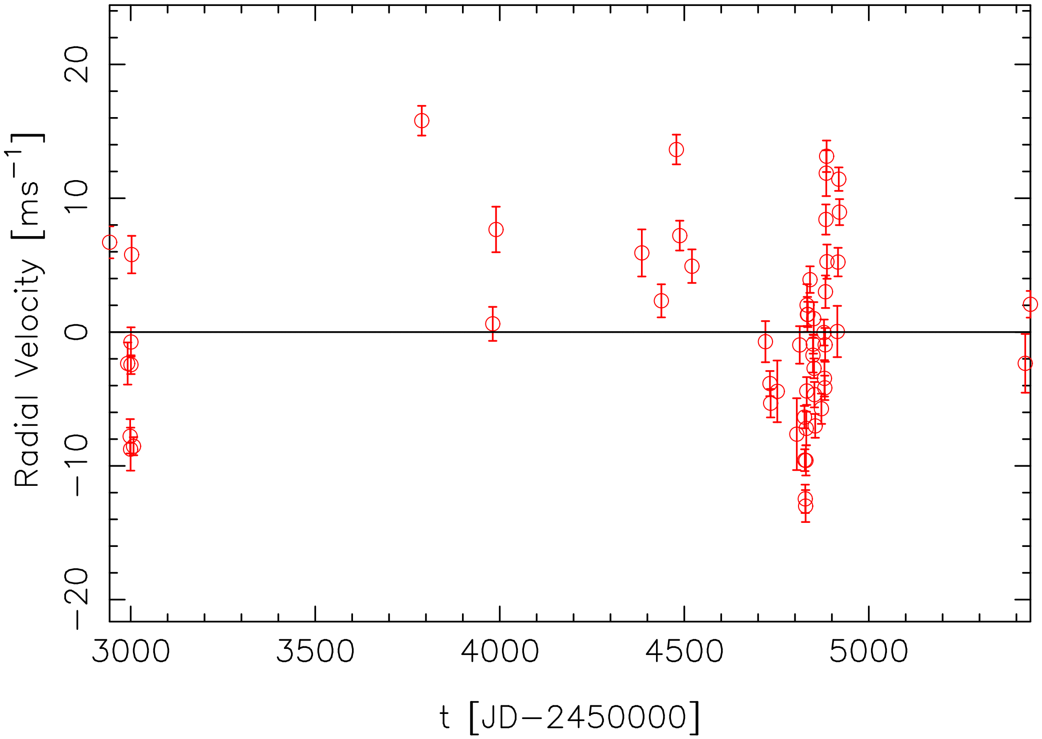
How to detect a signal (planet) from RVs?

There are k periodic signals in the radial velocities if

- $P(\mathcal{M}_k|m) > \alpha P(\mathcal{M}_{k-j}|m)$ for a selected threshold $\alpha > 1$ and for all $j \geq 1$.
- The radial velocity amplitudes of all signals are statistically distinguishable from zero, i.e. their BCSs (*) do not overlap with zero for a selected (but sufficiently high) threshold $\delta \in [0, 1]$.
- All periodicities get well constrained from above and below.
- The planetary system corresponding to the solution (parameters k and θ) is stable.

$$(*) \text{ Set } \mathcal{D}_\delta = \Big\{ \theta \in C \subset \Theta : \int_{\theta \in C} \pi(\theta|m) = \delta, \pi(\theta|m)|_{\theta \in \partial C} = c \Big\}. \qquad (27)$$

Periodic variations in XXXX RVs



Are there planetary signals in these velocities?

Moreover, how many signals, and what are the corresponding orbital parameters?

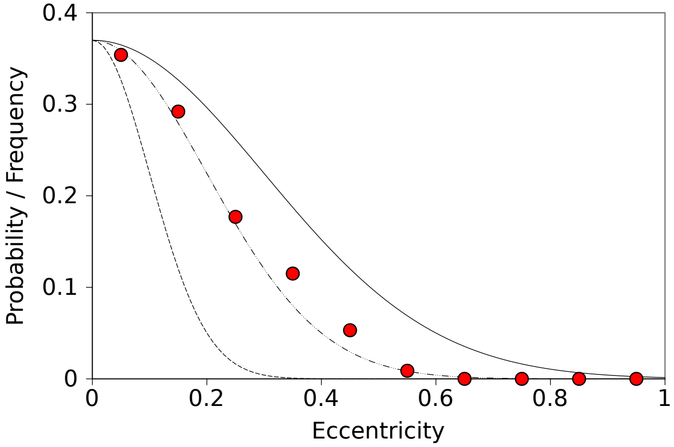
Bayesian model comparisons and posterior samplings are the perfect tool for answering these questions.

But what about priors?

Eccentricity prior

When searching for low-mass planets around nearby M dwarfs, can we use informative priors?

Selecting all low-mass ($m_p \leq 0.1\mathrm{M}_{\mathrm{Jup}}$) planets in the Extrasolar Planets Encyclopaedia gives a distribution of eccentricities or such planets.



Eccentricity distribution of low-mass planets. The didstribution appears to peak ata zero, but that could be a bias caused by data analysis methods.

Gaussian curves of $\mathcal{N}(0, 0.1^2)$, $\mathcal{N}(0, 0.2^2)$, and $\mathcal{N}(0, 0.3^2)$.

The second one of these curves seems to coincide with the data.

Modelling radial velocity noise

- Usually RV data is binned (somehow) by calculating the average of few velocities within an hour or so.
- Binning will always result in loss of information (because the transformation called “binning” is not a bijective mapping).
- Instead, model the noise as realistically as possible.
- Possibility to have the “binning” procedure as a part of the statistical model, which enables comparisons of different procedures.

Modelling radial velocity noise

An effective analogue of “binning” is e.g. a noise model with moving average (MA) component. This statistical model can be written as

$$m_i = f_k(t_i) + \epsilon_i + \sum_{j=1}^{p} \phi_j \epsilon_{i-j}, \tag{28}$$

where measurement m_i at epoch t_i is modelled using the function f_k and some convenient white noise component ϵ_i.

The analyses of HARPS radial velocities indicate, that this noise model is much better than pure white noise – and information is not lost if the MA coefficients ϕ_i are selected conveniently (or even better, free parameters).

Planets around M dwarfs

Search for periodic signals in RV data is basically a difficult task because the posterior is highly multimodal in the period space.

Because the noise is not white, periodogram analyses give biased results - spurious powers and lack of them where they should be.

Solution: global search of the period space using temperate samplings.

1) Draw a sample from $\pi(\theta|m)^{\beta}$ with $\beta \in [0, 1]$.

2) Calculate $\log \pi(\theta|m)$ as a function of this chain.

3) Plot the chain as a function of period and find the global maximum (and local ones).

Tuomi et al. in preparation

Definition of a planet candidate

All RV planets are only “planet candidates” because only minimum mass ($m_p \sin i$) is available.

1) Detection criteria are satisfied, including the dynamical stability criterion.

2) Not a false positive: supporting evidence from at least two independent data sets from different instruments.

3) No periodicities in the activity data at orbital periods or the first harmonics/aliases of them.

4) The significance of the signal increases (on average) when adding more measurements to the data set.

Thank You for Your attention

E-mail: mikko.tuomi@utu.fi

Web: users.utu.fi/miptuom