Multidata inverse problems and Bayesian solution methods in astronomy

Mikko Tuomi

University of Turku, Tuorla Observatory, Department of Physics and Astronomy, Väisäläntie 20, FI-21500, Piikkiö, Finland

Abstract

Statistical analyses of measurements that can be described by statistical models are of essence in astronomy and in scientific inquiry in general. The sensitivity of such analyses, modelling approaches, and the consequent predictions, is sometimes highly dependent on the exact techniques applied, and improvements therein can result in significantly better understanding of the observed system of interest. Particularly, optimising the sensitivity of statistical techniques in detecting the faint signatures of low-mass planets orbiting the nearby stars is, together with improvements in instrumentation, essential in estimating the properties of the population of such planets, and in the race to detect Earth-analogs, i.e. planets that could support liquid water and, perhaps, life on their surfaces. We review the developments in Bayesian statistical techniques applicable to detections planets orbiting nearby stars and astronomical data analysis problems in general. We also discuss these techniques and demonstrate their usefulness by using various examples and detailed descriptions of the respective mathematics involved. We demonstrate the practical aspects of Bayesian statistical techniques by describing several algorithms and numerical techniques, as well as theoretical constructions, in the estimation of model parameters and in hypothesis testing. We also apply these algorithms to Doppler measurements of nearby stars to show how they can be used in practice to obtain as much information from the noisy data as possible. Bayesian statistical techniques are powerful tools in analysing and interpreting noisy data and should be preferred in practice whenever computational limitations are not too restrictive.

Department of Physics and Astronomy
University of Turku
Turku, Finland

Supervised by

Prof. Harry Lehto
Prof. Esko Valtaoja
University of Turku

Reviewed by

Dr. Alexis Brandeker
Stockholms Universitet

Dr. Johanna Tamminen
Finnish Meteorological Institute

Sarja A I 470
ISBN 978-951-29-5505-3 (PDF)
ISSN 0082-7002

Contents

To Maikki

List of articles

This thesis consists of four articles (Tuomi, 2011, 2012; Tuomi et al., 2011; Tuomi & Jones, 2012) and an introductory text describing the methodology behind these articles in greater detail than is possible to present in standard astronomical publications.

I Tuomi, M. 2011. Bayesian re-analysis of the radial velocities of Gliese 581. Evidence in favour of only four planetary companions. Astronomy & Astrophysics, 528, L5.

II Tuomi, M., Pinfield, D., & Jones, H. R. A. 2011. Application of Bayesian model inadequacy criterion for multiple datasets to radial velocity models of exoplanet systems. Astronomy & Astrophysics, 532, A116.

III Tuomi, M. 2012. Evidence for nine planets orbiting HD 10180. Astronomy & Astrophysics, 543, A52.

IV Tuomi, M. & Jones, H. R. A. 2012. Probabilities of exoplanet signals from posterior samplings. Astronomy & Astrophysics, 544, A116.

Articles are listed in the order of appearance.

1. Introduction

Detections of extra-solar planets orbiting, and exoplanet systems around, nearby stars has been a hot topic in astronomy during the last two decades. Ever since the first discovery of such a planet in 1992 (Wolszczan & Frail, 1992), and especially, after the discovery of the first exoplanet orbiting a Solar-type star in 1995 (Mayor & Queloz, 1995), several hundred planets have been found orbiting a variety of different stellar targets in the Solar neighbourhood. References to the latest developments in exoplanet searches can be found in e.g. The Extrasolar Planets Encyclopaedia[1] (Schneider et al., 2011) and Exoplanet Orbit Database[2] (Wright et al., 2011a). Furthermore, due to improvements in instrumentation and observational facilities, and rapid accumulation of data due to several ongoing surveys, the rate of such discoveries can only be expected to increase as a function of time.

Detections of planets indirectly by observing their effects on the stellar radiation is difficult and typically requires applications of sophisticated statistical techniques in order to distinguish the planetary fingerprints from various sources of noise, such as photon noise, stellar activity, and Earth's atmosphere; other types of variation mimicking Keplerian signals, such as daily and annual constraints to the visibility of the star in the sky, stellar activity cycles and rotation coupled with starspots and magnetic phenomena; and biases, such as instrument instabilities and additional biases caused by the fact that the statistical model used to describe the data might be suboptimal. Furthermore, planetary occurrence rates have been shown to increase dramatically as a function of decreasing mass (e.g. Howard et al., 2012; Bonfils et al., 2013; Dressing & Charbonneau, 2013), which means that a large population of planets remains at or below the current detection limits of planet surveys and detecting such planets is not only a matter of instrumentation and observational facilities, but to a great extent of optimising the detection techniques and obtaining as much information from the valuable measurements as possible. As such low-mass planets are also among the most interesting ones in astrobiological terms because they could host biospheres on their rocky surfaces (e.g. Anglada-Escudé et al., 2013; Tuomi et al., 2013a,b; Tuomi & Anglada-Escudé, 2013), statistical techniques, together with state-of-the-art instrumentation, have a key role in the searches

[1] `exoplanet.eu`
[2] `exoplanets.org`

for other Earths orbiting the nearby stars.

Implementation of more efficient statistical techniques has been responsible for several recent observations of planets with the Doppler spectroscopy method such as the detection of a potential habitable-zone super-Earth orbiting HD 40307 (Tuomi et al., 2013a) and a diverse system of low-mass planets around GJ 667C (Anglada-Escudé et al., 2013) in the stellar habitable zone. Together with improvements in data reduction procedures and statistical modelling (Anglada-Escudé & Butler, 2012; Baluev, 2012; Tuomi et al., 2013b), the respective improvements in the sensitivity of exoplanet surveys have come with almost non-existent cost compared to the potentially considerable investments in instrumentation that would have been required to produce the same results had the traditional statistical techniques been relied on.

The greatest improvements in the statistical techniques are due to a paradigm shift from the so-called frequentist interpretation of probabilities to a Bayesian one. The former is based on the idea that a probability of an event occurring in an (scientific) experiment reflects the frequency of how often such an event happens out of all possible events that could have happened if the experiment was repeated infinitely many (or sufficiently many) times. The corresponding statistical analysis techniques have relied on the pioneering work of Pearson (1901) and Neyman & Pearson (1928) and have been used, with suitable improvements and modifications, to detect a majority of planets around nearby stars. The improvements have been made regarding the search of periodic signals and are essentially based on Fourier analysis techniques that take advantage of studying the data in frequency domain instead of the time domain (Lomb, 1976; Scargle, 1982; Cumming, 2004). However, a common strategy of assessing the significance of periodic signals detected using these periodogram methods typically rely on resampling techniques such as Bootstrapping (e.g. Efron, 1979) that are an attempt of artificially creating a statistically representative sample of data sets with statistical properties approximately equal to those of the one that has been detected to enable estimating what were the chances of obtaining the result out of several trials. For obvious reasons, in particular, because such a sample of data sets does not exist, and all experiments cannot be repeated arbitrarily many times, the latter, Bayesian, interpretation of probability is a much more practical in addition to being more solidly based on probability theory.

In the Bayesian framework, the goal is to calculate probabilities for dif-

ferent hypotheses give the data that was observed. This refers directly to the law of conditional probabilities, or the Bayes' rule. Assuming that the data consists of random numbers drawn from some underlying (and unknown) statistical distribution, the key feature of Bayesian techniques is to make assumptions regarding the nature of this distribution, i.e. formulating statistical models, and to calculate probabilities for different events, e.g. that the data are drawn from a Gaussian distribution with mean μ and variance of σ^2 or that the value of μ is in the interval $[a, b]$, given the data and the model. Similarly, it is possible to calculate probabilities of a given hypothesis, or statistical model, being a good description of the data with respect to other hypotheses. This latter process is referred to as Bayesian model selection. There is a vast literature discussing the problems, advantages, and implementation of various Bayesian methods (e.g. Green, 1995; Kass & Raftery, 1995; Spiegelhalter et al., 2002, and references therein) and such techniques have been applied to exoplanet detections during the recent years (e.g. Ford, 2005, 2006; Feroz et al., 2011; Gregory, 2011; Loredo et al., 2012; Tuomi, 2012; Tuomi et al., 2013a,b).

In this thesis, we discuss the statistical methods that are being used to detect planetary signals of low-mass companions to stars in the Solar neighbourhood. However, instead of giving direct ready-to-use recipes for analysing such data that comes in various forms, we describe the basic statistical techniques that can be used to obtain as much of the important information from the valuable measurements as possible with logical consistency and mathematical rigour. While we explain the rationales behind the various statitical tehcniques and computational methods as simply as possible, we also attempt to express them with mathematical precision that is sufficient for replicating the results we present in the various applications of the methods. Should we fail to do so, the reader is encouraged to contact the author and report such shortcomings. Throughout this thesis, with few exceptions, we concentrate on the Doppler spectroscopy method used to detect a large fraction of planet candidates (Schneider et al., 2011; Wright et al., 2011a), although the methods we describe are mostly completely general and can be readily applied to any detection technique, and in fact, to any statistical data analysis problem in astronomy and beyond.

In particular, we discuss the various statistical techniques based on the Bayes' rule of conditional probabilities that are, in many cases, superior to the classic frequentist techniques whose applicability is more often that not very limited. For this reason, this thesis can also be seen as consisting of

criticism of the traditional solution methods, such as statistical hypothesis testing methods or standard computations of point estimates and correlation coefficients. This criticism is only partially intentional. Although the frequentist solution methods do have their place in the toolbox of a professional statistician, their assumptions and the resulting restrictions have to be understood in order to be able to use them properly. Bayesian statistical techniques are simply more general and applicable to a wider variety of problems and are therefore preferred. They also yield results that are based solely on the theory of conditional probabilities and are therefore, in most cases, more trustworthy.

The methods we describe, discuss, and apply, are by no means an exhaustive collection of statistical and numerical developments. They are simply a collection of methods that we are familiar with and/or that have been applied to astronomical problems such that it is possible to cite such applications appearing in astronomy journals. For this reason, plenty of useful methods will be neglected but that does not mean that such methods are not applicable or have not been applied to astronomical problems. The choice of the methods we describe is therefore a subjective choice of the author, which is rather convenient in the Bayesian context where subjective choices always have an effect on the obtained results.

The outline of this thesis is as follows. While in Sections 2 and 3 we discuss the statistical challenges and ideas behind the solution methods in general terms and present the basic principles of Bayesian statistics, we present some simple posterior sampling algorithms and Bayesian model comparison techniques in Sections 4 and 5, respectively. In Section 6 we discuss model adequacy and inadequacy briefly, and apply the methods to astronomical data in Section 7. Finally, we discuss the methods and their applications in Section 8.

2. Measurements and inverse problems

As is the case with astronomy in general, and searches for extrasolar planets in particular, all science is based on measurements of some kind. Measurements are always the driving force of theoretical considerations – a theory, model, or a hypothesis either remains the best available description of the reality, possibly gaining additional support, or is falsified and replaced by a better description when it is compared with other such descriptions given some available measurements. One is then entitled to ask what is the proce-

dure of falsification, or more accurately: when is a theoretical construction or a hypothesis falsified and when not? These questions are usually addressed by defining a measure of goodness whose values are first obtained for several descriptions of the measurements and then compared to one another. Underlying these comparisons are the only two true things in science: the measured quantities and the logical rules within the models described using mathematical relations. And even out of these, the former are corrupted by uncertainties of, usually, unknown type and magnitude whereas the latter might not be the best available descriptions, which leaves room for biases and misinterpretations. Yet, despite such difficulties, measurements and models that have been constructed in an attempt to describing them are the starting points of any scientific studies – perhaps apart from considerations that are purely theoretical.

However, even such theoretical considerations that are formulated using the language of mathematics, have to be compared to measurements to assess their explanatory qualities[3]. This requires the ability to quantify the relations between measurements and some variables of interest that are generally called the model parameters. More often than not, this process is far from simple and straightforward and requires the most advanced mathematical constructions to lead to the desired results: the discovery and quantification of a mathematical model describing the statistical properties of the measurements. This is the process of finding a solution to an inverse problem.

Throughout the vast fields of science to which astronomy is by no means an exception, theories, if any exist, are usually very complicated due to various complicated (non-deterministic) interactions in the system of interest and cannot be used as such but an appropriate approximation or empirical description needs to be found. Examples of such processes are not difficult to find. One can consider e.g. the formation process of planetary systems (e.g. Boss, 1997; Ida & Lin, 2010; Hansen & Murray, 2012) or estimation of stellar habitable zones (e.g. Selsis et al., 2007; Kopparapu et al., 2013) in the context of extra-solar planets. Furthermore, measurements are typically corrupted by uncertainties, usually containing systematic components due to an insufficient statistical modelling or simply because the measurements do not

[3]Comparing theoretical predictions to measurements is a fundamental requirement for all models. If this comparison is not possible due to lack of quantifiable predictions, such models can be considered to be "not even wrong", as was famously expressed by Wolfgang Pauli.

correspond to the modelled quantities well enough. The statistical challenges that arise from these grounds are called inverse problems. They are problems of finding the process that produces the observed features and are present whenever measurements are being analysed. A common forward problem occurs when one knows the cause and wants to know the consequence. This is usually a straightforward and easy calculation given e.g. some well-known laws of physics. The inverse problem is then that of knowing the consequence (i.e. the observed data) and being completely or partially ignorant about the cause – to a great extent a much more complicated problem to solve. With respect to inverse problems and solution techniques, we refer to the introductory text of Kaipio & Somersalo (2005).

Discrete inverse problems are the most common class of statistical problems in natural sciences. The discreteness means simply that the model used to describe the measurements is assumed to be fixed. Hence, there is a discrete amount of numbers, the model parameters, instead of a continuous spectrum of values, that describe the measured quantities. For practical reasons, any statistical problem is always discrete – it is not possible to save an infinite number of values to computer memory or any other storage media.

2.1. The relationship between measurements and models

Measurements are always indirect in the sense that a given measurement, random variable $m \in \Upsilon \subset \mathbb{R}^N$, where Υ is the measurement space[4], cannot be expected to be equal to the quantity of interest that we call parameter, random variable $\theta \in \Omega \subset \mathbb{R}^K$, where Ω is the parameter space, but these two can be assumed to be related by a statistical model. We use this formal notation to emphasise the fact that while measurements are typically random variables in the real line, to represent physical reality they cannot have arbitrarily large or small values and are therefore restricted to a subset of the whole real line that might, occasionally, consist of only two points if the measurements are logical in nature. In one of the simplest possible cases, this relation is $m = g(\theta) + \epsilon$, where ϵ is a random variable with unknown properties and $g : \Omega \to \Upsilon$ is a mapping relating the measurements to the parameters of interest. In this context, the inverse problem can be stated as a problem of finding the function g and the parameter values θ when m

[4]What we call a measurement space here, set Υ, consists of all the possible values the measurements could have and is therefore a small subset of the set $\mathbb{R}^N$, although its exact definition depends on the interpretation of the measurements.

has been measured in the precense of uncertainty expressed by the random variable ϵ whose properties are also unknown.

Because the measurements are random variables, they are drawn from some probability distribution: the true model describing the state and evolution of the system of interest. This density is the one the modeller would like to find by analysing the data to be able to predict the future measurements. However, there is no way of knowing whether any given probability density function (PDF) is this desired density or not. According to Kolmogorov (1968), when interpreted in terms of probability densities representing different models, the true model can never be found. This implies that for practical purposes there is no true model. At least, there is no way of knowing, no matter how sophisticated and accurate the models in hand are, whether one of these models represent the true PDF and not only some approximation of it. It is only possible to label all the tested PDFs corresponding to different models by some probability values, indicating how close they are to the true model with respect to one another. Hence, as stated by Box (1976): "All models are wrong but some are useful." This is also the philosophy adopted throughout this thesis while keeping in mind that the true model does not even exist. This philosophical view is evidently correct when the measurements describe a complex system whose behaviour cannot be derived from fundamental physical principles. However, it provides useful insights to simpler systems governed by physical theories, for instance, exoplanet detections. The reason is that regardless of the nature of the measurements, they are always corrupted by some sources of systematic errors that cannot be fully accounted for by the model.

What is then the philosophical approach a statistician should adopt when analysing the measurements in hand? If all the models can be labelled by a number describing their relative goodness, the modelling problem can be reduced to finding a collection of models with the highest relative goodnesses. Methods for this purpose are described in Section 5. The task of the modeller is then to select the most suitable set or class of models that are to be compared. This task is the one where machines cannot yet beat human intuition and imagination. And it is this task that is in a crucial role in all the solutions to statistical problems involving analysis of measured quantities.

2.2. Philosophical aspects of Bayesian methodology

In the Bayesian framework, probabilities are interpreted as measures of the degree of belief in an event, not as frequencies of events occurring when

repeating the experiment sufficiently many times. The former interpretation is therefore clearly a more general one because not all experiments can be repeated because they correspond to phenomena that only occurred once (e.g. formation of the Solar system and other historical events) and because repeating an experiment requires resources that are not always available in abundance. The immediately obvious shortcoming, although it is a shortcoming only when not understood properly, in the Bayesian framework is that prior information makes all Bayesian data analyses subjective processes. Bayesian statisticians update the prior information they might possess on the properties of the system of interest with information from the measurements and calculate the combined information, the posterior information, by using the famous Bayes' theorem named after the English mathematician Thomas Bayes. It could be argued that this kind of inference is biased because of the combination of the valuable information from the measurements with subjective initial beliefs, or prior information. However, as we will see, this argument cannot be justified because all science is based on such subjective beliefs and abandoning it would leave us without any useful statistical tools.

A Bayesian statistician does not see the posterior information any differently from the prior. The new posterior can always be used as prior information when new measurements are being analysed. After several new measurements, the original subjective prior does not play a crucial role anymore because the information from the measurements "overwhelms" the information from the original prior. In fact, after a sufficient amount of measurements, any Bayesian statistician with any (reasonable) prior beliefs will end up having asymptotically the same posterior information and hence they all agree even though they may have disagreed severely initially. Furthermore, if the idea of a prior belief seems counterscientific, it is always possible to define a noninformative prior, i.e. no prior information or maximum *a priori* ignorance, although such definitions are necessarily subjective as well. Typically a preferred choice would be a uniform distribution that corresponds to e.g. that the random variable θ can be found in all possible[5] intervals of similar length with the same probability, or equivalently, that the chances of the variable θ having a value in a given interval is proportional to the length of the interval. In reality, it is actually difficult to find cases where such a

[5]This refers to the parameter space Ω whose choice is only one aspect of the prior choice.

prior density would be realistic choice given the physical interpretations of the parameter θ. Therefore, a uniform prior density is only one subjective belief among others and does not technically differ from other realistic priors, except that in many cases it makes the computations relatively easy. A more detailed justification of priors can be found e.g. in Ford & Gregory (2007) and Tuomi & Anglada-Escudé (2013) in the context of exoplanet detections.

There are also practical differences. Unlike the frequentist approach, Bayesian methods do not differentiate between the comparison of two and more than two competing hypotheses or statistical models. In fact, the selection between competing models or hypotheses is not different from the selection between parameter values within a single model. The reason is that all the models, and all the combinations of parameter values within these models, can be arranged to a linear order using the corresponding Bayesian model probabilities and parameter probability densities. This is not possible with classical hypothesis testing methods where the goal is to test whether a simpler null hypothesis can be rejected in favour of a more complicated hypothesis – a method that can lead to undesired results if the alternative hypothesis does not represent the data well either. Going one step further, it is in fact possible to interpret the index describing the chosen model, e.g. $i = 0, ..., k$, as only another free parameter – one that has only integer values.

3. Inverse solution and Bayesian inference

An inverse solution is commonly defined as the full multidimensional conditional probability density of model parameter vector given the measurements (e.g. Kaipio & Somersalo, 2005). This solution contains all the information available in the measurements used to calculate the solution with respect to the selected model. It is commonly presented using a Bayesian credibility set (BCS) and a maximum *a posteriori* (MAP) estimate of the model parameter. This definition is generalised here to take into account the model selection problem as well. Hence, the inverse solution contains the densities of the model parameters of all the models in the *a priori* selected model set accompanied by their respective model probabilities that is actually a discrete density of the index parameters i describing which one of the models is being used.

We use the term inverse solution when discussing solutions to discrete inverse problems. The structure of the statistical model (denoted as $\mathcal{M}$) used to describe the measurements is assumed to be fixed, including the ex-

act expression for the function g and the properties of the random variable ϵ. Therefore, the only unknown for a given model is the posterior density of parameter vector θ. Since this parameter vector has a limited amount of components with $\dim \Omega = K$, the inverse problem is discrete. The validity of the different models $\mathcal{M}_0, \mathcal{M}_1, \mathcal{M}_2, ...$, is then analysed using Bayesian model comparison methodology, that is, by calculating the Bayesian model probabilities of all the models given the available measurements. The posterior density of the model parameters given the measurements m is written simply as a non-negative function $\pi(\theta|m)$ that satisfies the condition

$$\int_\Omega \pi(\theta|m)d\theta = 1. \tag{1}$$

According to the Bayes' theorem, the density $\pi(\theta|m)$ can be written as

$$\pi(\theta|m) = \frac{l(m|\theta)\pi(\theta)}{P(m)}, \tag{2}$$

where $l(m|\theta)$ is the likelihood function of the measurements and $\pi(\theta)$ is the prior density of θ containing all the information on the parameter known prior to obtaining the measurement. Function $P(m)$ is simply a scaling factor that is used to scale the integral of $\pi(\theta|m)$ over the parameter space to unity and can be written as

$$P(m) = \int_\Omega l(m|\theta)\pi(\theta)d\theta. \tag{3}$$

It is also called the marginal density of m and the integral in Eq. (3) is called the marginalisation of the parameter θ.

The situation is not different when there are two or more measurements or sets of measurements available. For N measurements, the probability density of θ given these measurements $m = (m_1, ..., m_N)$ can be written as

$$\pi(\theta|m) = \frac{l(m_1, ..., m_N|\theta)\pi(\theta)}{P(m)} = \frac{\pi(\theta)\prod_{i=1}^{N} l(m_i|\theta)}{P(m)}, \tag{4}$$

where the last equality is valid if the measurements are independent. Therefore, the Bayesian inference is simply the process of combining the likelihoods corresponding to different measurements with the prior density according to the Bayes' rule. This corresponds to interpreting the information in the measurements in terms of the selected model and expressing it as a posterior probability density of the model parameters.

3.1. Likelihood function

The likelihood function is a probabilistic representation of the measurements given the parameters. It is the probability density from which the measurements would have been drawn if the distribution described by a statistical model with parameter θ was the correct description of the data.

A very common practical choice is to model the measurements as Gaussian random variables. In this case, the likelihood of measurements $m = (m_1, ..., m_N)$ is written simply as

$$l(m|\theta) = l(m|\mu, \Sigma) = (2\pi)^{-\frac{N}{2}} |\Sigma|^{-\frac{1}{2}} \exp\Big\{ -\frac{1}{2}[\mu - m]^T \Sigma^{-1} [\mu - m] \Big\}, \quad (5)$$

where $\Sigma \in \mathbb{R}^{N \times N}$ is the covariance matrix of the measurement vector, μ is the modelled mean of the measurements, and $|\cdot|$ denotes the matrix determinant. This likelihood can be written briefly as a multivariate Gaussian density $\mathcal{N}(\mu, \Sigma)$. All the components of matrix Σ and vector μ are components of the parameter vector θ and therefore free parameters of the model, although simplifying assumptions are commonly made e.g. that $\Sigma = \sigma^2 I$, where σ is one of the components in θ and I is the identity matrix.

The parameter Σ (or σ), and any other parameters of no direct interest to the modeller, are usually referred to as nuisance parameters because they have to be included in the statistical model but describe features not essential for understanding the system of interest. However, parameter μ is not a nuisance parameter but consists of the quantities in the mathematical description of the measurements that are of major interest and significance to the modeller and whose PDFs are valuable in order to understand the behaviour and features of the modelled system. However, the division of parameters to such nuisance parameters and parameters of interest is completely arbitrary (subjective) and therefore we do not use such expressions but instead refer to them both as parameters.

3.2. Prior probability densities

The prior knowledge is contained in the prior density $\pi(\theta)$, sometimes called the prior model. In classical statistics there is no such thing as a prior, but its existence is a natural consequence of conditional probabilities in Eq. (2) and it is necessarily an integrated part of the scientific decision making process. For instance, all statistical methods based on the likelihood function, such as the maximum likelihood estimation, in fact assume that

the underlying prior density is uniform in the parameter space, although this is rarely expressed and its validity cannot be assessed generally. Similarly, all statistical tests based on the commonly used χ^2 statistics actually assume a uniform prior – on top of assuming a Gaussian likelihood function with commonly a fixed variance. This kind of a choice of fixed parameters is such a common in statistical analyses, that we discuss it briefly before going any further with prior choice in general.

Suppose that a model $\mathcal{M}$ consists of two parameters such that $\theta = (\omega, \phi)$. However, a simpler model $\mathcal{M}_0$, for which $\phi = \phi_0$ is fixed, is also used to describe the measurements m. Therefore, it can be seen that

$$\pi(\theta|m, \mathcal{M}_0) = \frac{l(m|\omega, \phi, \mathcal{M}_0)\pi(\omega, \mathcal{M}_0)\pi(\phi, \mathcal{M}_0)}{P(m|\mathcal{M}_0)}, \tag{6}$$

when assuming that the priors of ω and ϕ are independent such that $\pi(\omega, \phi) = \pi(\omega)\pi(\phi)$. But in this equation, $\pi(\phi, \mathcal{M}_0) = \delta(\phi - \phi_0)$, where δ denotes the Dirac delta-function, and it can be seen that the only difference between models $\mathcal{M}$ and $\mathcal{M}_0$ is that the former has a prior $\pi(\phi)$ whereas the latter has a prior $\delta(\phi - \phi_0)$ for parameter ϕ. Thus, models that are nested in more general descriptions are in fact only models with different prior densities. Therefore, comparison of a sequence of nested models is equivalent to comparing different prior models. This realisation has significant implications that we will discuss further in the Section 5.

Prior densities have additional properties that need to be accounted for in Bayesian analysis of scientific data. For instance, consider a coordinate transformation from θ to θ' described by using an invertible mapping $f : \Omega \to \Omega'$ such that f^{-1} exists. This means that $\theta' = f(\theta)$. However, it is easy to see that selecting e.g. $\pi(\theta) = \mathcal{U}(a, b)$, where $\mathcal{U}$ denotes a uniform density in an interval, leads to a transformation in the prior such that

$$\pi(\theta') = \pi(\theta)\Big|\frac{d\theta}{d\theta'}\Big|, \tag{7}$$

where $|\frac{d\theta}{d\theta'}|$ is the Jacobian of the transformation. If the mapping f is not a linear one in which case it would correspond to a change in the unit system, it necessarily leads to the inconvenient conclusion that if $\pi(\theta)$ is a uniform distribution, $\pi(\theta')$ is not and choosing uniform distributions in both coordinate systems leads to analysis results that are different and whose difference depends of the selected f. Therefore, priors are an in-built property of statistical analyses and cannot be neglected in any statistical problem. This

example also demonstrates that the Bayesian framework of data analysis is the only logically consistent one as the different results arising from priors corresponding to different subjectively selected parameter systems can also be taken into account by modifying the priors in suitable ways implied by the Eq. (7).

3.3. Point and uncertainty estimates

Point estimates (θ_P) are simply vectors of the parameter space that can be used to roughly describe the modelled system with as few numbers as possible. These vectors contain no information on the shape of the parameter density, and should therefore be accompanied by the corresponding uncertainty estimates describing the width of the density, such as standard deviations or the Bayesian credibility sets, and perhaps by some other measures of the shape of the density, such as skewness and kurtosis.

However, point estimates can be misleading in a variety of situations. For instance, if the probability density is skewed or has long tails, the mean is a poor estimate and is typically very different from the maximum likelihood (ML) or maximum *a posteriori* (MAP) estimates. In these cases, the standard deviation is also a poor estimate for parameter uncertainty. Furthermore, if the density has more than one maxima, the ML and MAP estimates are also misleading and should not be used as such but the full inverse solution is needed to describe the system with a sufficient accuracy. For instance, it is easy to see that if the posterior density can be described by using a sum of two densities such that $\pi(\theta) = \lambda\delta(\theta - \theta_0) + (1 - \lambda)\mathcal{N}(\mu, \sigma^2)$, the MAP density is equal to θ_0 but in fact a fraction $1 - \lambda$ of the posterior density is found around $\theta = \mu$, which makes the MAP estimate a very biased description of the density when the parameter $\lambda \in [0, 1]$ is small. Clearly, for $\lambda \ll 1$ the mean estimate is $\theta_M \approx \mu$ and thus the mean and MAP estimates differ from one another as much as the difference between μ and θ_0 is. But this only demontrates that point estimates can lead to poor results when the properties of the posterior are not described well by such simple numbers.

The Bayesian credibility set (BCS) is a subset of the parameter space that contains all the parameter values with posterior probability higher than some given number $c \in \mathbb{R}^+$ (e.g. Kaipio & Somersalo, 2005; Tuomi & Kotiranta, 2009). The BCS is actually a hypervolume enclosing the most probable parts of the parameter space. Formally, the BCS of a posterior density $\pi(\theta|m)$ with

parameter $\delta \in [0, 1]$, is

$$\mathcal{D}_\delta = \left\{ C \subset \Omega : \int_{\theta \in C} \pi(\theta|m) d\theta = \delta, \pi(\theta|m)|_{\theta \in \partial C} = c \right\}, \tag{8}$$

where the set ∂C represents the edge of the set C. This edge is a hypersurface enclosing the BCS and consists of parameter values that have equal probability of c. The interpretation of the BCS is simple because the probability of finding a value inside the $\mathcal{D}_\delta$ set is simply δ.

In fact, the BCS is a much more general way of estimating the model parameters when the posterior has a complicated "landscape" in the parameter space. For instance, $\mathcal{D}_1$ is equivalent to the maximum *a posteriori* estimate if the posterior density has a unique maximum. Moreover, choosing a sequence $\delta_1, ..., \delta_n$ such that $\delta_i < \delta_{i+1}$ for all i can be used to determine the corresponding $\mathcal{D}_{\delta_i}$ for which it holds that $\mathcal{D}_{\delta_i} \supset \mathcal{D}_{\delta_{i+1}}$. This sequence of BCSs can then be used as a tool to describe the properties of the posterior density much more accurately than by using any point estimate.

3.4. Bayesian multidata inversion

Different measurements, or different sets of measurements – datasets from different sources – cannot generally be used in the process of finding the inverse solution on their own because they contain different amounts of information (e.g. Kaasalainen, 2011; Kaasalainen & Viikinkoski, 2012). For instance, if a dataset has plenty of measurements but only a little information, it should have a smaller weight coefficient than another dataset with a few measurements but plenty of information. This is apparent because when finding the model solution of the two datasets simultaneously, the smaller set, regardless of higher information content, would be overwhelmed by the larger number of measurements in the larger set. Hence, the high amount of information in the smaller set would not be inferenced correctly to the posterior density. There are no generally accepted rules for selecting these weights, but some principles can be derived nevertheless. For instance, a point estimate called the maximum compatibility estimate (Kaasalainen, 2011; Kaasalainen & Viikinkoski, 2012) takes the different information contents of different datasets, or data modes, into account by weighting them with optimal coefficients. However, there are also simpler ways of combining several datasets.

In Eq. (4), the posterior PDF is calculated using several measurements $m_i \in \Upsilon, i = 1, ..., N$. Let us assume that each m_i is a set of measurements

containing N_i individual measurements. Given some statistical model of the measurements, the posterior density of measurements m_i can be written as

$$\pi(\theta, \sigma_i|m_i) = c_i^{-1} l(m_i|\theta, \sigma_i)\pi(\theta, \sigma_i), \tag{9}$$

where $\pi(\theta, \sigma_i)$ is the joint prior density of all the parameters and σ_i contains the nuisance parameter of each of the measurements in m_i. Therefore, for Gaussian densities without covariance, Σ in Eq. (5) is $\Sigma = \sigma_i^2 I$ and I is the identity matrix. The constant c_i is the marginal density $P(m_i)$ in the Eq. (3) but the integral is clearly over both θ and σ_i.

In the Gaussian case, this posterior can be written explicitly as

$$\pi(\theta, \sigma_i|m_i) = c_i^{-1}(2\pi)^{-N_i/2}\sigma_i^{-N_i} \exp\Big\{ - \frac{||g_i(\theta) - m_i||_2^2}{2\sigma_i^2} \Big\}\pi(\theta, \sigma_i), \tag{10}$$

where $||\cdot||_2$ is the common Euclidean vector norm (the 2-norm). It is commonly assumed that the parameter σ_i is known *a priori* and its value is fixed. This is actually a special case of the above expression where the prior density of σ_i is a delta function that peaks at some positive value, say σ_0. In this case Eq. (10) simplifies considerably and the problem of finding the MAP solution of parameter θ becomes a minimisation problem that is commonly defined as

$$\theta_{MAP} = \arg\min_{\theta\in\Omega} \Big[||g_i(\theta) - m_i||_2^2 + \alpha||H\theta||_2^2 \Big], \tag{11}$$

where $\alpha \in \mathbb{R}$ is the so-called Tikhonov regularisation parameter (Tikhonov & Arsenin, 1977). Matrix H can be interpreted as representing the prior information on the parameter θ because $\alpha||H\theta||_2^2 = \log \pi(\theta, \sigma_i)$, although the prior cannot be always written in the matrix form of Eq. (11).

The expression in Eq. (11) becomes interesting if written for M datasets. In such a case, it becomes

$$\theta_{MAP} = \arg\min_{\theta\in\Omega} \Big[\alpha||H\theta||_2^2 + \sum_{i=1}^{M} \omega_i ||g_i(\theta) - m_i||_2^2 \Big]. \tag{12}$$

Now, the minimised function contains the weight coefficients $\omega_i \in \mathbb{R}^+$ of each of the datasets, whose values have to be adjusted according to the information content, or the lack of it, of the datasets. However, there is no generally accepted way of adjusting them, which is clearly an unsatisfactory feature and poses limitations to the applicability of Eq. (12).

This problem can be overcome if the parameter σ_i is not fixed. In that case the posterior density of the parameters given all the datasets and Gaussian measurements is

$$\pi(\theta, \sigma|m) = c^{-1}\pi(\theta, \sigma)\prod_{i=1}^{M}(2\pi)^{-N_i/2}\sigma_i^{-N_i}\exp\Big\{-\frac{||g_i(\theta) - m_i||_2^2}{2\sigma_i^2}\Big\}, \quad (13)$$

where $\sigma = (\sigma_1, ..., \sigma_M)$. Using this posterior, the MAP estimate of the parameters, including those in vector σ, can be calculated as

$$(\theta, \sigma)_{MAP} = \arg\max_{(\theta,\sigma)\in\Omega}\Big[\log\pi(\theta, \sigma) - \sum_{i=1}^{M} N_i \log(\sigma_i\sqrt{2\pi}) - \sum_{i=1}^{M}\frac{||g_i(\theta) - m_i||_2^2}{2\sigma_i^2}\Big]. \quad (14)$$

Now, the weight parameters in Eq. (12) are naturally expressed as $\omega_i = (2\sigma_i^2)^{-1}$ for all i but there is an additional term in the equation that is a function of σ. This term can be interpreted simply as the information entropy of the measurements.

Denoting $\hat{\theta} = (\theta, \sigma)$, Eq. (14) can be re-written as

$$\hat{\theta}_{MAP} = \arg\max_{\hat{\theta}\in\Omega}\Big[h(\hat{\theta}) - \sum_{i=1}^{M}\omega_i||g_i(\hat{\theta}) - m_i||_2^2\Big], \quad (15)$$

where $\omega_i = (2\sigma_i^2)^{-1}$ are now free parameters of the modelled system and h is some function of the parameters representing the prior information and the entropy of the σ parameter. This function can be called the regularisation function but it must be remembered that it contains model parameters, and therefore, only the prior density part of it can be selected subjectively.

We note that the MAP estimate cannot generally be written in a simple form of Eq. (15) because it assumes a Gaussian distribution for the measurement noise. However, it provides an example of the standard notation in the literature regarding solutions to inverse problems (e.g. Kaipio & Somersalo, 2005, and references therein). Therefore, while useful in a restricted set of statistical problems, it is not generally applicable.

3.5. Time series

When the measurements consist of a time series, it is usually more convenient to write the equations for the MAP estimate by using the standard notation of Eq. (2). In this case, there is a vector t that corresponds to the measurement m and is expected to explain the evolution of the system of interested according to some model. Typically, the vector t is interpreted as an explanatory variable that explains the behaviour of the measurements. We note that the vector t does not necessarily have the dimension of time but it can be any variable, or indeed several of them, whose values are measured together with m and are assumed to have an effect on the measurements. However, the values of t are not of direct interest and they do not therefore need to be modelled. This is the only difference between m and t.

Assuming that the values of m_i and t_i for $i = 1, ..., N$ have been measured, the posterior density of parameters can be written as

$$\pi(\theta|m,t) = \frac{l(m|\theta,t)l(t|\theta)\pi(\theta)}{P(m,t)}. \tag{16}$$

In this equation it is convenient to assume that the likelihood $l(t|\theta)$ can be approximated as a delta-function likelihood, i.e. that the value of t can be measured with such a high accuracy that its uncertainty can be neglected completely. The expression in Eq. (16) then becomes equivalent to Eq. (2) for a given t.

If the values of t cannot be assumed to be known with a sufficient accuracy, their likelihood functions would have to be broader than strict delta-functions. For instance, it could be realistic to express these likelihoods by using another model that has parameters of its own. This implies that t would have to be modelled as well in order to be able to take into account all the possible sources of uncertainty in the modelled system. But this would simply mean that we could write (m, t) instead of m in Eq. (2), making the situation equally simple in practice.

For instance, if t represents some activity measurements of stellar origin, such as bisector velocities in the case of radial velocity data, it cannot be assumed that its likelihood function is close to a delta-function in practice. In such cases, the likelihood $l(t|\theta)$ has to be taken into account according to Eq. (16).

4. Solutions from posterior samplings

The full inverse solution is defined as the full posterior probability density of the model parameters. This definition has several advantages over more traditional definitions for the solutions of inverse problems, such as pure point estimates and corresponding uncertainty estimates. But if traditional solution methods have been used successfully in a variety of inverse problems, why should one put effort into finding the full solutions in the first place? The answer to this question is simple. The full density contains all the information in the measurements given the selected model. Any point estimates, as well as any measures of uncertainty or shape of the density, can be calculated using this solution, which makes it a more general approach. There are also efficient methods for approximating the posterior densities, such as the collection of algorithms classified under the general title of Markov chain Monte Carlo (MCMC) methods.

4.1. Metropolis-Hastings algorithm

Out of all MCMC methods, the Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970) is one of the most widely used posterior sampling algorithm in astronomy (e.g. Balan & Lahav, 2009; Tuomi & Kotiranta, 2009; Wright et al., 2011b). This algorithm can be used to draw statistically representative samples from the parameter posterior density. Such samples can then be used to estimate the joint posterior density of the model parameters. The first steps are to choose an initial parameter value θ_0 and a proposal density $q(\theta_i, \theta_j)$, sometimes called a transition kernel (usually a symmetric density with a mean equal to θ_j) that is a probability density describing the probability of a transition from θ_j to θ_i in the parameter space. The second step is to calculate the conditional probability of having the initial parameter value given the measurements $\pi(\theta_0|m) \propto l(m|\theta_0)\pi(\theta_0)$. After this, the algorithm works by repeating the following two steps for $n = 1, 2, ..., n_C$, until the Markov chain converges (sufficiently close) to the posterior density $\pi(\theta|m)$.

1. Draw a new parameter value, $\theta^\star$, from $q(\theta^\star|\theta_n)$ and calculate the corresponding likelihood $l(m|\theta^\star)$.

2. If for a random number $\alpha \in [0, 1]$ it holds that

$$\alpha < \frac{\pi(\theta^\star|m)q(\theta_n, \theta^\star)}{\pi(\theta_n|m)q(\theta^\star, \theta_n)}, \tag{17}$$

   then set $\theta_{n+1} = \theta^\star$, otherwise set $\theta_{n+1} = \theta_n$.

It can be seen that the acceptance rule simplifies considerable when the proposal density is indeed symmetric. However, the ability to calculate the chain is not sufficient to be able to use the method. It needs to be known how long chain is required to obtain a statistically representative sample from the desired density $\pi(\theta|m)$. The question of how to select the proposal density needs to be addressed as well. Typically it is sufficient to ensure that the obtained chains are long enough in the sense that their statistics do not change significantly when adding new members to the chain. Similarly, whether the obtained samples are statistically representative of the posterior can be verified if several chains with different initial states result in the same posterior density.

4.2. Adaptive Metropolis algorithm

The Metropolis-Hastings algorithm can be improved by adapting the proposal as a function of the Markov chain member. If a chain calculated using some adaptive algorithm retains its ergodic properties[6] and thus has a stationary distribution, this algorithm can be used to modify the proposal as more information is gathered from the posterior density. Potentially, an adaptive algorithm will be computationally faster in applications, because the proposal adapts closer and closer to the posterior. Also, it will provide better mixing properties in the sense that all sections of considerable probability in the parameter space will be visited by the chain frequently enough. After a sufficient burn-in period, such a chain will "forget" its initial state and make the initial selection of the proposal density and initial state θ_0 irrelevant – unless this initial selection is unfortunate and the initial state is close to a very high local maxima in the posterior density. In such a case, the convergence of the chain might take longer than a typical duration of an academic career and be of little practical importance. This also emphasises the fact that several different initial states should always be explored in practice.

The adaptive Metropolis algorithm presented by Haario et al. (2001) retains the ergodicity of the chain, despite the fact that it is no longer exactly Markovian but only asymptotically so. This algorithm is constructed by assuming that the proposal is a Gaussian multivariate density, which is updated given the information accumulated so far. Although this assumption

[6]These properties are 1) that the chain is aperiodic such that $\nexists n$ such that $\theta_{i+n} = \theta_i$ and 2) it is recurrent such that the probability of the chain returning to state θ_i is zero.

can lead to difficulties in the case of multimodality and high skewness of the posterior, it appears to work well in several applications (e.g. Tuomi, 2012; Tuomi et al., 2013a,b). This simple algorithm for updating the proposal density in the Metropolis-Hastings sampling can be described in the following way.

If the covariance matrix V_n of the model parameters is known for the chain up to the nth member, there is a recursive formula for this covariance at the next step. This formula is

$$V_{n+1} = \frac{n+1}{n} V_n + \frac{s}{n}[n\bar{\theta}_{n-1}\bar{\theta}_{n-1}^T - (n+1)\bar{\theta}_n\bar{\theta}_n^T + \theta_n\theta_n^T + \epsilon I], \qquad (18)$$

where $\bar{\theta}_n$ denotes the mean of n members of the chain and ϵ is some small number that ensures the positivity of the matrix. Parameter s is commonly set equal to $(2.38)^2/K$, where K is the number of parameters (Gelman et al., 1996). Since the mean is also trivially expressed using a recursive formula, this equation enables the updating of the covariance matrix at each step of the chain, which makes the proposal adapt to the information gathered.

The samplings from the posterior are very efficient when the amount of parameters in the model is low and the chains converge readily to the posterior in the sense that statistics of several chains do not change significantly and are consistent with one another after they have become long enough. However, these methods become increasingly inefficient in cases of multimodal posterior density and/or when there are significant non-linear correlations between the parameters. In such cases, generalisations of the above methods, such as the delayed-rejection adaptive Metropolis algorithm (Haario et al., 2006), which improves the efficiency of the sampling in case of multimodal posterior, and reversible jump MCMC (Green, 1995) that can be used to draw a sample from several models simultaneously, or jumping between different subsets of the parameter space, are important generalisations of the standard Metropolis algorithm.

5. Bayes factors and model selection

When selecting between two or more competing models or hypotheses, a way has to be found of balancing between "good fitting" and parsimony (e.g. Cavanaugh, 1999; Liddle, 2007). There are several ways of achieving this goal, but the Bayesian methods provide the most general and trustworthy framework.

Bayesian model selection theory can be used effectively when assessing the relative probabilities of two or more hypotheses, or mathematical models describing the measurements. When comparing this methodology with the frequentist approach, it can be found to have several advantages. First of all, the methodology is independent of the number of hypotheses or models tested. There can be two or more (k) models that need to be tested against measurements and the methodology is the same unlike in the frequentist approach which is usually designed to compare only two hypotheses: the null hypothesis and some alternative one based on the pioneering work of e.g. Pearson (1901), Fisher (1922), and Neyman & Pearson (1928).

When comparing models that are constructed to represent some measured quantities, the Bayesian model comparison procedure can be described in terms of few simple equations. In our notation, $\mathcal{M}_j, j = 1, ..., k$, are models defined as functions of their respective parameter vectors θ_j. Because the probability of a model being "more correct" than some other model can only be determined with respect to a measurement m, this probability can be written as

$$P(\mathcal{M}_j|m) = \frac{P(m|\mathcal{M}_j)P(\mathcal{M}_j)}{\sum_{i=1}^{k} P(m|\mathcal{M}_i)P(\mathcal{M}_i)} \tag{19}$$

where

$$P(m|\mathcal{M}_j) = \int \pi(m|\theta_j, \mathcal{M}_j)\pi(\theta_j|\mathcal{M}_j)d\theta_j \tag{20}$$

is the marginal integral of Eq. (3) with the only exception that the dependence on the selected model is written in the equation explicitly. The probability in Eq. (19) can be written shortly as

$$P(\mathcal{M}_j|m) = P(\mathcal{M}_j)\left[\sum_{i=1}^{k} B_{i,j}(m)P(\mathcal{M}_i)\right]^{-1}, \tag{21}$$

where

$$B_{i,j}(m) = \frac{P(m|\mathcal{M}_i)}{P(m|\mathcal{M}_j)} \tag{22}$$

is the Bayes factor in favour of model $\mathcal{M}_i$ and against model $\mathcal{M}_j$.

Based on the arguments of Jeffreys (1961), Kass & Raftery (1995) proposed interpreting the Bayes factors and the corresponding model probabilities according to Table 1, although Evett (1991) suggested that the Bayes factor should have a value of at least 1000 for decisive evidence. Basically,

Table 1: The interpretation of Bayes factors and the corresponding model selection according to Jeffreys (1961); Kass & Raftery (1995).

$B_{i,j}$	Evidence in favour of the ith model
1-3	Not worth mentioning
3-20	Positive
20-150	Strong
>150	Decisive

this Jeffreys scale corresponds roughly to the usual interpretation of probabilities as measures of confidence or degree of belief in whether an event takes place or not. In this case, the event is that the measurements m have been drawn from a probability density described using the model $\mathcal{M}_j$, i.e. the corresponding statistical likelihood $l(m|\theta_j, \mathcal{M}_j)$ and the prior density $\pi(\theta_j|\mathcal{M}_j)$.

5.1. Computation of Bayes factors from posterior samples

When calculating the value of Bayes factor, the integral in Eq. (20) has to be evaluated. There are several ways of approximating this integral (e.g. Newton & Raftery, 1994; Kass & Raftery, 1995; Chib, 1995; Chib & Jeliazkov, 2001; Clyde et al., 2007; Tuomi & Jones, 2012) and computationally demanding direct numerical integrations are not always necessary. If the posterior probability density of the model parameters has been sampled, i.e. that a sample has been drawn from it using some MCMC method, simple estimates can be obtained by using a method called importance sampling that is based on one of the most valuable mathematical operations: expressing an obvious issue in a less obvious way.

The idea behind importance sampling is to choose functions g and w such that $\pi(\theta) = w(\theta)g(\theta)$ and write the marginal integral as an expectation $\mathbb{E}_g$ with respect to the probability density g. Thus we have

$$\mathbb{E}_g[w(\theta)l(m|\theta)] = \int g(\theta)w(\theta)l(m|\theta)d\theta = P(m), \tag{23}$$

where we call function g the importance sampling function. Now, given a sample of N members drawn from the density g, i.e. that we have $\theta_i \sim g(\theta)$ for all $i = 1, ..., N$, it is possible to estimate the expectation in (23) as (e.g.

Kass & Raftery, 1995)

$$\hat{P} = \left[\sum_{i=1}^{N} \frac{\pi(\theta_i) l(m|\theta_i)}{g(\theta_i)} \right] \left[\sum_{i=1}^{N} \frac{\pi(\theta_i)}{g(\theta_i)} \right]^{-1}. \tag{24}$$

To be able to use this estimate, the function g has to be selected appropriately in such a way that it is possible to draw a sample from it, and that the estimate converges to the correct value of the marginal integral rapidly and reliably as a function of N.

It is easy to verify that choosing $g(\theta) = \pi(\theta)$ or $g(\theta) = \pi(\theta|m)$, i.e. choosing g equal to the prior or the posterior densities, respectively, leads to the mean estimate ($\hat{P}_M$) and the harmonic mean estimate ($\hat{P}_{HM}$), although these simple estimates are biased and/or have poor convergence properties (Newton & Raftery, 1994; Kass & Raftery, 1995; Tuomi & Jones, 2012). For instance, the most significant problem with the harmonic mean estimate is that occasional small values in the likelihood dominate and can result in a bias in the resulting Bayes factor. Also, the harmonic mean estimate does not necessarily satisfy the Gaussian central limit theorem (Kass & Raftery, 1995).

In an attempt to overcome these problems, Tuomi & Jones (2012) proposed a truncated posterior mixture estimate (TPM) that appears to work reasonable well in practice (Tuomi, 2012; Tuomi et al., 2013a,b). This estimate is obtained by setting each $g(\theta_i) = (1-\lambda)\pi(\theta_i|m) + \lambda\pi(\theta_{i-h}|m)$, which lead to the estimate

$$\begin{aligned} \hat{P}_{TPM} = & \left[\sum_{i=1}^{N} \frac{l_i p_i}{(1-\lambda) l_i p_i + \lambda l_{i-h} p_{i-h}} \right] \\ & \times \left[\sum_{i=1}^{N} \frac{p_i}{(1-\lambda) l_i p_i + \lambda l_{i-h} p_{i-h}} \right]^{-1}, \end{aligned} \tag{25}$$

where we have denoted $l_i = l(\theta_i|m)$ and $p_i = \pi(\theta_i)$ for short.

There are also more accurate (and more complicated) methods for estimating the marginal likelihood using the output of an MCMC algorithm directly (e.g. Chib, 1995; Kass & Raftery, 1995; Chib & Jeliazkov, 2001). For the Metropolis-Hastings algorithm, an estimate called the one-block Metropolis-Hastings (OBMH) estimate for the marginal integral can be calculated in the following way (Chib & Jeliazkov, 2001). The random variable

α in the M-H algorithm (Eq. (17)) can in fact be considered a function of the parameter vector at the two points θ and θ'. Hence, it can be written as

$$\alpha(\theta, \theta') = \min \left\{ 1, \frac{\pi(\theta'|m)q(\theta', \theta)}{\pi(\theta|m)q(\theta, \theta')}) \right\} \tag{26}$$

Denoting $p(\theta, \theta') = \alpha(\theta, \theta')q(\theta, \theta')$, it follows that for any point $\theta^\star \in \Omega$

$$p(\theta, \theta^\star)\pi(\theta|m) = p(\theta^\star, \theta)\pi(\theta^\star|m). \tag{27}$$

Integrating both sides, it follows that the posterior density at $\theta^\star$can be expressed as

$$\begin{aligned} \pi(\theta^\star|m) &= \frac{\int_{\theta\in\Omega} \alpha(\theta, \theta^\star)q(\theta, \theta^\star)\pi(\theta|m)d\theta}{\int_{\theta\in\Omega} \alpha(\theta^\star, \theta)q(\theta^\star, \theta)d\theta} \\ &= \frac{\mathbb{E}_\pi[\alpha(\theta, \theta^\star)q(\theta, \theta^\star)]}{\mathbb{E}_q[\alpha(\theta_\star, \theta)]}, \end{aligned} \tag{28}$$

where the expectations $\mathbb{E}_\pi$ and $\mathbb{E}_q$ are with respect to densities $\pi(\theta|m)$ and $q(\theta^\star, \theta)$, respectively. Now, these expectations can be estimated because after the computation of the Markov chain, there is a sample from the posterior available and it is easy to draw a sample from the proposal density to estimate $\mathbb{E}_q$. Hence, using the mean estimate based on the importance sampling, an estimate for $\pi(\theta^\star|m)$ can be calculated as

$$\hat{\pi}(\theta^\star|m) = \frac{K^{-1}\sum_{k=1}^{K} \alpha(\theta^{(k)}, \theta^\star)q(\theta^{(k)}, \theta^\star)}{J^{-1}\sum_{j=1}^{J} \alpha(\theta^\star, \theta^{(j)})}, \tag{29}$$

where $\theta^{(k)}$ are drawn from the posterior and $\theta^{(j)}$ are drawn from q. Finally, an estimate for the marginal integral can be written, according to the Bayes rule and writing the equations explicitly as a function of the given model $\mathcal{M}_k$, as

$$\begin{aligned} &\log \hat{P}(m|\mathcal{M}_k) = \log l(m|\theta_k^\star, \mathcal{M}_k) + \log \pi(\theta_k^\star|\mathcal{M}_k) \\ &- \log \hat{\pi}(\theta_k^\star|m, \mathcal{M}_k). \end{aligned} \tag{30}$$

Now, if the $\theta_k^\star$ is chosen as e.g. $\theta_k^\star = \theta_{k,MAP}$, it is easy to calculate the estimate $\hat{P}(m|\mathcal{M}_k)$ that is an estimate for the marginal integral in Eq. (20).

5.2. Other methods

There is also a diverse collection of other methods for estimating the marginal integrals. These include the reversible-jump MCMC (Green, 1995) that enables the determination of the relative probabilities of the solutions given several models simultaneously and can be readily combined with the Metropolis algorithm and its modifications (Green & Mira, 2001); methods relying on thermodynamic integration (Gelman & Meng, 1998) and its application to a so-called parallel tempering algorithm (Gregory, 2005; Ford & Gregory, 2007), and the nested sampling method of Skilling (2006). This list is by no means representative and there are various other techniques for such integral estimations but we do not discuss them here in detail. However, a class of simple estimates called information criteria, although somewhat "less Bayesian", can be very useful because they are extremely simple and provide fast means of assessing the magnitudes of the marginal integrals under certain assumptions.

5.3. Model selection based on information criteria

There are various model selection methods based on different information criteria developed to estimate the marginal integrals under some simplifying assumptions. The different criteria have been used successfully in several fields but little is known of their relative performance, although several comparisons of different criteria have been conducted (e.g. Burnham & Anderson, 1998; Spiegelhalter et al., 2002; Liddle, 2007). The rationale behind the various information criteria is to provide simple means of estimating the marginal likelihoods without the need to directly estimate the corresponding complicated multidinensional integrals. These criteria are therefore based on approximations and simplifications of the underlying equations, and suitable choices of prior densities, which in fact makes the obtained estimates sub-Bayesian in the sense that the information criteria cannot be used to compare different prior models but only different likelihood models given a fixed prior density.

5.3.1. Bayesian information criterion

The Bayesian information criterion (BIC; Schwarz, 1978), sometimes called the Schwarz information criterion (SIC), is a way of approximating the integrals in Bayes factors under some simplifying assumptions.

The integral in Eq. (20) contains the likelihood of the model parameters and the prior density. If the likelihood function resembles a multimodal

Gaussian density in the vicinity of its maximum likelihood value θ_{ML}, such that $\theta_{ML} = \arg\max_{\theta\in\Omega} l(m_i|\theta)$, the likelihood function can be approximated as

$$l(m_i|\theta) \approx l(m_i|\theta_{ML}) \exp\left[-\frac{1}{2}(\theta-\theta_{ML})^T V(\theta_{ML})^{-1}(\theta-\theta_{ML})\right]. \quad (31)$$

This is called the Laplace's method of approximation. Now, assuming a uninformative constant prior – that is, assuming at least that the prior is approximately constant in the hypervolume of the parameter space that contains the highest values of the likelihood, i.e. that $\pi(\theta) = 1$ – the integral in Eq. (20) becomes

$$\begin{aligned} P(m_i) &\approx l(m_i|\theta_{ML}) \\ &\times \int_{\theta\in\Omega} \exp\left[-\frac{1}{2}(\theta-\theta_{ML})^T V(\theta_{ML})^{-1}(\theta-\theta_{ML})\right] d\theta \\ &= l(m_i|\theta_{ML})(2\pi)^{K/2} \left\| V(\theta_{ML})^{-1} \right\|^{-1/2}, \end{aligned} \quad (32)$$

where K is the number of parameters in the parameter vector θ ($K = \dim\Omega$) and V is the covariance matrix of the ML estimate. Now, for i.i.d. measurements, the covariance matrix can be written as $V(\theta_{ML})^{-1} = N_i V_1(\theta_{ML})^{-1}$, where V_1 is the covariance matrix from only one measurement and N_i is the amount of the measurements in m_i, as before. Finally,

$$\begin{aligned} 2\log P(m_i) &\approx 2\log l(m_i|\theta_{ML}) - K\log N_i + K\log 2\pi \\ &+ \log\left\| V_1(\theta_{ML})^{-1} \right\|. \end{aligned} \quad (33)$$

Dropping the last two terms, because they are negligible for large N_i, yields the traditional form for BIC as

$$-2\log P(m_i) = -2\log l(m_i|\theta_{ML}) + K\log N_i. \quad (34)$$

Now, the smaller this value is, the better the model. The first term in the BIC can be thought of as a measure of goodness of the fit – in fact, for Gaussian likelihoods, this term is the common sum of squared residuals. The second term is sometimes referred to as a penalising term that increases the BIC value for more complicated models. Hence, the model with the smallest BIC value is balanced between good fitting and simplicity, as a good model should.

Under these assumptions, the Bayes factors can be approximated and the corresponding model probabilities calculated as

$$P(\mathcal{M}_j|m) = \frac{P(m|\mathcal{M}_j)P(\mathcal{M}_j)}{\sum_{k=1}^{q} P(m|\mathcal{M}_k)P(\mathcal{M}_k)}$$
$$= \frac{l(m|\theta_{ML,j}, \mathcal{M}_j)N^{-\frac{1}{2}K_j}P(\mathcal{M}_j)}{\sum_{k=1}^{q} l(m|\theta_{ML,k}, \mathcal{M}_k)N^{-\frac{1}{2}K_k}P(\mathcal{M}_k)}. \quad (35)$$

However, it needs to be remembered that this approximation is only valid if the amount of measurements is large and if the parameter PDF is close to a Gaussian multivariate density in the vicinity of the ML estimate. If these conditions do not hold, the Bayes factors cannot be approximated by using the BIC values.

We note that if the prior does not have an uninformative density, an information criterion similar to the BIC can still be obtained. In that case, the ML estimate of the parameters can be simply replaced by the corresponding MAP estimate without any loss of generality.

5.3.2. The Akaike information criterion

The Akaike information criterion (AIC; Akaike, 1974; Hurvich & Tsai, 1989) is a model selection criterion based on the Kullback-Leibler divergence. The Kullback-Leibler (KL) divergence (Kullback & Leibler, 1951), sometimes called the relative entropy, $D_{KL}\{g||f\}$, is defined for continuous probability density functions f and g as

$$D_{KL}\{f||g\} = -\int_{\theta\in\Omega} f \log \frac{f}{g} d\theta. \quad (36)$$

This expression is commonly interpreted as consisting of the PDF f that is being compared to some 'true' density g that actually produces the measurements. However, KL divergence is not symmetric. Generally, for PDFs f and g, $D_{KL}\{f||g\} \neq D_{KL}\{g||f\}$. Moreover, the K-L divergence does not satisfy the triangle inequality, which means that it is not a metric and cannot therefore be thought of as a measure in the strict meaning of the word[7].

The K-L divergence between a candidate model with likelihood function $l(\theta|m)$ and an the underlying "true" model with likelihood function

[7]The fact that it is not a measure is the reason it is called a "divergence", which implies a way of estimating how different two densities are but not measuring this difference.

$l(m|\theta')$ leads to the AIC that states that the model for which the quantity $2\log l(m|\hat{\theta}_{ML}) - 2K$ has the greatest value should be referred to as the model with the greatest amount of support by the data (Akaike, 1974). As increasing the likelihood function clearly increases this quantity, increasing the number of parameters (K) decreases it providing the Occam's razor to the AIC. From the Bayesian perspective, this criterion is very naive as different priors cannot used when applying it and becasue it is only valid when the amount of measurements is much greater than the amount of free parameters in the model. A formal derivation of the AIC can be found in e.g. Burnham & Anderson (1998) and a general derivation of the AIC and its small-sample approximation (Hurvich & Tsai, 1989) can be found in (Cavanaugh, 1997).

The AIC and BIC are discussed in the astronomical context in e.g. Liddle (2007).

5.3.3. Other information criteria

Several other information criteria, based on different assumptions, have been introduced (see e.g. Bozdogan, 1989; Spiegelhalter et al., 2002; Konishi & Kitagawa, 2008). Among these are the Kullback information criterion (KIC; Cavanaugh, 1999) and its modification for small sample size (KICc; Hafidi & Mkhadri, 2006), the Takeuchi information criterion (TIC; Takeuchi, 1976), and the deviance information criterion (DIC; Spiegelhalter et al., 2002). However, the relative performance of these criteria is not generally known and therefore, they should be used with care in applications. If it is possible to calculate the Bayes factors directly, it should always be preferred over the various criteria.

5.4. Prior probabilities

In Eq. (19), there are prior probabilities of the different models denoted as $P(\mathcal{M}_i)$ for $i = 1, ..., k$. These probabilities represent the prior beliefs, or probabilities, of how well the different models were expected to describe the data before it was obtained. It is a rather common practice to set these probabilities equal for all i, but such a choice is only one subjective (discrete) distribution for parameter i. In fact, it would be very unusual if all the models were indeed expected to describe the data equaly well *a priori* in practice because the collections of models $\mathcal{M}_i$ is not chosen randomly out of the set of all possible models, but the models that are being compared have their interpretations as reference or baseline models, including a possible "null hypothesis", alternative models providing descriptions of one or some

phenomena whose existence is under investigation, and even models that might not have clear physical interpretations or models that describe some very speculative theoretical aspects whose validity is considered to be rather low *a priori*[8]. We do not discuss these prior probabilities here in detail because they depend heavily on the exact application and the corresponding interpretations of the models and their parameters.

As an example, we consider a problem of selecting between models that explain the data as arising from a superposition of j Keplerian signals, as is commonly the case when e.g. searching for planets by using the radial velocity or transit photometry data. It can readily be argued that when comparing models with $j = k$ and $j = k + 1$, the latter should have a lower prior probability because if there are k planets orbiting a given star, there are less dynamically stable orbits left in the system for the $k+1$th planet (Tuomi, 2012). However, one could equally well argue that when k low-mass planets have been detected in a given exoplanetary system, it is more probable *a priori* that there are more planets that have not yet been detected because low-mass planets are commonlly found in systems with high multiplicity (e.g. Tuomi, 2012; Anglada-Escudé et al., 2013; Tuomi et al., 2013a). Therefore, it is clear that such model probabilities are completely subjective choices.

6. Is the model good enough?

When comparing different models in the Bayesian context, it is only possible to determine which model *out of the selected collection of models* represents the data the best in terms of having the greatest posterior probability as described in Eq. (19). In fact, such model comparison results could be obtained even in the case that none of the models in this *a priori* selected collection of candidate models describes the data very well. This problem has been pointed out recently in Tuomi et al. (2011) and it has significant implications to model comparison problems.

For instance, regardless of how good a given candidate model $\mathcal{M}$ is, i.e. how high its posterior probability is, it is always possible that given a different

[8]Hence the common proverb: extraordinary claims require extraordinary evidence; which refers to the fact that when the prior probability of a given hypothesis is very low, it is necessary to obtain considerable amounts of evidence in favour of it in terms of Bayesian evidences to overcome the *a priori* low probability and to give the hypothesis a high posterior probability.

collection of candidate models, this model $\mathcal{M}$ would have been among the most poorly performing models. In other words, Bayesian model comparison results are only valid with respect to the other (subjectively selected) models and they do not have any general interpretation because a general measure for model goodness does not exist in the Bayesian context. Thus, it is the responsibility of the statistician to make sure that the set of candidate models contains at least few models that are realistic and represent the data well in practice.

Following Tuomi et al. (2011), we assume that a model $\mathcal{M}$ has been constructed such that it describes measurements m_i and m_j with a likelihood function $l(m_i, m_j|\theta)$. We also assume that there is another model $\mathcal{M}'$ that describes the two measurements with the likelihood function $l(m_i, m_j|\theta_i, \theta_j) = l(m_i|\theta_i)l(m_j|\theta_j)$, i.e. that the measurement m_i is described by the parameter θ_i and m_j by parameter θ_j and that the two measurements and parameters are independent. As was demonstrated by Tuomi et al. (2011), there can be other parameters as well, but they do not affect the results and we do not discuss that possibility for simplicity.

With these assumptions, it can be seen that the marginal integral in Eq. (20) given model $\mathcal{M}'$ can be written as

$$\begin{aligned} & P(m_i, m_j|\mathcal{M}') \\ &= \int_{\Omega\times\Omega} l(m_i, m_j|\theta_i, \theta_j, \mathcal{M}')\pi(\theta_i, \theta_j|\mathcal{M}')d(\theta_i, \theta_j) \\ &= \prod_{k=i,j} \int_{\Omega} l(m_k|\theta_k, \mathcal{M}')\pi(\theta_k|\mathcal{M}')d\theta_k = P(m_i|\mathcal{M})P(m_j|\mathcal{M}), \end{aligned} \quad (37)$$

where the last equality follows from the fact that the models $\mathcal{M}$ and $\mathcal{M}'$ are identical given only one measurement. Given the Eq. (37), and using the common comparison of the two models, it can be seen that if

$$P(m_i, m_j|\mathcal{M}) < \frac{s}{1-s} P(m_i|\mathcal{M})P(m_j|\mathcal{M}) \quad (38)$$

holds for some small threshold probability s, it can be concluded that the measurements m_i and m_j cannot be modelled with the same parameter θ, but different parameters (corresponding to the model $\mathcal{M}'$) should be used instead. Therefore, we say that a model $\mathcal{M}$ is an inadequate description of two data sets if it holds that

$$B(m_i, m_j) = \frac{P(m_i, m_j)}{P(m_i)P(m_j)} < r, \quad (39)$$

where we have dropped the model from the notation and define $r = s(1-s)^{-1}$ (Tuomi et al., 2011). This condition, called the Bayesian model inadequacy criterion (BMIC), can be generalised for several measurements as

$$B(m_1, ..., m_N) = \frac{P(m_1, ..., m_N)}{\prod_i P(m_i)} < r. \tag{40}$$

The BMIC in Eqs. (39) and (40) has an interesting interpretation in terms of the K-L information discussed in the context of the Akaike information criterion in Section 5.3.2. We define the information loss, or the information lost when moving from the posterior density back to the prior, in the K-L sense as

$$D_{KL}\{\pi(\theta)||\pi(\theta|m)\} = \int_\Omega \pi(\theta) \log \frac{\pi(\theta)}{\pi(\theta|m)} d\theta. \tag{41}$$

However, writing this information loss for several measurements $m_i, i = 1, ..., N$, leads to

$$\begin{aligned} D_{KL}\{\pi(\theta)||\pi(\theta|m_1, ..., m_N)\} &= \sum_{i=1}^{N} D_{KL}\{\pi(\theta)||\pi(\theta|m_i)\} \\ &+ \log B(m_1, ..., m_N), \end{aligned} \tag{42}$$

where the Bayes factor B is the same one shown in Eq. (40). This means that the BMIC is related to the information content of the measurements in a fundamental way. It is easy to verify that if $B(m_1, ..., m_m) \leq 1$ holds, Eq. (42) implies immediately that

$$D_{KL}\{\pi(\theta)||\pi(\theta|m_1, ..., m_N)\} \geq \sum_{i=1}^{N} D_{KL}\{\pi(\theta)||\pi(\theta|m_i)\}, \tag{43}$$

i.e. that the combined set of measurements contains more information than the sum of the individual ones in terms of K-L information loss (Tuomi et al., 2011). However, such simple conclusions cannot be derived by using information gain, i.e. the K-L divergence $D_{KL}\{\pi(\theta|m)||\pi(\theta)\}$, although an expression that is equivalent to that in Eq. (43) can be obtained (Tuomi et al., 2011).

7. The inverse problem of exoplanet detection

Because of the curious nature of human beings, the question of whether there are planetary systems, and in particular habitable planets enabling the

existence of life elsewhere in the universe, has been asked by several great thinkers throughout history. However, trials of answering this question have remained speculative and sophisticated guesses until a few years ago. Ever since the discovery of the first extrasolar planet (Wolszczan & Frail, 1992) and the discovery of the first such planet orbiting a Solar-type star (Mayor & Queloz, 1995), depending on the exact detection criteria, several hundred planetary companion candidates to nearby stars have been found (see e.g. Schneider et al., 2011; Wright et al., 2011a). This has partially enabled scientific answers to the above question – a question that has been disturbingly difficult to answer in the past, and remains so even today. We can now confidently state that planets and planetary systems are very common in our galaxy and therefore likely elsewhere in the universe as well. Furthermore, based on the first examples of systems with several super-Earths, it can be said confidently even based on radial velocity surveys that planets of terrestrial size are abundant in the Solar neighbourhood and thus likely elsewhere as well (e.g. Mayor et al., 2009a,b; Bonfils et al., 2013; Anglada-Escudé et al., 2013; Tuomi et al., 2013a,b). This conclusion is reinforced when looking at the population of transiting planets in the Kepler field (e.g. Howard et al., 2012; Dressing & Charbonneau, 2013; Morton & Swift, 2013). However, we still do not know how commonly planetary systems are similar to ours and – despite some very recent promising results (Bonfils et al., 2013; Dressing & Charbonneau, 2013; Kopparapu et al., 2013; Tuomi et al., 2013c) – how commonly Earth-like planets are orbiting their host-stars within the limits of the local habitable zones, which could enable the existence of liquid water, and possibly life, on their surfaces.

A majority of the low-mass companions of stars discovered to date remain only planetary candidates because the radial velocity (RV) technique, used to detect most of them, can only be used to estimate the product of mass and the sine of orbital inclination, yielding the minimum mass for the candidate. Even though unlikely, as a consequence, the RV observations cannot rule out the possibility that some of these companion candidates are in fact brown dwarfs or low mass companion stars with inclinations close to zero. This turned out to be the case with one of the candidates, HD 33636 b (Bean et al., 2007), and is likely to be the case for other candidates as well based on pure statistical estimations. Fortunately, on average, the minimum masses are only slightly lower than the expected values of the masses – that is, if the inclinations of the planetary orbits are randomly oriented in space – and they cannot be much greater than the minimum masses in systems of two or more

planets with closely-spaced orbits because that would result in instabilities and, consequently, orbital configurations that are not physically viable in long term.

Other techniques capable of detecting extra-solar planets exist as well. Photometric transit observations have been used successfully to detect several planetary companions – not merely candidates but quite confidently real planets – orbiting their host stars on close-in orbits (e.g. Howard et al., 2012; Dressing & Charbonneau, 2013, and references therein). Also, gravitational microlensing has been successful in a few lucky instances (e.g. Gaudi et al., 2008; Dong et al., 2009); direct imaging has yielded the first pictures of extrasolar planets (e.g. Kalas et al., 2008; Lagrange et al., 2008; Marois et al., 2008); and the first astrometric detections, after decades of failure (e.g. van de Kamp, 1969), have been made successfully recently (Pravdo & Shaklan, 2009) by targeting a very low mass star in the Solar neighbourhood – although the discovery of (Pravdo & Shaklan, 2009) has been subsequently disputed by Anglada-Escudé et al. (2010) and Bean et al. (2010).

7.1. What is a positive detection?

As always when dealing with measurements, it needs to be defined objectively when the value of the desired quantity has been detected meaningfully as opposed to having uncertainties that do not enable any conclusions either way. When detecting extrasolar planets, this question is reduced to: when is the signal of a planetary companion conclusive and the companion can be said to have been detected? On Bayesian grounds, this question is equivalent to: when is $P(\mathcal{M}_{k+1}|m) > \alpha P(\mathcal{M}_k|m)$, or when is the Bayesian probability of a $k+1$ planet model ($\mathcal{M}_{k+1}$) greater than a corresponding probability of a model with k planetary companions ($\mathcal{M}_k$) given some confidence limit defined by parameter α?

This approach is different from the common criterion of positive detections by periodogram-based analyses, where the detection is considered positive if one of the periodogram peaks is higher than some false alarm probability (FAP)[9] (e.g. Cumming, 2004). However, this approach can yield false negative results if there are severe gaps in the data or if the observational baseline is longer than the orbital period. Furthermore, it can result in

[9]Authors using such methods do not typically discuss the possibility that there are several peaks that exceed a given FAP because periodogram analyses apply for one signal at the time.

detections of false positives when the underlying assumptions, such as Gaussianity and independence of the measurements, are not satisfied (Mayor et al., 2009b; Vogt et al., 2010; Gregory, 2011; Tuomi, 2011; Vogt et al., 2012). The situation is even worse when there are several signals with amplitudes comparable to the noise levels. In such cases, it is possible that none of the signals get detected confidently with periodograms. False positives are also possible if some alias of a periodic signal is mistakenly considered to be the real signal (e.g. Udry et al., 2007; Mayor et al., 2009b). However, it is still possible to find the inverse solution, the orbital parameters and minimum masses, in these cases using some global solution method, such as MCMC (e.g. Ford, 2005, 2006; Tuomi & Kotiranta, 2009; Tuomi, 2012; Tuomi et al., 2013a). Even though this could mean that the estimates of the probability densities of the orbital parameters are broad, at least some confidence limits would be available and the detection could be considered trustworthy.

Additional criteria have to be satisfied as well. According to Tuomi (2012), a signal can only be said to have been detected if its amplitude is statistically distinguishable from zero with some chosen confidence level and if the period can be well-constrained from above and below. The amplitude has to be constrained from below because otherwise it would remain consistent with zero implying that the signal is not statistically significant. Furthermore, the period has to be constrained from above and below because otherwise it would not be possible to call the signal periodic. Together with the probabilistic detection threshold (e.g. Ford & Gregory, 2007; Feroz et al., 2011; Tuomi, 2012), these additional criteria have been applied recently to radial velocity planet searches (e.g. Tuomi, 2012; Tuomi et al., 2013a,b; Tuomi & Anglada-Escudé, 2013; Anglada-Escudé et al., 2013). They are also applicable to astrometric and transit photometry data without modification.

7.2. Bayesian inference of different measurements

With different measurements available, it is possible to extract more information from the observed system by combining these measurements than when they are used separately. This fact was demonstrated for RV and astrometric exoplanet observations of a planetary companion in a circular orbit by Eisner & Kulkarni (2002) and is directly implied if the model used to analyse the measurements does not satisfy the BMIC (Tuomi et al., 2011). In Tuomi et al. (2009), it was shown that the observational baselines of astrometry and RV measurements can be as short as 25% of the orbital period of the companion for a positive detection, whereas they cannot generally be much

shorter than the orbital period if the measurements are used separately in the analysis.

In the case of RV and astrometric data, the Bayesian inference means simply the updating of the prior density by the likelihood functions of data from both sources. Hence, the posterior can be written as

$$\pi(\theta|m_{\rm RV}, m_{\rm A}) = \frac{l(m_{\rm RV}|\theta)l(m_{\rm A}|\theta)\pi(\theta)}{P(m_{\rm RV}, m_{\rm A})}, \tag{44}$$

where it has been assumed that the RV and astrometric measurements are independent.

The advantages of the Bayesian inference of these two sources of information with respect to using the sources separately are essentially caused by correlations between the inertial reference frame parameters and the orbital parameters (Eisner & Kulkarni, 2001a,b, 2002; Tuomi et al., 2009). In the inference these correlations cancel one another to some extent resulting in better constraints for the orbital parameters. This in turn makes it possible to have shorter observational baselines than assumed conventionally – usually it is assumed that these baselines have to exceed the orbital period. According to the results presented in Tuomi et al. (2009), this assumption is not exactly true when accurate RV and astrometric measurements are both available.

Because the generalisation of Eq. (44) to more than two sources of measurements is obvious, we do not write it explicitly here. However, if a nearby system for which astrometric and RV measurements are available has a favourable inclination such that planetary transits can be observed photometrically, this transit data can be naturally combined with the other data sources using Bayesian inference. Furthermore, as obtaining the planetary properties such as mass, semi-major axis, and radius, depends on the observed or estimated stellar properties, any information on these properties could be incorporated in the Bayesian inference in a natural way – after all, stellar mass and radius, together with effective temperature and luminosity that are needed in estimating the location of the habitable zone in the system (Selsis et al., 2007; Kopparapu et al., 2013), are simply common parameters whose estimation can be performed in a fully Bayesian manner.

7.3. Astrometric "snapshots"

There are potentially even more dramatic advantages in the inference of RV and astrometric measurements. Astrometric snapshots are defined as

astrometric measurements with an observational baseline $T_{\rm A} \ll P$, where P is the orbital period of a planet orbiting the target star. Clearly, with these kinds of astrometric observations available, it would be next to impossible to make a positive detection of a planetary companion with the orbital period of P. The planetary signal would resemble a linear trend in the data, with possibly little curvature, and the stellar wobble caused by the companion would be confused with the inertial reference frame parameters. However, if the observational baseline of RV measurements is at least $T_{\rm RV} \approx P$, it can be shown that $T_{\rm A} \approx \frac{1}{10}P$ is sufficient for the detection of the true mass of the planetary companion (Tuomi et al., 2009) given reasonable assumptions on the nature of the observational precision.

In Tuomi et al. (2009), it was also shown that in the case of the Jupiter twin HD 154345 b (Wright et al., 2008) that has an orbital period of 9.1 years, with high precision astrometric observations available, given a sufficient precision, $T_{\rm A} \approx 0.8$ years is sufficient for the determination of the true mass of the companion – a demonstration of the usefulness of Bayesian inference of RV and astrometric data in a snapshot scenario. The Bayesian inference of RV and astrometry remains to be tested with real data but in principle these advantages are available when appropriate data sets from future space-telescopes become available.

7.4. Dynamical information

Dynamical analyses can be used as an additional source of information if there are more than one planetary companions (or brown dwarfs; or if the target is a stellar binary or has even higher multiplicity) orbiting a given target star. Since close encounters necessarily make the system prone to instability, a too low orbital spacing can be shown to be unstable and the densities of the orbital parameters and the planetary masses can be constrained more accurately by eliminating unstable subspaces of the parameter space. This procedure was attempted in the case of HD 11506 (Tuomi & Kotiranta, 2009) but all the parameter values drawn from the posterior density of the two-planet model were found to be stable. The procedure was more successful when analysing the velocity data of HD 40307 (Tuomi et al., 2013a), but in that case only the highest eccentricities were excluded from the solution because they corresponded to orbital configurations that were not dynamically viable in the long-term.

Since the RV method can only be used to obtain a lower limit for the planetary masses, their planetary nature remains uncertain unless upper limits

can be set for the masses as well. Fortunately, if there are several planetary candidates orbiting a star, it is possible to derive an upper-estimate for the planetary masses and to confirm their planetary nature. This is possible because if the inclination of the system approaches zero, the masses of the planetary companions that produce the observed RV signal approach infinity. Therefore, at some inclination in between, the system becomes unstable due to the gravitational interactions of the massive companions, and this limiting inclination can be used to calculate an upper limit for the corresponding masses.

Adding dynamical constraints will thus help tightening the BCS of the orbital parameters. Denoting the dynamical information as $\mathcal{S}$, the Bayesian inference in Eq. (44) can then be written as

$$\pi(\theta|m,\mathcal{S}) = \frac{l(m|\theta)l(\mathcal{S}|\theta)\pi(\theta)}{P(m,\mathcal{S})}, \tag{45}$$

where it has been assumed that the measurements m and the dynamical information are independent, although this is not necessarily the case as it would be impossible to obtain measurements corresponding to unstable orbital configurations, unless the observed planetary system was in a chaotic state distinguished by e.g. close-encounters that would result in bodies escaping the system or collitions between them. Therefore, Bayesian inference is a powerful tool for combining the information in several sources of data, whether this data consists of measurements or dynamical analyses, or in fact, any kind of additional information from any available source.

In Tuomi (2012), a simple analytical approximation of Lagrange stability (e.g. Barnes & Greenberg, 2006) was used to estimate the stability of a given parameter vector. Although this criterion does not take into account orbital resonances and is only valid for two planets at the time, it can still be used to define the "dynamical likelihood" $l(\mathcal{S}|\theta)$ in Eq. (45). According to this criterion, two planets with masses of μ_1 and μ_2 as fractions of the total mass of the system (M) are on stable orbits if it holds that

$$\alpha^{-3}\Big(\mu_1 - \frac{\mu_2}{\delta^2}\Big)\left(\mu_1\gamma_1 + \mu_2\gamma_2\delta\right)^2 > 1 + \mu_1\mu_2\Big(\frac{3}{\alpha}\Big)^{4/3}, \tag{46}$$

where $\alpha = \mu_1 + \mu_2$, $\gamma_i = \sqrt{1-e_i^2}$, $\delta = \sqrt{a_2/a_1}$, e_i is the eccentricity, and a_i is the semimajor axis.

If, for instance, we set the likelihood such that $l(\mathcal{S}|\theta) = c$ when the criterion in Eq. (46) holds and zero otherwise, it is possible to use this

criterion to rule out likely unstable orbital solutions from the BCS. Clearly, it is possible to consider the product of $l(\mathcal{S}|\theta)\pi(\theta)$ as a prior instead of only $\pi(\theta)$, but this is simply a matter of taste and does not affect the conclusions in any way.

7.5. Modelling low-amplitude signals in RV data

Finally, we describe briefly the currently used modelling strategies in analyses of radial velocities of nearby stars. Because prior densities are discussed extensively in Ford & Gregory (2007), Tuomi (2012), Tuomi & Anglada-Escudé (2013), and Anglada-Escudé et al. (2013), we do not discuss them here. Instead, we discuss the choice of likelihood models that has been a subject of significant improvements during the past year.

We start from the observation of Baluev (2012) and Tuomi et al. (2013b) that radial velocity noise is neither white nor Gaussian in general. Therefore, instead of applying a simple white noise model with a Gaussian distribution that leads to least-squares minimisations, we consider other options. For instance, Baluev (2012) and Tuomi et al. (2013b) observed that high-precision RV noise contains correlations that can give rise to red noise. The approach of Tuomi et al. (2013b), as well as that of Tuomi et al. (2013a), was to use moving average (MA) models to take these correlations into account. Generalising this approach by following the considerations of Scargle (1981) and Tuomi et al. (2013b), we write the general RV model as

$$m_{i,l} = f_k(t_i) + \gamma_l + \dot{\gamma} t_i + \epsilon_{i,l} + \sum_{j=1}^{q} \omega_j m_{i-1,l} + \sum_{j=1}^{p} \phi_j \epsilon_{i-1,l}, \qquad (47)$$

where f_k is a function describing the superposition of k planetary signals[10]; parameters γ_l are the reference velocities of each telescope-instrument combination denoted by using the subindex l; $\dot{\gamma}$ represents the possible linear acceleration in the data set due to a stellar or substellar companion on a long-period orbit, or caused by secular or perspective acceleration if it has not been removed from the data; $\epsilon_{i,l}$ is a random variable describing the excess white noise that is usually represented as consisting of two parts as $\epsilon_{i,l} = \epsilon_i + \epsilon_l$, where the former is referred to as instrument noise and the

[10]We note that this description can be fully Newtonian or post-Newtonian, and it can also take into account planet-planet interactions.

latter as the "stellar jitter" or noise caused by the stellar surface; ω_j are autoregressive (AR) components of the model and ϕ_j are the corresponding moving average components. Generally, this model is an ARMA(p,q) model (see e.g. Tuomi et al., 2013b).

We note that while the model defined in Eq. (47) is certainly more general than the commonly used model that assumes that the measurements are independent and identically distributed according to the Gaussian distribution, it is not by any means perfect in the sense that unevenly spaced data cannot necessarily be described very accurately by using such ARMA models because the time-difference of the subsequent measurements is not constant. This problem can be overcome by using exponential smoothing such that $\phi_j \propto \exp(-\delta t/\tau)$, where δt is the time-difference of two measurements and τ is the time-scale of this smoothing function (Baluev, 2012; Tuomi et al., 2013b). However, it is not certain whether the correlations have only one time-scale instead of several; whether the ARMA process is a suitable description of RV data at all; and how should the white noise component $\epsilon_{i,l}$ be selected in practice, because even though such a white noise component might indeed exist, it is not certain whether the common Gaussian density is a sufficiently accurate description in all cases.

Together with the problems of assessing the numbers of planetary companions in the data (k) and the numbers of ARMA components (p and q) needed in the analyses, finding the best statistical model for radial velocity data is a complicated statistical problem in practice. Problems in finding suitable likelihood models might be one of the reasons behind the controversy in the number of planet orbiting GJ 581 (Bonfils et al., 2005; Udry et al., 2007; Mayor et al., 2009b; Vogt et al., 2010, 2012; Tuomi, 2011; Baluev, 2012). As it is currently uncertain whether the number of planets orbiting GJ 581 is three or six or something in between, this problem can only be solved by refining the modelling strategies and by combining all the existing data from HARPS (Mayor et al., 2009b), HIRES (Vogt et al., 2010) and other spectrographs in the Bayesian manner to obtain trustworthy results. This only demonstrates that it is crucial to improve the statistical descriptions of the data and to apply the best possible statistical techniques when using them in practice.

8. Conclusions and discussion

Discrete inverse problems arise whenever measurements are made and statistical tools are used to extract information from them. Clearly, inverse problems are then everywhere – the only differences between them arise from the detailed aspects of the measurements: what kind of a system they describe and what is the primary interest of the researcher. If several sources of information are available, such as different kinds of measurements describing different features of the same system, the information in these measurements can be combined by using Bayesian inference of the different measurements. This procedure is called the problem of finding the solution to the multidata inverse problem.

When finding this solution, it is essential to have a model that describes the measurements of all the available sources in a consistent manner. Therefore, finding a suitable model is of essence when solving multidata inverse problems.

For this reason, the problem of finding the full inverse solution reduces to two problems that need to be solved to reach a solution. The first one is that of finding the most suitable model. By "most suitable" we mean that the model has to have the highest posterior probability out of the *a priori* selected set of models expected to work (reasonably) well. If there is a large number of measurements available for this task, and if the assumptions within are not too limiting, the simple information criteria discussed in Section 5.3 can be used for this purpose. If not, the probability densities of the parameters of all the model in the set can be used to calculate the relative probabilities of the models given the measurements (Section 5). The second problem is the problem of finding the full solution given the selected model, although in practice this solution is already available if the probability densities of the model parameters have been sampled using MCMC when finding the best model.

It is necessary to model the measurements as realistically as possible. For instance, all the unknowns, including the nuisance parameters, have to be treated as free parameters of the model. This ensures that the uncertainty in these nuisance parameters is correctly transferred to the uncertainties of the other parameters via possible correlations in their PDFs. Conversely, if some parameters are fixed, i.e. given delta-function priors, the estimates of the other parameters can be severely biased because the possible correlations between the PDFs of these parameters are not being taken into account

properly.

The Bayesian methods have recently – within the last decades – started to gain increasing amounts of popularity in astronomy for several important reasons. First, the increased computational capabilities have made it possible to use e.g. MCMC method that require large amounts of computer memory in high-dimensional problems. Second, the availability of efficient algorithms, such as the Metropolis-Hastings (Metropolis et al., 1953; Hastings, 1970) and the adaptive Metropolis algorithm (Haario et al., 2001), have made it possible to calculate the chains in an efficient manner. Third, nowadays measurements of great scientific value are commonly difficult to make, increasing the requirements for financial resources. This means that the analysis of the valuable measurements has to be as efficient as possible, and the Bayesian methods provide the required power and generality for these purposes.

With powerful statistical methods available, astronomers can focus on the most important problem left – the problem of choosing the set of models to be compared. This task is by far the most challenging in the chain of analyses leading to the inverse solution, because it cannot be automatised and solved by computers[11]. It requires to a great extent good scientific intuition, fundamental knowledge of the system of interest, and the ability to identify *a priori* the most important features in the system. If the selected model set or model class is poor, the solution will be limited to the poor solutions enabled by the models. On the other hand, if the selected class is good enough, it will be possible to draw significant conclusions about the system after the analyses because, at least, features that cannot exist in the system can be ruled out when their posterior probabilities turn out to be negligible. However, this problem of selecting a suitable class of models is not necessarily hard to solve. For instance, when detecting exoplanets with the RV method, the model class consists of models with k Keplerian signals in the RV data (Tuomi & Kotiranta, 2009; Tuomi, 2011, 2012), and possibly diferent noise descriptions (Tuomi et al., 2013b). In such reasonably simple systems, there are not many options available and the model set can be safely assumed to be well selected.

Finally, with a full solution to an inverse problem available, the parame-

[11] At least this is the case without considerable innovations in machine learning and artificial intelligence.

ter posterior densities can be used to predict the behaviour of the modelled system. The most trustworthy way of calculating these predictions is to generate a sample by drawing values from the parameter density and by calculating the corresponding sample of a density of the predicted quantity. In this manner, the full solutions can be used most efficiently in practice, and it can be made sure that the information in the measurements is inferenced directly to the prediction densities. This has enabled the detection of the most populated planetary system known to date (Tuomi, 2012) together with the detections of the first candidate habitable planets in the Solar neighbourhood (Tuomi et al., 2013a,b).

References

Akaike, H. 1974, IEEE Transactions on Automatic Control, 19, 716

Anglada-Escudé, G., Skholnik, E. L., Weinberger, A. J., et al. 2010, ApJ, 711, L24

Anglada-Escudé, G. & Butler, R. P. 2012, ApJS, 200, 15

Anglada-Escudé, G., Tuomi, M., Gerlach, E., et al. 2013, A&A, 556, A126

Balan, S. T. & Lahav, O 2009, MNRAS, 34, 1936

Baluev, R. V., 2012, MNRAS, 429, 2052

Barnes, R. & Greenberg, R. 2006, ApJ, 647, L163

Bean, J. L., McArthur, B. E., Benedict, G. F., et al. 2007, AJ, 134, 749

Bean, J. L., Seifahrt, A., Hartman, H., et al. 2010, ApJ, 711, L19

Bonfils, X., Forveille, T., Delfosse, X., et al. 2005, A&A, 443, L15

Bonfils, X., Delfosse, X., Udry, S., et al. 2013, A&A, 549, A109

Boss, A. P. 1997, Science, 276, 1836

Box, G. E. P. 1976, J. Am. Stat. Ass., 71, 791

Bozdogan, H. 1989, Psychometrika, 52, 345

Burnham, K. P. & Anderson, D. R. 1998, Model selection and multimodel inference: A practical information-theoretic approach, Springer Verlag, New York

Cavanaugh, J. E. 1997, Stat. Prob. Lett., 33, 201

Cavanaugh, J. E. 1999, Stat. Prob. Lett., 42, 333

Chib S. 1995, J. Am. Stat. Ass., 90, 1313

Chib S. & Jeliaszkov I. 2001, J. Am. Stat. Ass., 96, 270

Clyde, M. A., Berger, J. O., Bullard, F., et al. 2007, Statistical Challenges in Modern Astronomy IV, Babu, G. J. & Feigelson, E. D. (eds.), ASP Conf. Ser., 371, 224

Cumming, A. 2004, MNRAS, 354, 1165

Dong, S., Bond, I. A., Gould, A., et al. 2009, ApJ, 698, 1826

Dressing, C. D. & Charbonneau, D. 2013, ApJ, accepted (arXiv:1302.1647)

Efron, B. 1979, Ann. Stat., 7, 1

Eisner, J. A. & Kulkarni, S. R. 2001a, ApJ, 550, 871

Eisner, J. A. & Kulkarni, S. R. 2001b, ApJ, 561, 1107

Eisner, J. A. & Kulkarni, S. R. 2002, ApJ, 574, 426

Evett, I. W. 1991, Implementing Bayesian methods in forensic science, Paper presented at the Fourth Valencia International Meeting on Bayesian Statistics

Feroz, F., Balan, S. T., & Hobson, M. P. 2011, MNRAS, 415, 3462

Fisher, R. A. 1922, JRSS, 85, 597

Ford, E. B. 2005, AJ, 129, 1706

Ford, E. B. 2006, ApJ, 642, 505

Ford, E. B. & Gregory, P. C. 2007, Statistical Challenges in Modern Astronomy IV, Babu, G. J. & Feigelson, E. D. (eds.), ASP Conf. Ser., 371, 189

Gaudi, B. S., Bennett, D. P., Udalski, A., et al. 2008, Science, 319, 927

Gelman, A. G., Roberth, G. O., and Gilks, W. R. 1996. Efficient Metropolis jumping rules. In. Bernardo, J. M., Berger, J. O., David, A. F., and Smith, A. F. M. (eds), Bayesian Statistics V, pp. 599-608. (Oxford: Oxford University Press).

Gelman, A. & Meng, X.-L. 1998, Stat. Sci., 13, 163

Gray, R. O., Corbally, C. J., Garrison, R. F., et al. 2006, ApJ, 132, 161

Green, P. J. 1995, Biometrika, 82, 711

Green, P. J. & Mira, A. 2001, Biometrika, 88, 1035

Gregory, P. C. 2005, ApJ, 631, 1198

Gregory, P. C. 2007a, MNRAS, 381, 1607

Gregory P. C. 2007b, MNRAS, 374, 1321

Gregory, P. C. 2011, MNRAS, 410, 94

Haario, H., Saksman, E., & Tamminen, J. 2001, Bernoulli, 7, 223

Haario, H., Laine, M., Mira, A., & Saksman, E. 2006, Stat. Comp., 16, 339

Hafidi, B. & Mkhadri, A. 2006, Comp. Stat. Data An., 50, 154

Hansen B. M. & Murray, N. 2012, ApJ, 751, 158

Hastings, W. 1970, Biometrika 57, 97

Howard, A. W., Marcy, G. W., Bryson, S. T., et al. 2012, ApJS, 201, 15

Hurvich, C. M. & Tsai, C. L. 1989, Biometrika, 76, 297

Ida, S. & Lin, D. N. C. 2010, ApJ, 719, 810

Jeffreys, H. 1961, The Theory of Probability (The Oxford University Press)

Kaasalainen, M. 2011, Inverse Problems and Imaging, 5, 37

Kaasalainen, M. 2012, A&A, 543, A97

Kaipio, J. & Somersalo, E. 2005, Statistical and Computational Inverse Problems, Applied Mathematical Sciences 160, Springer, New York

Kalas, P., Graham, J. R., Chiang, E., et al. 2008, Science, 322, 1345

Kass, R. E. & Raftery, A. E. 1995, J. Am. Stat. Ass., 430, 773

Kolmogorov, A. N. 1968, IEEE Transactions on Information Theory, Vol. IT-14, 662

Konishi, S. & Kitagawa, G. 2008, Information Criteria and Statistical Modeling, Springer (Springer Series in Statistics), New York

Kopparapu, R. K., Ramirez, R., Kasting, J. F., et al. 2013, ApJ, 765, 131

Kullback, S. & Leibler, R. A. 1951, Ann. Math. Stat., 22, 76

Lagrange, A.-M., Gratadour, D., Chauvin, G., et al. 2008, A&A, 493, L21

Liddle, A. R. 2007, MNRAS, 377, L74

Lomb, N. R. 1976, Astrophys. Space Sci., 39, 447

Loredo, T. J., Berger, J. O., Chernoff, D. F., et al. 2012, Stat. Met., 9, 101

Marois, C., Macintosh, B., Barnman, T., et al. 2008, Science, 322, 1348

Mayor, M. & Queloz, D. 1995, Nature, 378, 355

Mayor, M., Udry, S., Lovis, C., et al. 2009a, A&A, 493, 639

Mayor, M., Bonfils, X., Forveille, T., et al. 2009b, A&A, 507, 4870

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., et al. 1953, J. Chem. Phys., 21, 1087

Morton, T. D. & Swift, J. 2013, ApJ, accepted (arXiv:1303.3013)

Newton, M. A. & Raftery, A. E. 1994, J. Roy. Stat. Soc. B, 56, 3

Neyman, J. & Pearson, E. S. 1928, Biometrika, 20, 175

Pearson, K. 1901, Philosophical Magazine Series 6, 11, 559

Pravdo, S. H. & Shaklan, S. B. 2009, ApJ, 700, 623

Scargle, J. D. 1981, ApJS, 45, 1

Scargle, J. D. 1982, ApJ, 263, 835

Schneider, J., Dedieu, C., Le Sinader, P., et al. 2011, A&A, 532, A79

Schwarz, G. E. 1978, Ann. Stat., 6, 461

Selsis, F., Kasting, J. F., Levrard, B., et al. 2007, A&A, 476, 1373

Skilling, J. 2006, Bayesian Statistics 8, eds. Bernardo, J. M., Bayarri, M. J., Berger, J. O., Dawid, A. P., Heckerman, D., Smith, A. F., & West, M., Oxford, UK, Oxford Univ. Press

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. 2002, J. R. Statist. Soc. B, 64, 583

Takeuchi, K. 1976, Suri-Kagaku (Math. Sci.), 153, 12 (in Japanese)

Tikhonov, A. N. & Arsenin, V. Y. 1977, Solutions of Ill-posed problems (New York: Wiley).

Tuomi, M. & Anglada-Escudé 2013, A&A, 556, A111

Tuomi, M. & Jones, H. R. A. 2012, A&A, 544, A116

Tuomi, M. & Kotiranta, S. 2009, A&A, 496, L13

Tuomi, M. 2011, A&A, 528, L5

Tuomi, M. 2012, A&A, 543, A52

Tuomi, M., Kotiranta, S., & Kaasalainen, M. 2009, A&A, 494, 769

Tuomi, M., Pinfield, D., & Jones, H. R. A. 2011, A&A, 532, A116

Tuomi, M., Anglada-Escudé, G., Gerlach, E., et al. 2013a A&A, 549, A48

Tuomi, M., Jones, H. R. A., Jenkins, J. S., et al. 2013b, A&A, 551, A79

Tuomi, M., Jones, H. R. A., Barnes, J. R., et al. 2013c, A&A, submitted

Udry, S., Bonfils, X., Delfosse, X., et al. 2007, A&A, 469, L43

van de Kamp, P. 1969, AJ, 1371, 757

Vogt, S., Butler, P., Rivera, E., et al. 2010. ApJ, 723, 954.

Vogt, S. S., Butler, R. P., & Haghighipour, R. N. 2012, AN, 333, 1

Wolszczan, A. & Frail, D. A. 1992, Nature, 355, 145

Wright, J. T., Marcy, G. W., Butler, R. P. & Vogt, S. S. 2008, ApJ, 683, L63

Wright, J. T., Fakhori, O., Marcy, G. W., et al. 2011, PASP, 123, 412

Wright, J. T., Veras, D., Ford, E. B., et al. 2011, ApJ, 730, 93