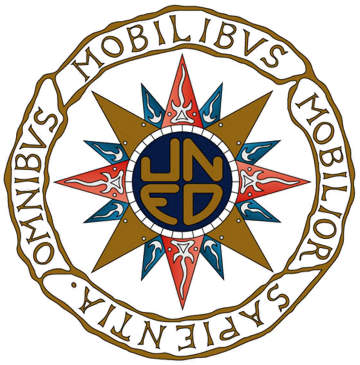
A Bayesian Graphical Model for Frequency Recovery of Periodic Variable Stars



Héctor Delgado-Ureña Poirier
ETSI Informática
UNED

A thesis submitted for the degree of
Master in Advanced Artificial Intelligence

Madrid, 2014

Dedicated to my grandparents.

Acknowledgements

I would like to express my gratitude to my supervisor and domain expert, Dr. Luis Manuel Sarro Baro, for his support, patience and assistance in writing this thesis in English.

Abstract

This thesis has been developed in the context of the recently launched European Space Agency's Gaia mission. The thesis has addressed the problem of determining the probability distributions of the real physical parameters for a variable star population, given their recovered values by the Data Processing and Analysis Consortium (DPAC) from the telemetry of the satellite. These recovered values are affected by a number of stochastic errors and systematic biases due to the aliasing phenomenon as a product of the Gaia scanning law, the optical and photometric resolution of the satellite and the algorithms used in the recovery process. The purpose of the thesis has been to model the data recovery process and infer the real distributions for the frequencies, apparent G-magnitudes and amplitudes for a Large Magellanic Cloud (LMC) classic Cepheid star population. A two level Bayesian graphical model was constructed with the aid of a domain expert to model the recovery process and a Markov chain Monte Carlo (MCMC) algorithm specified to perform the inference. The system was implemented in the declarative BUGS language. The system was trained from a set of recovered data from an artificially generated real distribution of LMC Cepheids. The system was tested by comparing the parameters of the artificially generated real distributions with the distributions inferred by the MCMC algorithm. The results obtained have shown that the system remove successfully the systematic biases and is able to infer correctly the real frequency distribution. The results have also shown a correct inference for the real apparent magnitudes in the G band. Nevertheless, the results obtained for the case of the real amplitude distribution have not allowed to establish significant conclusions.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

1.1.1 The Framework

The framework of probabilistic graphical models (PGMs) in artificial intelligence (Pearl, 1988; Lauritzen, 1996) constitutes a powerful formalism for knowledge representation and reasoning under uncertainty which combines graph and probability theory to construct probabilistic expert systems whose dependency structure between variables can be graphically represented and visually analyzed. In particular, the sub-framework of directed acyclic graphical (DAG) models, also known as Bayesian networks (Pearl, 1985) constitutes a consolidated formalism that has been traditionally and successfully applied to diagnostic problems in which it is necessary to represent causal or influence relationships between a set of discrete (categorical) domain entities. From a statistical analysis perspective the above classical DAG framework has been enriched, on the one hand, by the application of statistical inference techniques for learning both the structure and the parameters (probabilities) of the network (Spiegelhalter and Lauritzen, 1990; Cooper and Herskovits, 1992; Heckerman, 1996; Spirtes et al., 2000; Neapolitan, 2004). On the other hand, the ability of DAGs to represent the repetitive pattern of samples generation in stochastic domains with random variables belonging to different families and complex dependency relationships between them has propitiated to use this knowledge representation formalism as a standard (basis) for statistical inference. In particular, when the Bayesian paradigm[1] for statistical inference

[1] Kuhn (2012).

(Gelman and Shalizi, 2013) is applied, parameters to infer are represented explicitly in the network as random variables with a distribution which represents the prior knowledge about them and the inference is seen as the updating of this knowledge from concrete samples (the evidence). This latter approach has been so-called the Bayesian graphical models (BGMs) framework (Højsgaard et al., 2012) and nowadays is a growing area of research in many domains. This new framework is supplemented by the use of Markov chain Monte Carlo (MCMC) simulation techniques (Robert and Casella, 2004) in those situations in which inference about parameters of interest is difficult or impossible to develop based in analytic methods.

1.1.2 Astronomy: An Application Domain for Graphical Models

With the advent of the 21st century the domain of observational astronomy (Léna et al., 2012) has experienced an authentic revolution due to a sustained increase in quality, complexity, heterogeneity and volume of data (*surveys*) collected by a number of ambitious terrestrial telescopes and space missions. This has enforced a paradigm shift in the way to do astronomy and imposed the challenge of presenting all information to the researcher in an organized, centralized, preprocessed and easily accessible way. Thus, the area has become naturally an increasing application field for advanced techniques of statistics, data mining and AI (Feigelson and Babu, 2003) and, in particular, Bayesian methods (Hobson et al., 2010). The ultimate goal, if any exists in science, would be to have all significant information from all *surveys* accessible in the form of astronomical virtual observatories (AVOs) (Djorgovski et al., 2003; Solano, 2006; López Del Fresno et al., 2011).

1.1.3 The Mission

Gaia (Lindegren et al., 2008) is a recently launched space mission, by the European Space Agency (ESA), whose main objective is to make a large-scale astronomical survey of about one billion stars (≈1%) of our Galaxy and its Local Group. The satellite will scan the entire sky from a Lissajous orbit around the Sun–Earth L2 Lagrangian point for about 5 years with an unprecedented precision (microarcseconds) in position and motion measures (astrometry) for stars brighter than the 20th magnitude in the G band. It will also be able to perform multi-epoch photometry, with a total mean of 70 transits per object, and measurements of radial velocity, being the former suitable for studies of stellar variability.

During its lifetime the satellite will send to ground stations a huge amount of raw data (100TB) that shall be processed before becoming available to the research community in the form of a true scientific archive. The complex task of designing and implementing such archive and the tools to access it has been entrusted to a international software consortium, named the DPAC (Data Processing and Analysis Consortium) (Mignard et al., 2008) with the hardware support of six Data Processing Centres (DPCs) to develop their activities. The DPAC is composed by scientists, software engineers and academic institutions from more than 20 countries and organized in nine Coordination Units (CUs) each dedicated to a different aspect of the data processing. The Department of Artificial Intelligence at UNED participates in three of these units: CU7 (Variability Processing), CU8 (Astrophysical Parameters) and CU9 (Catalogue Access). In particular, CU7 is devoted to the analysis and knowledge discovery related with variable astronomical sources and consists of four working groups, involved in the development of the corresponding software packages. These packages perform tasks of Characterization, Classification and Bias Estimation for each recovered source and Statistical Quality Assessment (QA) for all sources (the complete survey) recovered in each category.

1.2 Problem Statement, Objectives and Scope

Physical parameters recovered by CU7 from preprocessed data provided by other CUs for each source belonging to a particular variable star survey are affected by a number of stochastic errors and systematic biases. These errors and biases mainly arise due to the way the satellite scans the sky, its optical and photometric resolution and the algorithms used in the recovery process. Under these conditions, the quality of the recovered survey is obviously degraded and the problem arises of how to rebuild, as far as possible, the actual statistical distribution of the survey parameters. This problem statement suggests us that its resolution could be addressed by means of the BGMs formalism and leads to state the general objective of this thesis as follows:

- To develop a Bayesian graphical model for representing the (subject to biases) generative process of an output variable star survey in CU7 and inferring the parameters of variables distributions in the input survey observed by Gaia.

In particular, we aim to achieve the following specific objectives:

- To model the dependency structure between the input physical parameters amplitude, frequency/period and apparent magnitude for a variable star observed by Gaia and the corresponding output parameters recovered by CU7.
- To state a parametric model of the probabilistic density functions (PDFs) for the above dependency structure, including an explicit representation of the biases associated to the recovery process.
- To state a parametric model based in categories to constrain the probabilities of correct/incorrect recovery for the frequency/period of a star.
- To represent the repetitive pattern associated to the generation of a sample of recovered amplitudes, frequencies/periods and apparent magnitudes from the corresponding input parameters.
- To infer the parameters (hyparameters) of the PDFs corresponding to the input physical parameters and the parameters of the submodel which constrains the probabilities of categories of recovered frequencies/periods.

With regard to the scope of the thesis the first consideration that must be made is that Gaia is not yet operational, forcing us to work with simulated data. Second, it should be noted that inputs in CU7 are preprocessed (not raw) data consisting in an astronomical time series for each star. Finally, we should note that QA is in the final phase of data analysis in CU7. In this context we assume that inputs are the real physical parameters (frequency/period, amplitude and mean apparent magnitude) which characterizes these time series but not the time series themselves. Also and in particular, we circumscribe the analysis to a simulated survey consisting in Cepheids variables stars from the Large Magellanic Cloud (LMC) to which the Deeming algorithm for frequency/period recovery has been applied.

1.3 Methodology and Resources

To achieve the objectives stated in Section 1.2 we apply the classical methodology in the Symbolic AI paradigm based on a top-down decomposition of the problem in terms of three levels of abstraction: Computational Theory, Representation and Algorithm and Implementation (Marr et al., 2010; Mira and Delgado, 2001; Newell, 1982). This methodological

approach, supplemented with specific aspects for the development of PGMs (Díez, 2010) and an evaluation phase, is developed in the following stages:

1. **Problem Analysis**. This phase is developed with the aid of an expert in the astrophysical domain and involves a manual selection (identification) of entities in the Input and Output Representation Spaces associated with the problem we aim to solve, and identification of transformations between them. The knowledge gathered in this stage includes astrophysical laws, design features of the satellite, aspects of the processing on CU7 and experimental results.

2. **System Design**. The qualitative and quantitative knowledge gathered in the first stage is reduced to the Bayesian graphical model used as knowledge representation formalism for the problem. An MCMC algorithm (a Gibbs sampling scheme) for the inference mechanism is also specified.

3. **System Implementation**. The reduction to the programming language level is made directly by using the BUGS (Lunn et al., 2009) declarative language. The inference mechanism is activated by selecting the appropriated sampling algorithms from a library accessible via the OpenBUGS software tool.

4. **Evaluation**. The model is trained from a sample of the output survey obtained by CU7 applying its processing tools to a set of instances belonging to a simulated input space. The convergence of MCMC chains for the induced parameters of the input space is assessed. Finally, a comparison between the induced parameters and the real (simulated) ones is done.

With regard to the resources employed in the thesis, besides the aforementioned OpenBUGS environment used to implement the system, we use the statistic R environment (R Core Team, 2013) for both phases of analysis and evaluation, including the CODA (Plummer et al., 2006) package for MCMC chains analysis and diagnostic of convergence .

1.4 Thesis Structure

The rest of the work is organized as follows. In Chapter 2 we review the astrophysical background underlying the problem posed in the thesis. It includes a general description of

the implied astrophysical entities, their physical parameters, the transformations between their representation spaces and the errors and biases associated to these transformations. In Chapter 3 we review the Bayesian graphical modeling framework in which is based the model proposed in the thesis and its applications to the area of astrostatistics. In Chapter 4 we complete the analysis stage started in Chapter 2 and proceed to the system design and implementation by using the mathematical and programing tools studied in Chapter 3. In Chapter 5 we evaluate our model. Finally, in Chapter 6 we summarize the work, state our conclusions and depict future lines of research.

Chapter 2

Literature Review I: Astrophysical Background

In this chapter we review the astrophysical framework underlying to the Bayesian graphical model proposed in the thesis. It is structured in two sections as follows. Section 2.1 introduces the essential astrophysical concepts necessary to understand, from a general perspective, the nature of the underlying entities, representation spaces and transformations between them. Finally, Section 2.2 is devoted to particularize the analysis for the case of Gaia and the processing tools developed in the CU7 for stellar variability studies.

2.1 Preliminary Concepts

In this section we review some basic concepts about photometry (Karttunen et al., 2007; Ashdown and Eng, 2002) and digital signal processing theory (Proakis and Manolakis, 1996; De Meyer, 2003) in the context of the observational astrophysical domain (Léna et al., 2012).

2.1.1 Photometric systems

Definition 2.1. (Spectral luminosity, intrinsic brightness). The *intrinsic brightness* is the radiant energy per unit time (power) per unit bandwidth at a frequency ν:

$$L_\nu \left[W \cdot Hz^{-1} = J \cdot s^{-1} \cdot Hz^{-1} \right] = \frac{dQ}{dtd\nu}$$

It is an intrinsic magnitude of the source independent of the distance from which the source is observed.

Definition 2.2. (Spectral flux density, apparent brightness). The *apparent brightness* is the spectral luminosity per unit area at a point on a surface:

$$F_\upsilon \left[W \cdot m^{-2} \cdot Hz^{-1}\right] = \frac{dL_\nu}{dA}$$

It is an extrinsic magnitude of the source dependent of the distance from which the source is observed.

Theorem 2.1. ***(Inverse square law)***. *For a punctual source that radiates isotropically (its radiation at a distance r is distributed evenly on a spherical surface S) the spectral flux density F_v passing through S is directly proportional to the spectral luminosity L_ν of the source and inversely proportional to the square of r:*

$$F_v = \frac{\int_S F_v ds}{4\pi r^2} = \frac{L_\nu}{4\pi r^2} \tag{2.1}$$

Definition 2.3. (Photometric filter). A *photometric filter* is a photoelectric device used to allow to enter into the detector only a determinate wavelength band (the so called *passband*) of the electromagnetic spectrum of the source.

Definition 2.4. (Photometric system, multicolor system). A *photometric system* is a set of photometric filters. Relevant to our work is the Johnson-Cousins UBVRI broadband system (Bessell, 1990) with mean wavelengths from 361 (U, Ultraviolet) to 806 nm (I_C, Infrared) and 551 nm for the V (visible) passband. In the Johnson-Cousins system, the zero magnitude are defined through Vega ($m_V = 0.030$ mag for the V passband).

Definition 2.5. (Apparent magnitude). The *apparent magnitude* in the x-band is defined by

$$m_x = -2.5 \cdot log_{10}\left(\frac{F_x}{F_{x,0}}\right) \tag{2.2}$$

, where F_x is the spectral flux density of the source for the x-band and $F_{x,0}$ is a normalizing constant (*zero-point*) equal to the spectral flux density when $m_x = 0$. It is a measure of the apparent brightness of the source.

In practice, the flux on the receptor is measured within a finite band of wavelengths and modulated by a factor $R_x(\lambda)$ due to the optical transmission of the instrument, its sensibility in the band, etc. Therefore, a more realistic expression for Eq. 2.2 is

$$m_x = -2.5 \cdot log_{10}\left(\frac{\int R_x(\lambda) F_\lambda d\lambda}{\int R_x(\lambda) F_{\lambda,0} d\lambda}\right) \tag{2.3}$$

, with $\int R_x(\lambda) d\lambda = 1$. For example, in the Johnson-Cousins UBVRI broad-band system the zero magnitude for the V (visible) band (470-700 nm) is defined through Vega (α Lyr) by (Bessell, 1990; Bessell et al., 1998)[1]

$$m_V - 0.03 = -2.5 \cdot log_{10}\left(\frac{\int V(\lambda) F_\lambda d\lambda}{\int V(\lambda) F_{\lambda,\alpha\ \mathrm{Lyr}} d\lambda}\right) \tag{2.4}$$

, being $V(\lambda)$ the V-band filter sensibility curve.

Definition 2.6. (Absolute magnitude). The *absolute magnitude* in the x-band, denoted by M_x, is defined (in a free space without stellar absorption) as the apparent magnitude at a distance of 10 parsecs ($1pc \equiv 206.26 \times 10^3 \mathrm{AU} \equiv 3.26156$ light years $\equiv 30.857 \times 10^{15} m$) from the source. It is a measure of the luminosity (intrinsic brightness) of the source.

Definition 2.7. (Distance module). Given a source at a distance r from the observer, the *distance module* in the x-band is defined as the difference $m_x - M_x$ between the apparent magnitude and the absolute magnitude of the source in the band.

Proposition 2.1. *Given a source at a distance r from the observer, the distance modulus in x-band is given by*

$$m_x - M_x = -5 + 5 \cdot log(r) \tag{2.5}$$

Definition 2.8. (Colour index). A *colour index* is the difference between two (apparent) magnitudes for a determinate multicolor system.

Definition 2.9. (Colour–colour transformation). A *colour-colour transformation* an expression that relates the colour indices of two different colour systems.

[1]Note that α Lyr has an V-mag equal to 0.03 but not equal to zero.

2.1.2 Periodic Light Curves

Definition 2.10. (Astronomical light curve). An *astronomical light curve* is a continuous function $m_x(t)$ which relates the apparent magnitude of an astronomical object (for a frequency band x) with the time.

Definition 2.11. (Astronomical time series). An *astronomical time series* is a sample $\{m_x(t_k)\}_{k=1}^{o}$ from an astronomical light curve $m_x(t)$ corresponding to a sequence $\{t_k\}_{k=1}^{o}$ of o *observation times* (*epochs*).

Definition 2.12. (Periodic folded light curve model). A *periodic folded light curve model* is a model of a periodic variable star light curve in the interval corresponding to one period, defined by

$$m_x(\varphi \mid S; \overline{m}_x, A, P) \tag{2.6}$$

where,

- The parameters $\overline{m}_x$, A and P are the mean apparent magnitude of the star, its (peak to peak) variability amplitude around $\overline{m}_x$ and its period, respectively.
- φ, verifying $0 \leq \varphi \leq 1$, is the *phase*, i.e. the variation of time in one period, defined as $\varphi(t) = \frac{\mathrm{mod}(t-t_0,P)}{P}$ with t_0 the reference epoch (e.g. $t_0 = 0$ for the phase at origin).
- S is the shape of the curve in one period. S can be: i) a periodic function like e.g. $S(\varphi) = \sin(\varphi)$ or ii) a template consisting in a (discrete) time series that provides the magnitudes as a function of the phases, in which case a function $S(\varphi)$ that can be reconstructed by interpolation is assumed.
- A is the peak-to-peak amplitude, i.e. the difference between $m_x^{\max}(\varphi) - m_x^{\min}(\varphi)$.

For our work are relevant the periodic light curves of a particular type of periodic variable stars, namely the classic Cepheids. Their main characteristics and the particular aspects of its processing in the context of the Gaia mission and the CU7 will be Studied in Section 2.3.

2.1.3 The Sampling Process in the Temporal Domain

Consider a deterministic and strictly periodic light curve $x_{\mathrm{input}}(t)$ from a periodic variable source. Seen by the telescope focal plane this deterministic temporal process becomes, in

principle, in a continuous random (stochastic) process which we can model by

$$x(t) = x_{\text{input}}(t) + \epsilon(t) \tag{2.7}$$

, where $\epsilon(t)$ is the random noise due to the measurement error for the band (given in our case by Eq.2.24) . Nevertheless, in practice, the telescope and particularly a large scale survey like Gaia, only recover a discrete realization of the process which is a sample of magnitudes $\{x(t_k)\}_{k=1}^{N}$ or *astronomical time series* with expression

$$x(t_k) = x_{\text{input}}(t_k) + \epsilon(t_k) \tag{2.8}$$

In general, the sampling process can be summarized, in the temporal domain, by the product

$$w_N(t) \cdot x(t) \tag{2.9}$$

, where$w_N(t)$ is a *data window* function defined by a finite combination of generalized Dirac delta functions as

$$w_N(t) = \sum_{k=1}^{N} \delta(t - t_k) \tag{2.10}$$

Depending on the characteristics of the observation program, the amplitude $\Delta t_i = t_{i+1} - t_i$ of temporal intervals between successive samples may vary from the uniform (even) case, with a sampling frequency $f_s = 1/\Delta t$, to completely arbitrary values. There is also a range of intermediate situations in which these amplitudes present certain patterns, say a semi-regular sampling.

Keeping in mind the analysis of the periodicity of the light curve $x_{\text{input}}(t)$ we can "extend" the definition of a strictly periodic function to the stochastic case in Equation 2.7 by means of the following two definitions.

Definition 2.13. (weak stationary stochastic process). A stochastic process $x(t)$ is *weak stationary* if the two following conditions hold:

1. $\mathrm{E}[x(t)] = \mu \quad , \forall t$
2. $\mathrm{Cov}[x(t), x(t-\tau)] = \mathrm{Cov}[x(t+l), x(t-\tau+l)] = \gamma(\tau) \quad , \forall t, l, \tau$

, that is, if the first order moment is constant for any time and the second order moments depend on the lag between times.

Definition 2.14. (autocorrelation). Given a weak stationary stochastic process $x(t)$, the *autocorrelation function* is defined by

$$\rho_x(\tau) = \frac{Cov[x(t), x(t-\tau)]}{\sqrt{Cov[x(t), x(t)]} \cdot \sqrt{Cov[x(t-\tau), x(t-\tau)]}} = \frac{\gamma(\tau)}{\gamma(0)} \tag{2.11}$$

, which is interpreted as the similarity between observations as a function of the time lag between them.

2.1.4 Analysis in the Frequency Domain

In the frequency domain, the sampling process of Equation 2.9 is represented by applying the continuous Fourier transform (CFD)

$$X_N(\nu) = \int_{-\infty}^{+\infty} x(t) \cdot w_N(t) e^{j2\pi\nu t} dt = \sum_{k=1}^{N} x(t_k) e^{j2\pi\nu t_k} \tag{2.12}$$

, whose square $|X_N(\nu)|^2 = X_N(\nu) \cdot X_N^*(\nu)$ is the *observed power*.

Similarly, for the data window of Equation 2.10 we obtain the so called *spectral window*

$$W_N(\nu) = \sum_{k=1}^{N} \int_{-\infty}^{+\infty} \delta(t - t_k) e^{j2\pi\nu t} dt = \sum_{k=1}^{N} e^{j2\pi\nu t_k} \tag{2.13}$$

, and multiplying by its conjugate we obtain the *power spectral window*

$$V_N(\nu) = W_N(\nu) \cdot W_N^*(\nu) = |W_N(\nu)|^2 \tag{2.14}$$

Finally, for the autocorrelation given by Equation 2.11 we obtain the *power spectrum,* a real and even function function defined by

$$P(\nu) = \int_{-\infty}^{+\infty} \rho_x(\tau) e^{j2\pi\nu t} dt = 2 \int_{0}^{+\infty} \rho_x(\tau) \cos t dt \tag{2.15}$$

Then, it is possible to demonstrate Deeming (1975) that the observed power is proportional to the convolution of the power spectrum of the stochastic process with the power

spectral window

$$N^{-1}\mathrm{E}\left[\left|X_N(\nu)\right|^2\right] = \mathrm{Var}(x)\cdot P(\nu)\otimes\gamma_N(\nu) = \mathrm{Var}(x)\int_{-\infty}^{+\infty} P(\nu)\,\gamma_N(\nu-\nu')\,d\nu' \tag{2.16}$$

, where $\gamma_N(\nu) = N^{-1}V_N(\nu)$ is the power spectral window normalized such that $\gamma_N(0) = 1$.

Recalling the discrete Fourier transform (DFT) of Equation 2.12 the expression

$$p(\nu) = \frac{1}{N}\left|X_N(\nu)\right|^2 = \frac{1}{N}\left|\sum_{k=1}^{N} x(t_k)\,e^{j2\pi\nu t_k}\right|^2 \tag{2.17}$$

is the so called *periodogram*. It is the tool that has been used to recover the frequencies/periods of the simulated time series in the context of our work.

2.1.5 The Aliasing Phenomenon

This phenomenon arises in its more extreme expression in a scenario of even sampling. To illustrate it, we consider then that for every k $\Delta t_k = \Delta t$ and $t_k = k\Delta t$. We assume also that N is odd and consider that sampling times are centered around $t_0 = 0$. For simplicity, we assume too that the signal $x(t)$ is deterministic. In this latter case the equivalent expression for Equation 2.16 is given by

$$X_N(\nu) = X(\nu)\otimes W_N(\nu) = \int_{-\infty}^{+\infty} X(\nu)\,W_N(\nu-\nu')\,d\nu' \tag{2.18}$$

With the imposed conditions, in the limit, for $N\to\infty$, the data windows $w_N(t) = \sum_{k=-N/2}^{N/2}\delta(t-k\Delta t)$ tends, by proper definition, to a Shah function $\mathrm{III}(t;\Delta t) = \sum_{k\in\mathbb{Z}}\delta(t+k\Delta t)$, that is, to an infinite comb of Dirac delta functions in the temporal domain with periodicity equals to Δt. Otherwise, the spectral window $W_N(\nu)$, which is a real, odd and periodic function, tends also to a Shah function, but now in the frequency domain and with period-

icity Δt^{-1}. Finally, the convolution integral of Equation 2.18 acquires the form

$$X^A(\nu) = X(\nu) \otimes \text{III}\left(\nu; \Delta t^{-1}\right) = \int_{-\infty}^{+\infty} X(\nu') \sum_{k\in\mathbb{Z}} \delta\left(\nu + k\Delta t^{-1} - \nu'\right) d\nu' = \tag{2.19}$$

$$\sum_{k\in\mathbb{Z}} X\left(\nu + k\Delta t^{-1}\right) = X(\nu) + \sum_{k\in\mathbb{N}^+} \left[X\left(\nu - k\Delta t^{-1}\right) + X\left(\nu + k\Delta t^{-1}\right)\right]$$

, expression which says that the ideal process of obtaining an infinite number of samples from a signal $x(t)$ in the temporal domain is represented in the frequency domain by means of an "alias" CFT $X^A(\nu)$ which replicate the spectrum $X(\nu)$ of $x(t)$ with periodicity equal to the sampling frequency Δt^{-1} in intervals $[(2k-1)\nu_N, (2k+1)\nu_N]$ with $k \in \mathbb{Z}$ and where $\nu_{\text{N}} = \frac{1}{2\Delta t}$ is the so called Nyquist frequency, Nevertheless, a perfect replication requires a signal limited in Band for which its maximum frequency is lower than half of the sampling frequency, that is $\nu_{\max} < \nu_{\text{N}}$. If these conditions do not hold, the aliasing phenomenon arises.

To depict the problem, let us consider two even samplings, both with frequency $\Delta t^{-1} =$ 4KHz, for the signals $x_1(t) = \cos(2\pi\nu_1 t)$ and $x_2(t) = \cos(2\pi\nu_2 t)$. Let us assume, too, that the respective frequencies of the signals are $\nu_1 = 0.75$ KHz and $\nu_2 = 2.5$ KHz. For both cases the Nyquist band $[-\nu_{\text{N}}, \nu_{\text{N}}]$ is the same and $\nu_{\text{N}} = 2$ KHz. The respective scenarios in the temporal domain are depicted in Figure 2.1a.

Now, let us to shift to the frequency domain. Taking into account that the CFT for a signal $x(t) = \cos(2\pi\nu_0 t)$ is $X(\nu) = \frac{1}{2}[\delta(\nu - \nu_0) + \delta(\nu + \nu_0)]$ and particularizing Equation 2.19 we have that

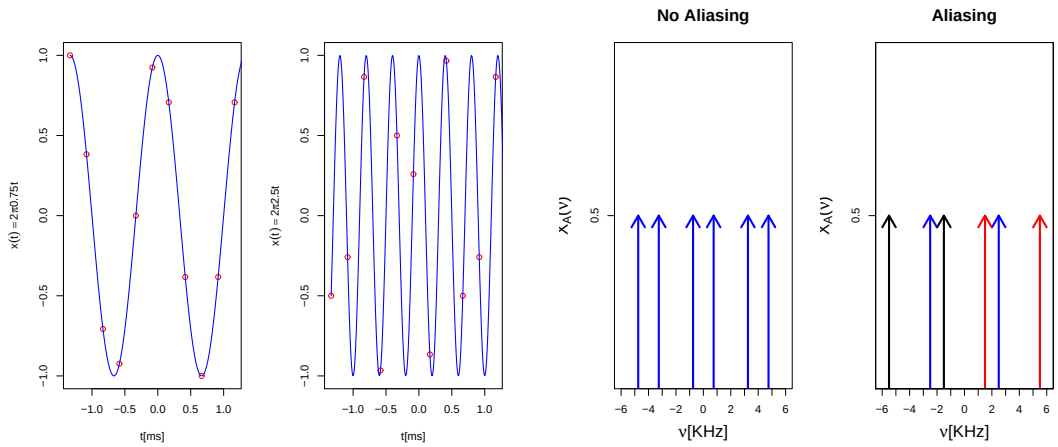
$$X^A(\nu) = \sum_{k\in\mathbb{Z}} X\left(\nu + k\Delta t^{-1}\right) = \tag{2.20}$$

$$\sum_{k\in\mathbb{Z}} \delta\left(\nu - \nu_0 + k\Delta t^{-1}\right) + \delta\left(\nu + \nu_0 + k\Delta t^{-1}\right)$$

The corresponding scenarios in this latter domain are depicted in Figure 2.1b. We conclude that:

- For $x_1(t) = \cos(2\pi 0.75t)$ (left), $X^A(\nu) = X(\nu)$ within the Nyquist band $[-\nu_{\text{N}}, \nu_{\text{N}}] = [-2, 2]$ and consequently there is no aliasing.

- For $x_2(t) = \cos(2\pi 2.5t)$ (right), $-\nu_2 = -2.5$ KHz and $+\nu_2 = 2.5$ KHz are outside of the



(a) (b)

Figure 2.1: Illustration of the aliasing phenomenon. See the text for a description.

Nyquist band but "folded" into the band. Therefore:

- $\nu_2' = 2.5 - k4 = -1.5$ KHz $(k = 1)$ (in red line) is an alias of the real frequency $\nu_2 = 2.5$ KHz.
- $-\nu_2' = -2.5 + k4 = 1.5$ KHz $(k = 1)$ (in black line) is an alias of the real frequency $-\nu_2 = -2.5$ KHz.

2.2 Gaia

2.2.1 Instrumentation and Observational Principle

We summarize the main characteristics of the Gaia payload and the way in which the satellite scan the sky as follows:

- Gaia simultaneously observes the sky, by means of two telescopes, in two viewing directions (a.k.a. fields of view, FoVs or lines of sight) separated by a basic angle of 106.5°.
- The light from both telescopes is combined into a common focal plane of 106 CCDs and a total resolution of about one billion pixels. This focal plane is constituted by three instruments: a sky mapper (SM), an astrometric field (AF), a blue (BP) and a red (RP) photometer and a radial-velocity instrument (RVS).

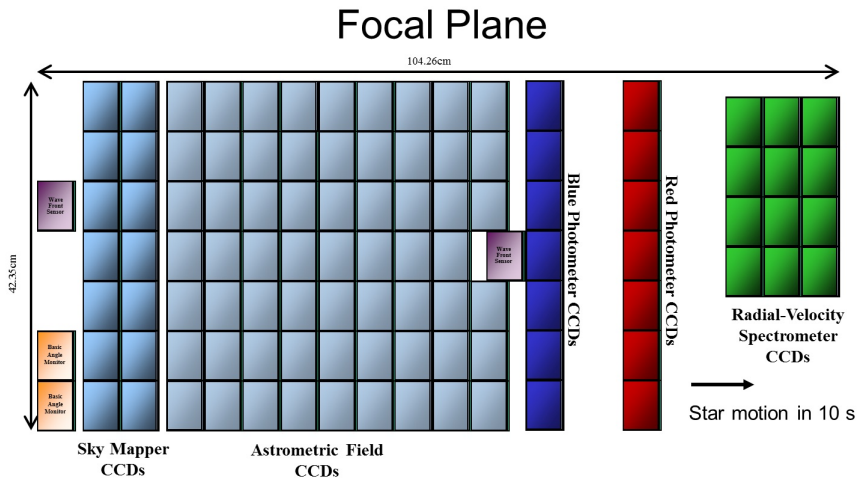


Figure 2.2: Focal Plane of Gaia. ©ESA

- The SM is composed by 14 CCDs arranged in two columns, each of them associated to one FoV. So, the SM determines from what telescope the light of a source, transiting the focal plane, is coming from. It detects objects up to 20th magnitude.

- The AF is composed by 63-1 CCDs arranged in a matrix of nine columns (strips). The first strip serves for confirmation purposes, namely to discard out false detections and to prevent for ulterior processing. If the object is accepted, eight additional measures are done in the resting 8 strips. The basic measurements are done within a rectangular 1D *data window*, with the same center as the image, which traverses the focal plane in an along-scan (horizontal) direction (AL). They compute the time when the window is entering into each CCD strip and the pixel values (counts of photo-electrons). The number of samples in the window varies from 6 to 18 depending on the magnitude of the object. The information provided by this instrument is transmitted to ground stations for further analysis about position of stars and stellar variability studies.It is convenient, however, to note that there is also a movement perpendicular to the AL displacement, namely in an across-scan direction (AC). If the pixel values are not binned in this AC direction, a 2D data window is transmitted. In this 2D windows are gathered 12 samples in the AC direction (for AF2-9).

- The mean number of transits of a detected object across the focal plane is expected to be about 70 during the 5 years mission duration.

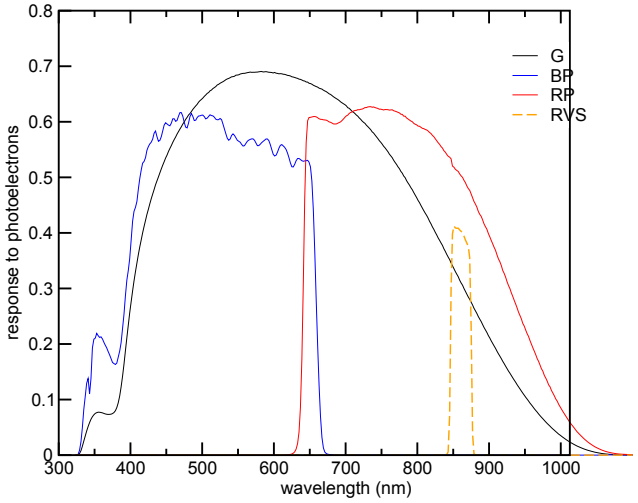


Figure 2.3: The Gaia G-Band. Source: (Jordi, 2012).

- The satellite spins around its axis with a constant rate (rotational frequency) ν_s of about $60\,\text{arcsec} \times s^{-1} \equiv 4\,\text{days}^{-1}$ or, equivalently, with a rotational period of 6 hours.
- The spin axis maintains a fixed aspect angle of 45º with the Sun and *precess* around the solar direction with a frequency ν_p of about $1/63 = 0.01587\,\text{days}^{-1}$.
- Most of times, an object transiting through one FoV is measured again in the focal plane after 106.5 or $360 - 106.5 = 253.5$ minutes, which corresponds respectively to the frequencies $\frac{24 \cdot 60}{106.5} \approx 13.5211\,\text{days}^{-1}$ and $\frac{24 \cdot 60}{253.5} \approx 5.6805\,\text{days}^{-1}$.

2.2.2 Photometric System, Error Model and Transformations

Gaia makes the astrometric observations and variability studies in the so called *G-Band*. This broad passband uses unfiltered (white) light measured in the AF of the telescope and goes from about 350 to 1000 nm (see (Jordi et al., 2010) and Figure 2.3). Unlike a conventional magnitude system, like those expressed by Eqs. 2.3 and 2.4, in the G-band system the fluxes are expressed directly as photo-electrons N_λ (per unit time per unit area per unit wavelength) integrated over the G-band in the form (De Bruijne, 2003)

$$N\left[e^- s^{-1} m^{-2}\right] = \int T(\lambda)\,\text{QE}(\lambda)\,N_\lambda(\lambda)\,d\lambda \tag{2.21}$$

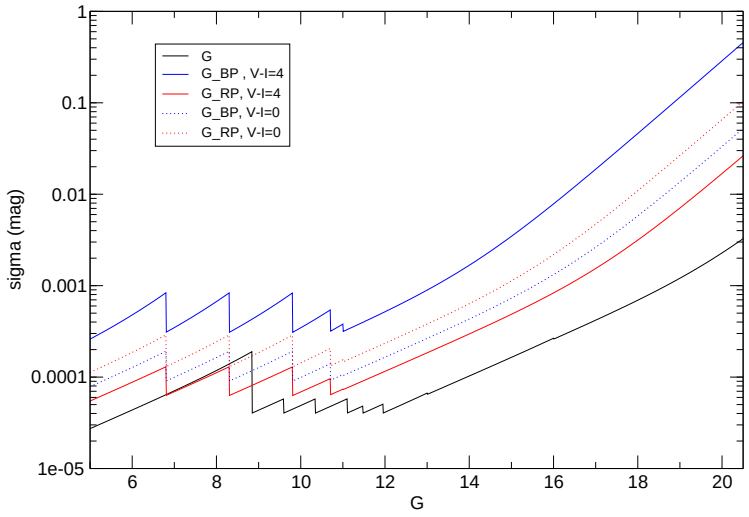


Figure 2.4: Precision for one transit in logarithmic scale. Source: (Jordi et al., 2009).

, where T and QE denote, respectively, the total transmission of the telescope optics and the CCD quantum efficiency. The G-magnitude system is formally defined (De Bruijne, 2003) (see also Jordi et al. 2006, p. 302) by

$$m_G = -2.5 \cdot log\left(\frac{N}{N_0}\right) \tag{2.22}$$

, where N_0 denotes the flux N corresponding to an unreddened A0V star (like α Lyr) with $m_V = 0$.

The quality of the sampling process performed in the AF is determined by a number of factors which cause an uncertainty in the measure of the G-magnitude. The magnitude error or standard deviation for one transit has been modeled in (Jordi et al., 2010, 2009) (see Figure 2.4) by

$$\sigma^{\mathrm{tr}}_{G(\mathrm{rec})} = m \cdot \frac{1}{\sqrt{n_{\mathrm{strips}}}} \left[\sigma^2_{\mathrm{cal}} + \left\{ 2.5 \cdot \log e \cdot \frac{\left[f_{\mathrm{aper}} \cdot f_G + (b_G + n_{\mathrm{AC}} \cdot r^2) \cdot n_s \cdot (1 + {}^{n_s}/_{n_b}) \right]^{1/2}}{f_{\mathrm{aper}} \cdot f_G} \right\}^2 \right]^{1/2} \tag{2.23}$$

, with the following meaning for its parameters[2]:

[2]Quantities denoted by fixed-width fonts can be found in the Gaia parameters database http://gaia.esac.esa.int/gpdb/index.php .

- f_G : Object flux within the G passband, computed by $f_G = N \cdot t_{\text{exp}}$, where N is derived from Equation 2.22, by

$$m_{G,\text{ZP}} - m_G = 2.5 \cdot log\left(\frac{N/N_0}{1/N_0}\right) \Rightarrow N = 10^{\frac{m_{G,\text{ZP}} - m_G}{2.5}}$$

, where $m_{G,\text{ZP}}$ = `:Satellite:Magnitude_ZeroPoint` and t_{exp} = `:Satellite:CCD_ExposureTime`

.

- f_{aper}: Light loss factor due to the 'aperture' characteristics, $f_{\text{aper}} \leq 1$, (= 0.9).
- b_G: Sky background contribution assumed to be derived from n_b background samples. Telescope_Number indica el numero de espejos (fields of view) Computed as:
  `:Satellite:AF:Sky_NumberOfPhotoElectrons*:Satellite:Telescope_Number*`
  `:Satellite:CCD_PixelAngularArea_MilliArcsecondSquare*:Satellite:CCD_ExposureTime`
- r: Total detection noise per sample. Computed as `:Satellite:AF:CCD_DetectionNoise_TypicalTotal`
- n_s: Number of samples of the object flux within the 2D window in the AL scan direction,

$$n_s = \begin{cases} 18 & m_G \leq 13 \\ 12 & 13 < m_G \leq 16 \\ 6 & m_G > 16 \end{cases}$$

- n_{AC}: Number of samples of the object flux within the 2D window in the AC scan direction, (= 12).
- n_b : Number of samples from the background (=6).
- σ_{cal} : Calibration error per observation (=30 mmag).
- n_{strips}: Averaged total number of CCD strips (columns) in the AF, $\left(\frac{6 \cdot 9 + 8}{7} \approx 8.86\right)$.
- m : Safety margin of a 20% ($m = 1.2$) that accounts for sources of error not considered explicitly in the equation.

At the end of mission the magnitude error is given by

$$\sigma_{G(\text{rec})} = \sigma^{\text{tr}}_{G(\text{rec})} \cdot \frac{1}{\sqrt{n_{\text{obs}}}} \tag{2.24}$$

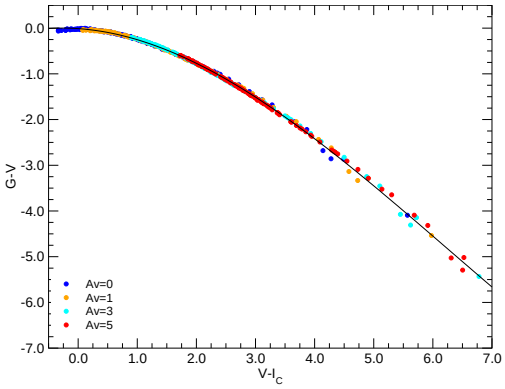


Figure 2.5: Relation between G-V and V-I. Source: (Jordi, 2012).

, where n_{obs} is the mean total number of transits (≈ 70), or equivalently, substituting the parameter n_{strips} in Equation 2.23 by the mean total number of observations $n_{\text{eff}} = n_{\text{obs}} \cdot n_{\text{strips}}$.

Finally, relevant to our work is also the colour-colour transformation between the G-magnitude, the V (visual) magnitude and the Johnson-Cousins colour index $m_V - m_{I_C}$ (Jordi et al., 2010) given by

$$m_G - m_V = -0.0257 - 0.0924 \cdot (m_V - m_{I_C}) - 0.1623 \cdot (m_V - m_{I_C})^2 + 0.0090 \cdot (m_V - m_{I_C})^3 \tag{2.25}$$

, with an error of $\sigma = 0.05$ mag.

2.3 Cepheid Variable Stars and CU7

2.3.1 Cepheids Variable Stars

A *variable star* is a star whose brightness varies across the time. A particular class is constituted by the *pulsating stars* in which these variations are the result of changes in the star radius (Aerts, 2007). The changes in the star radius are described as *pulsation modes*. In the simplest scenario, the so called *fundamental mode*, there exists a fixed (without movement) node at the center of the star and an anti-node in its surface with presents the maximum movement. In more complex scenarios the star pulsates also (or alternatively) in modes of higher order, namely in the first, second, etc. *overtones*. In this thesis we are interested in a particular type of pulsating stars, namely, the classic Cepheids. Classic Cepheids, or type I Cepheids, are giants or supergiants stars of spectral types between F5

and G5 which pulsate in the fundamental mode. In general, their pulsation period is in the range from 1 to 50 days. Their name is derived from the prototypical star δ Cephei.

Recall (see e.g. Proakis and Manolakis (1996), Section 4.1.1) that for an orthogonal basis $\{\cos(k\omega_0 t + \phi_k)\}_{k\in\mathbb{N}^+}$ the Fourier expansion of a real and periodic function $x(t)$ is given by

$$x(t) = \alpha_0 + \sum_{k\in\mathbb{N}^+} \alpha_k \cos(k\omega_0 t + \phi_k) \tag{2.26}$$

, where α_0 is the continuous component of the signal, the term into the summation are its *harmonic components*, α_k are the *amplitudes* and ϕ_k are the phase angles. It is verified that $\alpha_0 = c_0$ and $\alpha_k = 2|c_k|$ with

$$c_k = \frac{1}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}} x(t)\, e^{-jk\omega_0 t} dt. \tag{2.27}$$

Then, for a type I Cepheid we are saying that the predominant component is given by $k = 1$, that is, by the first harmonic or fundamental mode of vibration.

Period-luminosity relation This fundamental relation was inferred by (Leavitt and Pickering, 1912) from the study of 25 classic Cepheid of the SMC (Small Magellanic Cloud). Is a linear relationship of the form

$$\begin{aligned}\overline{M}_x = \alpha + \beta \cdot \log(P) = \\ \beta \cdot \log(P/10) + \alpha + 1\end{aligned} \tag{2.28}$$

, where M_x is the absolute magnitude of the star and P its period in days, being the slope $\beta < 0$. Its importance lies in that if the period of the star is known and one is able to determine its absolute brightness then, by applying Equation 2.5, one can estimate the distance to the star. Therefore, Cepheids allow to determine distances by photometry.

Important also are the period-colour and the period-luminosity-colour relationships given, respectively, by

$$\mathrm{CI} = \alpha + \beta \log(P) \tag{2.29}$$

$$\overline{M}_x = \alpha + \beta \log(P) + \gamma(\mathrm{CI}) \tag{2.30}$$

The two former relations are relevant to our work, particularized for Cepheids of the LMC (Large Magellanic Cloud), the V-Band and the colour index $m_V - m_{I_C}$ in the Johnson-Cousins UBVRI broad-band system. These latter relationships have been proposed by (Sandage et al., 2004). It is customary to note that the above equations can be expressed alternatively in terms of the logarithm of the frequency $log\,(\nu) = -log\,(P)$. E.g., Equation 2.28 can be alternatively expressed as

$$\overline{M}_x = \alpha - \beta \cdot \log\,(\nu) \tag{2.31}$$

2.3.2 Variability Processing in CU7

In this sub-section we summarize some aspects of the variability processing in CU7 related to our work. One of them are the folded light curve models (see Equation 2.6) for classic Cepheids provided by the DPAC package Lcmodels (Mowlavi et al., 2011) which use the templates given by (Bono et al., 2002) for the cepheids OGLE 56087 and OGLE 194103. The utility of such light curve models is that, given a sequence a $\{t_k\}_{k=1}^{o}$ of observation times, it is possible to generate a simulated astronomical time series for the Cepheid light curve which is done by the same package. In our case the sequence of observation times depends in turn on the way that Gaia scans the sky, the Gaia's scanning law (see Section 2.2.1). This dependence is modeled by the module Time Sampling Extraction of the Bias Estimation package ((Moitinho et al., 2011)) which, provided with the celestial coordinates of the star, generates such epoch sequences. The corresponding astronomical time series is then generated by the Time Series Generator module of the same package (which calls the Lcmodels package). It is important to note that the module Time Series generator can add noise to the simulated time series defined by $\sigma\left(\overline{m}_{G(\mathrm{rec})}\right)$ (see Equation 2.24)[3].

Once the time series has been simulated starts a post-processing stage which is done by the Characterization work group. First, a transformation to the frequency domain is done to extract the period/frequency of the time series. This task is accomplished by the Period Search package (Cuypers, 2013) by using the Deeming method (Deeming, 1975), that is, by the periodogram of Equation 2.17. Then, an inverse transformation is done to return to the time domain and adjust a polynomial to the time series. This fitting determines

[3]To be precise, the parameter that must be provided to the module is the signal-to-noise ratio *SNR* defined by $SNR = A/\sigma\left(\overline{m}_{G(\mathrm{rec})}\right)$ where A is the peak-to-peak amplitude of the light curve.

the first terms of a Fourier series as triplets of the form amplitude-phase-order $\{(\alpha_k, \phi_k, k)\}$ with indication of the residual. This latter task is performed by the Time Series Modeling module (De Ridder et al., 2009) of the characterization package (Cuypers and Guy, 2011). Note that the fitting is done to a discrete time series which is assumed to be periodic and not to the corresponding continuous light curve. That is, the fitting is done by using an expression similar to Equation 2.26 but with a finite number of terms.

2.3.3 The Aliasing Problem for Variable Stars in Gaia

We have stated the general objective of our work in Section 1.2. In particular, we aims to explicitly represent in our model the biases associated to the recovery process of the frequencies of our Cepheid population. Despite that the sampling performed by Gaia is not even, the aliasing phenomena still occurs due to the temporal recurring patterns analyzed in Section 2.2.1. The analysis of these patterns in the frequency domain is more complex that for the even sampling. Such an analysis has been accomplished for Gaia in (Mignard, 2005) demonstrating that some frequencies ν such that $\nu > \Delta t_{\min}^{-1}$, where $\Delta t_{\min}$ is the minimal lag between observations, can be successfully recovered. Otherwise, it has been showed (Eyer et al., 2009; Eyer and Mignard, 2005) that the period recovery success rate for a variable star depends on its ecliptic latitude and demonstrated that this dependence persists even for normalized signal-to-noise ratios. The classical solution for the aliased frequencies has been to discard them. For that, techniques that impose determinate constraints to accept as valid a recovered frequency have been applied (Koen and Eyer, 2002). For the case of Gaia, given that we know the regularities in its orbit, we could discard the corresponding aliased frequencies in the periodogram. Nevertheless, with this option we may be introducing biases in the real frequency distribution inferred by the model. Therefore, our proposal is to retain the complete set of recovered frequencies and model the biases.

Chapter 3

Literature Review II: Bayesian Graphical Models

In this chapter we review the mathematical framework in which is based the model proposed in this thesis and its applications to the area of astrostatistics. It is structured in five sections as follows. Section 3.1 introduces the essential concepts about graphs and probability distributions necessary to understand the Bayesian networks (BNs) knowledge representation formalism. Section 3.2 is devoted to define that representation formalism. Section 3 describes the application of BNs to the problem of statistic inference under the Bayesian paradigm, the so called Bayesian graphical models (BGMs) framework. Section 3.4 is devoted to the study of the framework for inference based in MCMC simulation techniques. Finally, in Section 3.5 we review the recent literature about application of BGMs to the area of astrostatistics.

3.1 Preliminary Concepts

In this Section we introduce some basic concepts about graphs and probability theory necessary to understand the formalism of Bayesian networks that will be described in Section 3.2. Definitions about graphs are based in those provided in (Lauritzen, 1996, Chapter 2) but we employ here an explicit notation based in indices. With regard to the concepts about probabilistic measure and particularly about conditional independence we have based our definitions in (Studenỳ, 2005).

3.1.1 Graphs

Definition 3.1. A directed graph (DG) is a pair $\mathcal{G} = (V, \mathcal{E})$ of:

- A finite set V of indices, where each index is referred to as a *vertex* (*node*).
- A subset $\mathcal{E} \subseteq V \times V$ in which each of its elements (i, j), called a *directed edge* (*arc*) and denoted $i \to j$, verify that $i \neq j \wedge (j, i) \notin \mathcal{E}$.

Definition 3.2. Given a DG $\mathcal{G} = (V, \mathcal{E})$, a *path* from vertex i to vertex j, denoted $s\,(i, j)$, is a sequence $\{i = i_0, ..., i_n = j\}$ in V such that:

- For every $k \in \{1, .., n\} : (i_{k-1}, i_k) \in \mathcal{E} \vee (i_k, i_{k-1}) \in \mathcal{E}$
- Nodes in the sub-sequence $\{i_1, ..., i_{n-1}\}$ are all distinct.

If the first condition of the definition is restricted to

- For every $k \in \{1, .., n\} : (i_{k-1}, i_k) \in \mathcal{E}$

, the path $s\,(i, j)$ is said to be a *directed path*. If additionally it holds that $i = j$ then the directed path $s\,(i, j)$ is said to be a *cycle*.

Definition 3.3. A DG $\mathcal{G} = (V, \mathcal{E})$ is a *directed acyclic graph* (DAG) if it has no cycles.

Definition 3.4. Given a DAG $\mathcal{G} = (V, \mathcal{E})$ and a vertex $j \in V$:

- The set of *parents* of j, denoted by $\mathrm{pa}\,(j)$, is the set of nodes i for which there exists an arc $i \to j$.
- The set of *ancestors* of j, denoted by $\mathrm{an}\,(j)$, is the set of nodes i such that there exists a directed path $s\,(i, j)$.
- The set of *descendants* of j, denoted by $\mathrm{de}\,(j)$, is the set of nodes i such that there exists a directed path $s\,(j, i)$.
- The set of *non descendants* of j, denoted by $\mathrm{nd}\,(j)$, is the set $V \setminus \mathrm{de}\,(j)$.

3.1.2 Probability Distributions

Definition 3.5. (Random variable). Let $(\Omega, \mathcal{F}, P)$ be a probability space, X be a real subset and $\mathcal{X}$ be a σ-algebra on X. Then, a function $X : \Omega \to \mathsf{X}$ is a (real valued) random variable if it satisfies

$$X^{-1}(\mathsf{B}) = \{w : X(w) \in \mathsf{B}\} \in \mathcal{F} \qquad , \forall \mathsf{B} \in \mathcal{X} \tag{3.1}$$

An example is the Borel σ-algebra $\mathcal{B}$ on $\mathbb{R}$. Note also that $X^{-1}(\mathcal{X})$ is a σ-algebra, called the σ-algebra *generated by* the r.v. X and denoted $\sigma(X)$.

Definition 3.6. (Random vector). Let $(\Omega, \mathcal{F}, P)$ be a probability space and V a finite set of indices such that $|V| = n$. A n-dimensional random vector is a collection $X_V = (X_i)_{i \in V}$ of random variables $X_i : \Omega \to \mathsf{X}_i$.

In practical applications the measurable space $(\Omega, \mathcal{F}, P)$ is often unspecified and the interest lies in computing probabilities directly over events into X or into the product space $\mathsf{X}_V = \times_{i \in V} \mathsf{X}_i$. The latter can be done by taking into account that $(\Omega, \mathcal{F}, P)$ induces via X and X_V, probability spaces on X and X_V, namely $(\mathsf{X}, \mathcal{X}, P_X)$ and $(\mathsf{X}_V, \mathcal{X}_V, P_{X_V})$, with $\mathcal{X}_V = \times_{i \in V} \mathcal{X}_i$, where the probability measures $P_X : \mathcal{X} \to [0,1]$ and $P_{X_V} : \mathcal{X}_V \to [0,1]$ are defined, respectively, by

$$P_X(\mathsf{B}) = P\left(X^{-1}(\mathsf{B})\right), \ \forall \mathsf{B} \in \mathcal{X} \tag{3.2}$$

$$P_{X_V}(\mathsf{B}) = P\left(X_V^{-1}(\mathsf{B})\right), \ \forall \mathsf{B} \in \mathcal{X}_V \tag{3.3}$$

Definition 3.7. (Marginal probability measure). Given a product space $(\mathsf{X}_V, \mathcal{X}_V, P_{X_V})$ and a subset of indices $A \subseteq V$, the *marginal probability measur*e associated to $\mathsf{X}_A = \times_{i \in A} \mathsf{X}_i$ is defined from P_{X_V} by

$$P_{X_A}(\mathsf{A}) = P_{X_V}(\mathsf{A} \times \mathsf{X}_{V-A}) \ , \forall \mathsf{A} \in \mathcal{X}_A \tag{3.4}$$

Definition 3.8. (Conditional probability measure). Let $A, C \subseteq V$ be disjoint sets of indices. The *conditional probability measure* on X_A given the σ-algebra $\mathcal{X}_C$ is a function $P_{X_A|X_C} : \mathcal{X}_A \times \mathsf{X}_C \to [0,1]$ such that, for every $\mathsf{A} \in \mathcal{X}_\mathsf{A}$:

1. $P_{X_A|X_C}(\mathsf{A} \mid \cdot) : \mathsf{X}_C \to [0,1]$ is a $\mathcal{X}_C$-measurable function.

2. $P_{X_{AC}}(\mathsf{A} \times \mathsf{C}) = \int_{\mathsf{C}} P_{X_A|X_C}(\mathsf{A} \mid x)\, dP_{X_C}(x)$, $\forall \mathsf{C} \in \mathcal{X}_C$.

In practical applications Definitions 3.7 and 3.8 are usually given in terms of pointwise functions, namely, Lebesgue integrable probability density functions (PDF's) or discrete probability mass functions (PMF's).

The important definition of conditional independence which follows was introduced by (Dawid, 1979) with an affordable treatment which also can be found in (Lauritzen, 1996, Chapter 2). A more formal approach is proposed and followed by (Dawid, 1980).

Definition 3.9. (Conditional independence). Let $A, B, C \subseteq V$ be pairwise disjoint subsets of indices. Random vectors X_A and X_B are *conditionally independents* given the r.v. X_C w.r.t. P, denoted by $I_P(X_A, X_B \mid X_C)$, if for every $\mathsf{A} \in \mathcal{X}_A$ and $\mathsf{B} \in \mathcal{X}_B$

$$P_{X_{AB}|X_C}(\mathsf{A} \times \mathsf{B} \mid x) = P_{X_A|X_C}(\mathsf{A} \mid x) \cdot P_{X_B|X_C}(\mathsf{B} \mid x) \ , P_{X_C} - a.e.\ x \in \mathsf{X}_C \tag{3.5}$$

In the particular case that $C = \varnothing$, Equation 3.5 defines the classic concept of independence between X_A and X_B

$$P_{X_{AB}}(\mathsf{A} \times \mathsf{B}) = P_{X_A}(\mathsf{A}) \cdot P_{X_B}(\mathsf{B}) \tag{3.6}$$

3.2 D-separation, Conditional Independence, and Bayes Nets

A probabilistic graphical model for DAG's, the so called Bayesian network (Pearl, 1985), is composed by a joint probability distribution defined over a set of variables whose independence relationships are dictated by the topology of a directed graph. To understand this representation formalism it is necessary to define clearly what topological properties of a DAG correspond to the different forms of stochastic independence. In this Section we review some definitions, introduced by (Pearl, 1988; Lauritzen, 1996), to establish such a correspondence and define the representation formalism itself.

Definition 3.10. (Active path) Given a DAG $\mathcal{G} = (V, \mathcal{E})$, a subset $C \subset V$ and two nodes i and j of $\mathcal{G}$ such that $\{i, j\} \subset V - C$, a path $s(i, j)$ is *active given* C, denoted by $\neg I^s_{\mathcal{G}}(i, j \mid C)$, when any of the following conditions holds:

- $s(i, j)$ is an arc: $i \to j$ or $j \to i$.

- $s\,(i,j)$ is of the type $i \to k \to j$ (*head-to-tail*) or $i \leftarrow k \to j$ (*tail-to-tail*), and $k \notin C$.
- $s\,(i,j)$ is of the type $i \to k \leftarrow j$ (*head-to-head*) and $k \in C$ or $\exists l \in \mathrm{de}\,(k) : l \in C$.
- $s\,(i,j)$ has more than three nodes and every subpath of $s\,(i,j)$ is active given C.

Otherwise the path $S\,(i,j)$ is *inactive* given C (or *blocked* by C), denoted by $I^s_{\mathcal{G}}\,(i,j \mid C)$.

Note that if $C = \varnothing$ head-to-tail or tail-to-tail paths are always active, i.e. $\neg I^s_G\,(i,j)$, and head-to-head paths are always inactive, $I^s_{\mathcal{G}}\,(i,j)$. So, to change the default status of a path of three nodes we should include the intermediate node k in a set C located to the right side of the bar, or in the case of a head-to-head path we could also change its status including in C any of the descendants of k.

Definition 3.11. (d-separation between nodes) Given a DAG $\mathcal{G} = (V, \mathcal{E})$ and a subset $C \subset V$, two nodes i and j of $\mathcal{G}$ are *d-separated by* C, denoted by $I_G\,(i,j \mid C)$, if all paths between i and j are inactive given C. Otherwise it is said that the nodes are *connected* given C, denoted by $\neg I_G\,(i,j \mid C)$.

Definition 3.12. (d-separation between sets of nodes) Given a DAG $\mathcal{G} = (V, \mathcal{E})$ and a tern of pairwise disjoint node subsets $A, B, C \subseteq V$, subsets A and B are *d-separated* (or *blocked*) by C, denoted by $I_G\,(A,B \mid C)$, if for every pair $(i,j) \in A \times B$ it holds that $I_G\,(i,j \mid C)$. Otherwise, A and B are said to be *connected* given C, denoted by $\neg I_G\,(A,B \mid C)$.

Note that the above definition do not include the concept of probabilistic independence at all. Nevertheless, it can be interpreted in a probabilistic sense by introducing the notion of independence map for DAGs (Pearl, 1988).

Definition 3.13. (I–map) Given a finite set of indices V, a DAG $\mathcal{G} = (V, \mathcal{E})$ and a r.v. $X_V = (X_i)_{i \in V}$ taking values in the probability space $(\mathsf{X}_V, \mathscr{X}_V, P_{X_V})$, $\mathcal{G}$ is an *I-map* of the probability distribution $P = P_{X_V}$ if for every tern of pairwise disjoints subsets $A, B, C \subseteq V$ it holds that

$$I_{\mathcal{G}}\,(A,B \mid C) \Rightarrow I_P\,(X_A, X_B \mid X_C) \tag{3.7}$$

To illustrate the sense of the latter definition let us to consider, for simplicity, an arbitrary DAG $\mathcal{G}$ of three nodes (variables) which is an I-map for a certain probability distribution P. Then, for the three paths (substructures) of three nodes that may exist in G, abstracting of node names we have:

- A head-to-tail path $i \to k \to j$ and an tail-to-tail path $i \leftarrow k \to j$ both verify, by Def. 3.10, that $I_G(i, j \mid k)$. Then, in both cases, by Def. 3.13 it holds that $P(X_i, X_j \mid X_k) = P(X_i \mid X_k) \cdot P(X_j \mid X_k)$, i.e. X_i and X_j are conditionally independent given X_k,

- A head-to-head path $i \to k \leftarrow j$ verify, by Def. 3.10, that $I_G(i, j)$, which is interpreted by applying Def. 3.13, as that X_i and X_j are marginally independent, i.e. $P(X_i, X_j) = P(X_i) \cdot P(X_j)$.

Note that the precedent definition states a sufficient condition for conditional (or unconditional) independence but do not says nothing about what happens if the nodes are connected. For an arc in the graph reflects a direct probabilistic dependency is required to state a necessary condition for independence. This necessary condition is formalized by means of the definition of independence map.

Definition 3.14. (Perfect I–map) Given a finite set of indices V, a DAG $\mathcal{G} = (V, \mathcal{E})$ and a r.v. $X_V = (X_i)_{i \in V}$ taking values in the probability space $(\mathsf{X}_V, \mathcal{X}_V, P_{X_V})$, $\mathcal{G}$ is a *perfect I-map* of $P = P_{X_V}$ or equivalently the pair $(\mathcal{G}, P)$ satisfy the *fidelity condition*, if for every tern of pairwise disjoints subsets $A, B, C \subseteq V$ it holds that

$$I_{\mathcal{G}}(A, B \mid C) \Leftrightarrow I_P(X_A, X_B \mid X_C) \tag{3.8}$$

Definition 3.15. (Local Markov property) Given a DAG $\mathcal{G} = (V, \mathcal{E})$ and a probability space $(\mathsf{X}_V, \mathcal{X}_V, P)$, P satisfies the *Local Markov Property* w.r.t. $\mathcal{G}$ if for every node $i \in V$ it holds that

$$I_P\left(X_i, X_{\mathrm{nd}(i)} \mid X_{\mathrm{pa}(i)}\right) \tag{3.9}$$

, that is, X_i is conditionally independent of the set of its non descendent given its parents.

Definition 3.16. (Recursive factorization property) Given a finite set of indices V, a DAG $\mathcal{G} = (V, \mathcal{E})$, a r.v. $X_V = (X_i)_{i \in V}$ taking values in the probability space $(\mathsf{X}_V, \mathcal{X}_V, P_{X_V})$ and the tern $(X_V, \mathcal{G}, P_{X_V})$, the probability measure P_{X_V} satisfies the *recursive factorization property* w.r.t. $\mathcal{G}$ if there exist σ-finite measures μ_i on X_i and functions $k_i : \mathsf{X}_i \times \mathsf{X}_{\mathrm{pa}(i)} \to \mathbb{R}^+$ such that

1. $\int_{\mathsf{X}_i} k_i\left(y_i, x_{\mathrm{pa}(i)}\right) d\mu_i(y_i) = 1$, for every $i \in V$ and $x_{\mathrm{pa}(i)} \in \mathsf{X}_{\mathrm{pa}(i)}$.

2. P_{X_V} has a density w.r.t. the product measure $\times_{i \in V} \mu_i$ given by $p(x) = \prod_{i \in V} k_i \left(x_i, x_{\mathrm{pa}(i)} \right)$.

Note that this definition connect with Def. 3.8 in the sense that k_i is the density for a kernel $K_i : \mathcal{X}_i \times \mathsf{X}_{\mathrm{pa}(i)} \to [0,1]$ which corresponds, almost sure, to the conditional probability measure on X_i given the σ-algebra $\mathcal{X}_{\mathrm{pa}(i)}$.

Definition 3.17. (Bayesian network). A tern $(X_V, \mathcal{G}, P_{X_V})$ is a *Bayesian network* if it satisfies any of the following properties, which are equivalent:

- $\mathcal{G}$ is an I-map of the probability distribution P_{X_V}.
- P_{X_V} satisfies the recursive factorization property w.r.t. $\mathcal{G}$.
- P_{X_V} satisfies the Local Markov Property w.r.t. $\mathcal{G}$.

The demonstration of the equivalence stated in the latter definition can be found in (Lauritzen, 1996) (see Proposition 3.25 and Theorem 3.27).

As we can see by the latter definition, a Bayesian network need not necessarily satisfy the fidelity condition stated by Def. 3.14. Keeping in mind the manual construction of a model based in such a representation formalism this fact suggests us to be cautious with the introduction of arcs between nodes, given that they should reflect only a direct probabilistic dependence between them.

To end the present Section we review an independence property of BNs that will be necessary to understand the MCMC sampling scheme that we will study in Section 3.4.5.

Definition 3.18. (**Markov blanket**). Given a finite set of indices V, a r.v. $X_V = (X_i)_{i \in V}$ taking values in the probability space $(\mathsf{X}_V, \mathcal{X}_V, P)$ and a r.v. X_j, a *Markov blanket* for X_j is any random subvector $X_{\mathrm{bl}(j)}$ such that

$$I_P \left(X_j, X_{V - (\mathrm{bl}(j) \cup \{j\})} \mid X_{\mathrm{bl}(j)} \right) \tag{3.10}$$

, i.e., such that X_j is conditionally independent of all the other variables in X_V given $X_{\mathrm{bl}(j)}$.

Note that Equation 3.10 can be expressed, with a slight abuse of the notation, in terms of PDFs f for P by

$$f \left(x_j \mid x_{V - \{j\}} \right) = f \left(x_j \mid x_{\mathrm{bl}(j)} \right) \tag{3.11}$$

Proposition 3.1. *Let $(X_V, \mathcal{G}, P_{X_V})$ be a Bayesian network and X_j be a component of X_V. Then, the random subvector $X_{\mathrm{bl}(j)}$whose subset of indices is given by*

$$\mathrm{bl}\,(j) = \mathrm{pa}\,(j) \cup \mathrm{ch}\,(j) \cup_{k \in \mathrm{ch}(j)} \{\mathrm{pa}\,(k) \setminus \{j\}\} \tag{3.12}$$

is a Markov blanket for X_j .

Corollary 3.1. *Let $(X_V, \mathcal{G}, P)$ be a Bayesian network, X_j be a component of X_V and f be a PDF for P. Then for every $x \in X_V$ it holds*

$$f\left(x_j \mid x_{\mathrm{bl}(j)}\right) \propto f\left(x_j \mid x_{\mathrm{pa}(j)}\right) \times \prod_{k \in \mathrm{ch}(j)} f\left(x_k \mid x_{\mathrm{pa}(k)}\right) \tag{3.13}$$

, i.e., the PDF of a random variable given its Markov blanket is proportional to the PDF of the variable given its parents in $\mathcal{G}$ times the product of the PDFs of each child given its respective parents.

3.3 Bayesian Graphical Models

3.3.1 Inference in a Classical Multinomial BN

To introduce the Bayesian graphical models framework let us start with the formulation of a classic multinomial Bayesian network as a tern $(\mathbf{X}, \mathcal{G}, p)$ where $p\,(\mathbf{X})$ is the discrete PDF for a categorical distribution which satisfies the recursive factorization property w.r.t. the DAG $\mathcal{G}$, namely

$$p\,(\mathbf{x}) = \prod_i p\left(x_i \mid x_{\mathrm{pa}(i)}; \boldsymbol{\theta}_i\right) \tag{3.14}$$

, where $p\,(x_i \mid \pi_i, \boldsymbol{\theta}_i)$ denotes the conditional PDF of the variable (node) X_i in $\mathcal{G}$ given a determinate configuration $x_{\mathrm{pa}(i)}$ of its parents. Note that we make explicit in this formula the dependence of the family of PDFs on the parameter vector $\boldsymbol{\theta}_i$ for all the possible parent configurations. So, the distribution in the i-th node given the j-th configuration of its parents is 1-Multinomial, $X_i \mid x^j_{pa(i)} \sim \mathcal{M}\,(1; \boldsymbol{\theta}_{ij})$, with parameters $\boldsymbol{\theta}_{ij} = (\theta_{ijk})_{k=1}^{r_i}$ where

$$\theta_{ijk} = p\left(X_i = x_i^k \mid x^j_{pa(i)}\right) \tag{3.15}$$

We can see that such a network has two elements: its structure, dictated by the DAG,

and its parameters (probabilities). These two elements correspond to the two types of knowledge (qualitative and quantitative) that the network encapsulates. They also depict the two main stages followed in its construction. In the case of a network already built and ready to make inferences, parameters are considered fixed and given by contingency tables. The problem of probabilistic inference in the network is then posed as the problem of computing the conditional posterior distribution of a set X_I of variables of interest, *given a particular case* x_E of a set X_E of observed variables (the evidence). By application of the definition of conditional probability this is formulated as

$$p\left(x_I \mid x_E\right) = \frac{p\left(x_I, x_E\right)}{p\left(x_E\right)} \tag{3.16}$$

It is important to note that, although in the setting of the inference problem stated by Equation 3.16 the definition of conditional probability is used, the term Bayesian network does not necessarily imply a statistical Bayesian approach to solve the problem Korb and Nicholson (2003). Really, if the two phases of construction of the net, qualitative (structure) and quantitative (parameters), are done completely with the aid of an expert the inference problem cannot even be considered a statistical inference problem.

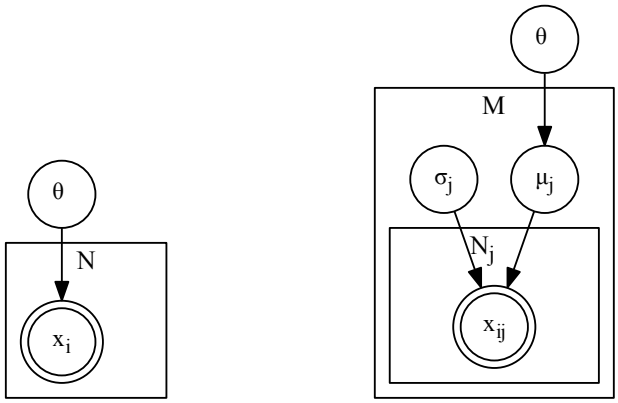
3.3.2 BGMs as a Representation Language for Statistical Inference

When the formalism of Bayesian networks is used as a representation language to perform statistical inference within the Bayesian paradigm (Gelman et al., 2004; Gelman and Shalizi, 2013), parameters of conditional distributions on each node are treated as random variables (on equal footing with the rest of random variables in the network) and the evidence is given by a set of independent samples[1] from the observed variables. In the simplest case we have a single parameter θ which parameterizes the distribution of a single observed variable X and have a sample $\mathcal{D} = \{x_i\}_{i=1}^{N}$ of X. So, the joint PDF of the network takes the form

$$p\left(\theta, \mathcal{D}\right) = \pi\left(\theta\right) \prod_{i=1}^{N} p\left(x_i \mid \theta\right) \tag{3.17}$$

, where the product in the second term of the equation reflects the fact that samples x_i are conditionally independent given θ. Although simple, the model expressed by Equation 3.17

[1] But not necessarily identically distributed.



(a) (b)

Figure 3.1: Graphs associated to the BGMs given by Equations 3.17 (a) and 3.18 (b). Rectangles indicate repeating patterns, by using the *plate* notation, and nodes enclosed in double circle denote observations.

contains the two main types of knowledge that we can find encapsulated into a BGM. This knowledge is a *prior distribution* $\pi(\theta)$ for the parameter and a component of *likelihood*. The prior reflects our previous knowledge about the parameter before the sample is observed and the likelihood give us the probability of observing the sample $\mathcal{D}$ given each value of θ. Therefore, unlike a manually constructed classical BN, here the encapsulated quantitative knowledge only includes the priors.

In a further level of complexity we can introduce the notion of *hierarchy*. This hierarchy, when applied to parameters, makes probabilities of the likelihood parameters depend on other parameters in turn, namely *hyperparameters* with their corresponding *hyperprior* distributions. When a hierarchy is used to model the generation of the observations, data are grouped and the likelihood is subfactorized by subsets of parameters. An example of BGM having these two types of hierarchy is

$$p(\theta, \boldsymbol{\sigma}, \boldsymbol{\mu}, \mathcal{D}) = \pi(\theta) \prod_{j=1}^{M} p(\mu_j \mid \theta) \prod_{i=1}^{N_j} p(x_{ij} \mid \mu_j, \sigma_j) \tag{3.18}$$

, where $\sum_{j=1}^{M} N_j = N$ and $\mathcal{D} = \{\mathcal{D}_j\}_{j=1}^{N}$ with $\mathcal{D}_j = \{x_{ij}\}_{i=1}^{N_j}$. For this second example the

likelihood component is given by

$$L(\boldsymbol{\sigma}, \boldsymbol{\mu}) = \prod_{j=1}^{M} \prod_{i=1}^{N_j} p(x_{ij} \mid \mu_j, \sigma_j) \tag{3.19}$$

To understand in a simple way the independence properties of a BGM, it is customary to draw the associated graph by using a *plate notation* to denote its repetition patterns. In Figure 3.1 we depict the graphs corresponding to the BGMs of Equations 3.17 and 3.18. We can see that for the graph in Fig. 3.1b data $\mathcal{D}_j$ inside the inner plate N_j, that is fixed j, are all conditionally independent given (μ_j, σ_j) and parameters (μ_j, σ_j) are all conditionally independent given the hyperparameter θ.

A BGM in which the notion of hierarchy is applied both to data and parameters, in the sense depicted in the latter example, are also called a *Hierarchical Bayesian Model* (see Gelman et al. 2004, Chapter 5). We prefer however the denomination of Bayesian graphical model employed by (Højsgaard et al., 2012) which seems to us more general and accurate. We can summarize the characteristics of the BGM representation language for statistic inference as follows:

- The random nodes (variables) of the graph are classified in two categories: *random parameters* and *observed variables*.
- A *random parameter* is a random variable in the graph whose values are unknown. Otherwise a *fixed parameter* is a parameter whose values are known a *priori*.[2]
- The distribution of a random parameter is called the *prior distribution* for the parameter. The prior can be a conditional distribution, if the node has parents, or an unconditional distribution if the parameter is an orphan node. This prior distribution encodes the stochastic knowledge that we have about the parameter *before* the data have been observed.
- For the set of observed variables exist a sample of observations which is independent but not always identically distributed. This sample constitutes the evidence and is

[2] – With the intention of simplifying the interpretation of the models proposed in this thesis, the fixed parameters normally will not be elicited in the corresponding graphs. Remark that this can vary depending on the source.

called *the data*.

- The conditional distribution of the data given their parent parameters is called the *likelihood of the model*. As the sample is a set of fixed values, the likelihood is a function of the data's parent parameters.

- Parameters, prior distributions and data can be organized into a hierarchy dictated by the structure of the graph. If the parents of a random parameter are also random parameters they are called *hyperparameters*. The corresponding distributions for the *hyperparameters* are called *hyperpriors*. The hierarchy reaches its top level when the conditional distribution of a random parameter depends only on fixed parameters.

As a summary, the representation framework adopted in this work can be represented by

$$\text{BGM=BN+Bayesian Statistical Framework} \tag{3.20}$$

3.3.3 The Inference Problem in a BGM

The inference problem in a BGM is the problem of updating our prior knowledge about some parameter (or parameters) of interest in the light of the evidence provided by the data (our statistical sample). This problem reduces to computing the posterior distribution of the parameter given the data, which presents two major difficulties. For example, for the model given by Eq. 3.17 by application of the Bayes Theorem, and for the only parameter we could be interested in, we have

$$p\left(\theta \mid \mathcal{D}\right) \propto \pi\left(\theta\right) p\left(\mathcal{D} \mid \theta\right) = \pi\left(\theta\right) \prod_{i=1}^{N} p\left(x_i \mid \theta\right) \tag{3.21}$$

, that is, the posterior for θ is known up to the normalizing constant $p\left(\mathcal{D}\right) = \int_{\theta} \pi\left(\theta\right) p\left(\mathcal{D} \mid \theta\right) d\theta$ that (i) must somehow determined. Now, for the two levels model given by Eq. 3.18 we can obtain the *joint posterior distribution* for all parameters by

$$p\left(\theta, \boldsymbol{\sigma}, \boldsymbol{\mu} \mid \mathcal{D}\right) \propto \pi\left(\theta\right) \prod_{j=1}^{M} p\left(\mu_j \mid \theta\right) \prod_{i=1}^{N_j} p\left(x_{ij} \mid \mu_j, \sigma_j\right) \tag{3.22}$$

But here we could be interested only in the *marginal a posteriori* for a determinate parameter, namely in $p\left(\theta \mid \mathcal{D}\right)$. Therefore this latter case presents the additional difficulty of (ii)

marginalizing over the remaining (*nuisance*) parameters:

$$p\left(\theta \mid \mathcal{D}\right) = \int_{(\boldsymbol{\sigma},\boldsymbol{\mu})} p\left(\theta, \boldsymbol{\sigma}, \boldsymbol{\mu} \mid \mathcal{D}\right) d\left(\boldsymbol{\sigma}, \boldsymbol{\mu}\right) \tag{3.23}$$

To summarize, in a general setting we will have a partition $(\Theta, \mathcal{D}) = (\Phi, \Lambda, \mathcal{D})$ of the variables in the network where Φ and Λ are, respectively, the set of parameters of interest and the nuisance parameters, being the objective to determine $p\left(\Phi \mid \mathcal{D}\right)$ somehow marginalizing over Λ, that is,

$$p\left(\Phi \mid \mathcal{D}\right) = \int_{\Lambda} p\left(\Phi, \Lambda \mid \mathcal{D}\right) d\Lambda \propto \int_{\Lambda} p\left(\Phi, \Lambda, \mathcal{D}\right) d\Lambda \tag{3.24}$$

In Section 3.4 we will see how that this task can be accomplished by application of MCMC simulation techniques. The BGM framework is thus completed by adding to Equation 3.20 the term

$$+\text{MCMC Inference} \tag{3.25}$$

3.4 Inference by MCMC Methods

Recall from Section 3.3.3 that we can partition variables in the network as $(\Theta, \mathcal{D}) = (\Phi, \Lambda, \mathcal{D})$. Recall also that the inference objective is to obtain the posterior distribution $p\left(\Phi \mid \mathcal{D}\right)$ for parameters of interest taking into account the presence of the set Λ of nuisance parameters in whose posterior distribution we are not interested. An easy-to-implement strategy to do this is to use a grid search where the posterior distribution in Equation 3.24 is estimated by a discrete approximation. The method has the following characteristics: i) the range of values for each parameter in Θ is divided into a number of discrete levels, ii) the joint density distribution to the right in Equation 3.24 is evaluated for every discrete combination of values of Θ, ii) the posterior density $p\left(\Phi, \Lambda \mid \mathcal{D}\right)$ for every combination is estimated by applying the discrete Bayes Theorem; and iii) the marginal distribution $p\left(\Phi \mid \mathcal{D}\right)$ is computed by marginalization. Nevertheless, this brute-force method is inefficient on a high-dimensional parameter space. In these latter cases, a better alternative is to use Markov chain Monte Carlo (MCMC) simulation techniques (Gentle et al., 2012; Roberts and Rosenthal, 2004; Robert and Casella, 2004; Neal, 1993).

3.4.1 Markov Chains

We dedicate this subsection to expose some basic concepts and results necessary to understand the fundamentals of the sampling algorithms described in the next subsections. The material presented here can be found in (Robert and Casella, 2004). We start with the definition of a transition kernel, concept which is the basis to define the temporal evolution of a Markov chain.

Definition 3.19. (**Transition kernel**). Given a topological space $\mathsf{X} \subseteq \mathbb{R}^d$ and the Borel σ-algebra $\mathcal{B}(\mathsf{X})$ induced on X, a *transition kernel* is a function K defined on $\mathsf{X} \times \mathcal{B}$ such that

1. $\forall x \in \mathsf{X}$, $K(x, \cdot)$ is a probabilistic measure.
2. $\forall A \in \mathcal{B}$, $K(\cdot, A)$ is a measurable function.

The first condition of the preceding definition defines the conditional probabilities of jumping in one step from a *source*, punctual and given *state* x to all the possible *destination states* or sets $A \in \mathcal{B}(\mathsf{X})$. The second condition guaranties that the probability of these jumps can be evaluated for any source point x of the space X. In the particular case that X is finite, with cardinality m the kernel is an m-by-m transition matrix with one row per source (2nd condition) and where the sum of the elements of each row, the transition probabilities, equals to 1 (1st condition).

Usually the kernel is defined by means of a density $k(x, y)$(w.r.t. the standard Lebesgue measure in $\mathbb{R}^d$) by

$$K(x, A) = \int_A k(x, y)\, dy \tag{3.26}$$

Now, a Markov chain is defined as a temporal sequence of random variables dictated by a transition kernel which additionally verifies the Markov. This property says "the past and the future states of the sequence are conditionally independent given its present state".

Definition 3.20. (**Markov chain**). Given a topological space$(\mathsf{X}, \mathcal{B}(\mathsf{X}))$ and a transition kernel K defined on $\mathsf{X} \times \mathcal{B}$, a *Markov chain* is a temporal sequence $\left\{X^{(t)}\right\}_{t \in \mathbb{N}}$ of random variables taking values in X such that

$$P\left(X^{(t+1)} \in A \mid x^{(t)}\right) = \int_A K\left(x^{(t)}, dx\right) \tag{3.27}$$

$$P\left(X^{(t+1)} \in A \mid x^{(t)}\right) = P\left(X^{(t+1)} \in A \mid \left\{x^{(k)}\right\}_{k=0}^{t}\right) \tag{3.28}$$

Really, to complete the preceding definition it is necessary to specify the "initial state" of the chain. It can be stated by means of a marginal distribution μ for X_0, or by means of the Dirac mass δ_{x_0} if the chain is started deterministically to a value $X_0 = x_0$.

Definition 3.21. (**Invariant probability measure**). A probability measure π is *invariant* (or *stationary*) w.r.t. a Markov chain $\left\{X^{(t)}\right\}_{t\in\mathbb{N}}$ with transition kernel $K(\cdot,\cdot)$ if

$$\pi(A) = \int_{\mathsf{X}} K(x, A)\, \pi(dx) \qquad ,\forall A \in B(\mathsf{X}) \tag{3.29}$$

, which implies, $\forall t \in \mathbb{N}$, that if $X^{(t)} \sim \pi$ then $X^{(t+1)} \sim \pi$.

Theorem 3.1. *(Roberts and Rosenthal, 2004) If a Markov chain on a state space with countable generated σ-algebra is ϕ-irreducible and aperiodic, and has a stationary distribution $\pi(\cdot)$, then for π-a.e. $x \in \mathsf{X}$,*

$$\lim_{t\to\infty} \left\| K^t(x, \cdot) - \pi(\cdot) \right\| = 0 \tag{3.30}$$

3.4.2 General Scheme of Inference

The general idea behind the MCMC methods to sample the posterior distribution for parameters of interest Φ in $\Theta = (\Phi, \Lambda)$ taking into account the presence of the set Λ of nuisance parameters can be summarized in the following steps:

1. Construct a Markov chain $\left\{\mathbf{X}^{(t)}\right\}_{t\in\mathbb{N}}$ such that:
   (a) The distribution of $\Theta \mid \mathcal{D}$ be invariant w.r.t. the chain $\left\{\mathbf{X}^{(t)}\right\}_{t\in\mathbb{N}}$ and
   (b) The chain be ergodic, i.e. it converges to the distribution of $\Theta \mid \mathcal{D}$ independently of its initial state.
2. State some criteria of convergence for the chain.
3. If the convergence has been reached in the step n_0, discard the first $n_0 - 1$ elements of the chain (*burn-in* phase), make $\Theta^{(n)} := \mathbf{X}^{(n+n_0)}$ and take the new chain $\left\{\Theta^{(n)}\right\}$ as a sample for $\Theta \mid \mathcal{D}$.
4. Estimate $\Phi \mid \mathcal{D}$ taking the subsample $\left\{\Phi^{(n)}\right\}$ into $\left\{\Theta^{(n)}\right\}$.

3.4.3 The Metropolis-Hastings Algorithm

Let be Θ a random vector with support a set $\mathsf{X} \subseteq \mathbb{R}^d$ whose PDF $\pi(\boldsymbol{\theta}) = \frac{f(\boldsymbol{\theta})}{k}$, so called the *target distribution*, is known up to the normalizing constant k. The objective of the Metropolis-Hasting (MH) algorithm (cited by Chib and Greenberg, 1995) is to generate samples of $\pi(\boldsymbol{\theta})$. For that, it generates a Markov chain from a conditional density $q(\mathbf{y} \mid \boldsymbol{\theta})$, so called the *proposal distribution*, and from an *acceptance probability* function $\alpha(\boldsymbol{\theta}, \mathbf{y})$. The proposal distribution is defined to be easy to simulate and the acceptance probability is defined such that the detailed balanced condition (see below) is verified by the transition kernel of the chain. In each step the algorithm generates a *candidate* $\mathbf{y}$ for the next state of the chain, given its present state $\boldsymbol{\theta}$, which is accepted or rejected depending of the value of $\alpha(\boldsymbol{\theta}, \mathbf{y})$. If rejected the chain remains in the present state $\boldsymbol{\theta}$; if accepted the chain transitions to $\mathbf{y}$.

For a transition $K(\boldsymbol{\theta}, d\mathbf{y})$ with $\boldsymbol{\theta} \notin d\mathbf{y}$, let us express the kernel density by

$$k(\boldsymbol{\theta}, \mathbf{y}) = q(\mathbf{y} \mid \boldsymbol{\theta})\, \alpha(\boldsymbol{\theta}, \mathbf{y}) \tag{3.31}$$

and assume that $k(\boldsymbol{\theta}, \boldsymbol{\theta}) = 0$. If it is verified that $\pi(\boldsymbol{\theta})\, q(\mathbf{y} \mid \boldsymbol{\theta}) > \pi(\mathbf{y})\, q(\boldsymbol{\theta} \mid \mathbf{y})$, it means, roughly speaking, that the chain transitions more often from the state $\boldsymbol{\theta}$ to the state $\mathbf{y}$ than from the state $\mathbf{y}$ to the state $\boldsymbol{\theta}$. Therefore, to compensate this fact, we set $\alpha(\mathbf{y}, \boldsymbol{\theta}) = 1$, and derive $\alpha(\boldsymbol{\theta}, \mathbf{y})$ imposing the *detailed balanced condition*, namely

$$\begin{gathered} \pi(\boldsymbol{\theta})\, q(\mathbf{y} \mid \boldsymbol{\theta})\, \alpha(\boldsymbol{\theta}, \mathbf{y}) = \pi(\mathbf{y})\, q(\boldsymbol{\theta} \mid \mathbf{y})\, \alpha(y, \boldsymbol{\theta}) = \pi(\mathbf{y})\, q(\boldsymbol{\theta} \mid \mathbf{y}) \\ \Rightarrow \alpha(\boldsymbol{\theta}, \mathbf{y}) = \frac{\pi(\mathbf{y})\, q(\boldsymbol{\theta} \mid \mathbf{y})}{\pi(\boldsymbol{\theta})\, q(y \mid \boldsymbol{\theta})} \end{gathered} \tag{3.32}$$

Note that if $\pi(\boldsymbol{\theta})\, q(\mathbf{y} \mid \boldsymbol{\theta}) < \pi(\mathbf{y})\, q(\boldsymbol{\theta} \mid \mathbf{y})$ the transition always is accepted and in this case it holds that $\alpha(\boldsymbol{\theta}, \mathbf{y}) = 1$. Therefore, the algorithm defines the acceptance probability by

$$\alpha(\boldsymbol{\theta}, \mathbf{y}) = \min\left[\frac{\pi(\mathbf{y})\, q(\boldsymbol{\theta} \mid \mathbf{y})}{\pi(\boldsymbol{\theta})\, q(\mathbf{y} \mid \boldsymbol{\theta})}, 1\right] \tag{3.33}$$

To complete the kernel definition rest to express the probability $r(\boldsymbol{\theta})$ of a transition $K(\boldsymbol{\theta}, d\mathbf{y})$ when $\boldsymbol{\theta} \in d\mathbf{y}$, that is, the probability that the transition not to be accepted and the chain remains in the present state, which is given by

$$r(\boldsymbol{\theta}) = 1 - \int_{\mathsf{X}} q(\mathbf{y} \mid \boldsymbol{\theta})\, \alpha(\boldsymbol{\theta}, \mathbf{y})\, d\mathbf{y} \tag{3.34}$$

Therefore the transition density for the MH algorithm is given by

$$k(\boldsymbol{\theta}, \mathbf{y}) = q(\mathbf{y} \mid \boldsymbol{\theta})\, \alpha(\boldsymbol{\theta}, \mathbf{y}) + \delta_{\boldsymbol{\theta}}(\mathbf{y})\, r(\boldsymbol{\theta}) \tag{3.35}$$

, where $\delta_{\boldsymbol{\theta}}(\mathbf{y})$ is the Dirac delta function.

The detailed balance condition 3.32 guarantees that $\pi(\boldsymbol{\theta})$ is invariant w.r.t. to the Markov chain defined by Eq. 3.35 (see e.g. Chib and Greenberg, 1995). Otherwise, a sufficient conditions for ergodicity can be found in Gentle et al. (2012).

3.4.4 Slice Sampling

Let be θ a random variable whose PDF $\pi(\theta) = \frac{f(\theta)}{k}$ is our target distribution. To sample $\pi(\theta)$ it is introduced an auxiliary random variable υ such that $\pi(\theta)$ is the marginal in θ of a joint distribution defined by

$$(\theta, \upsilon) \sim \mathsf{U}\{(\theta, \upsilon) : 0 < \upsilon < f(\theta)\} \tag{3.36}$$

The idea in to generate a Markov chain whose stationary distribution is equal to the distribution in Equation 3.36. For that, in the simplest version of this family of algorithms (Neal, 2003), for the t-th iteration a transition from the state $\left(\theta^{(t)}, \upsilon^{(t)}\right)$ to the state $\left(\theta^{(t+1)}, \upsilon^{(t+1)}\right)$ is done in two following steps

1. By sampling $\upsilon^{(t+1)}$ from
$$\upsilon \mid \theta^{(t)} \sim \mathsf{U}\left[0, f\left(\theta^{(t)}\right)\right] \tag{3.37}$$

2. By sampling $\theta^{(t+1)}$ from
$$\theta \mid \upsilon^{(t+1)} \sim \mathsf{U}\left[\upsilon^{(t+1)}, f(\theta)\right] \tag{3.38}$$

The transition density is then given by

$$k(\theta, \upsilon; \theta', \upsilon') = p(\upsilon' \mid \theta) \cdot p(\theta' \mid \upsilon') \tag{3.39}$$

3.4.5 The Gibbs Sampler

The Gibbs sampling algorithm can be considered as a particular case of MH (see Robert and Casella, 2004, Theorem 10.13). In the context of our work it is better seen as an algorithmic scheme that acts as a container for other univariate sampling algorithms. This is is in fact the approach followed by BUGS (see Lunn et al. (2012), Sec. 4.2). The algorithmic scheme uses a Markov chain whose transition density is given by the product of the full conditional densities for the BN parameters, that is, the product of the densities for each parameter given the rest of parameters and the data, namely

$$k\left(\boldsymbol{\theta}, \boldsymbol{\theta}'\right) = \prod_{k=1}^{P} p\left(\theta_k \mid \theta_1^{(t+1)}, ..., \theta_{k-1}^{(t+1)}, \theta_{k+1}^{(t)}, ..., \theta_P^{(t)}, D\right) \tag{3.40}$$

, where we have assumed that the network has P random parameters once its basic structure has been replicated. At first glance may seem that estimation of the joint posterior distribution for parameters by means of the kernel in Eq. 3.40 is computationally very expensive. Nevertheless the cost is significantly reduced if each full conditional is computed conditioning only by its Markov blanket. Therefore, accordingly to Eq. 3.11 and Cor. 3.1, we have that $p\left(\theta_k \mid \theta_k^{(t+1)}, ..., \theta_{k-1}^{(t+1)}, \theta_{k+1}^{(t)}, ..., \theta_N^{(t)}, D\right)$ can be reduced for every k to

$$p\left(\theta_k \mid \mathrm{bl}\left(\theta_k\right)\right) \propto f\left(\theta_k \mid \mathrm{pa}\left(\theta_k\right)\right) \times \prod_{v \in \mathrm{ch}(\theta_k)} f\left(v \mid \mathrm{pa}\left(v\right)\right) \tag{3.41}$$

Sometimes it is possible to obtain a closed form for the conditional PDFs in Eq. 3.41, by using conjugate distributions, making possible a direct sampling. If that is not possible, the distribution shall be simulated using some MCMC algorithm like 1D MH. Whatever the case, at the end of the t-th cycle the algorithm will provide a single sample $\left(\theta_1^{(t+1)}, ..., \theta_P^{(t+1)}\right)$ from the joint posterior distribution.

3.4.6 Convergence Criteria

3.4.6.1 Autocorrelation function

Given a realization $\left\{x^{(t)}\right\}_{t=1}^{n}$ of a Markov chain the *sample autocorrelation function* $\hat{\rho}\left(k\right)$ is defined as

$$\hat{\rho}\left(k\right) = \frac{\hat{\gamma}\left(k\right)}{\hat{\gamma}\left(0\right)} \tag{3.42}$$

, where k is the *time lag* between observations ($k < n$) and $\hat{\gamma}(k)$ is the sample autocovariance function, given by

$$\hat{\gamma}(k) = \frac{1}{n-k} \sum_{t=1}^{n-k} \left(x^{(t)} - \hat{\mu}\right)\left(x^{(t+k)} - \hat{\mu}\right) \tag{3.43}$$

, where $\hat{\mu} = \frac{1}{n}\sum_{t=1}^{n} x^{(t)}$ is the mean.

3.4.6.2 Corrected Gelman and Rubin statistic

Unlike the autocorrelation function, this statistic (Brooks and Gelman, 1998) gives a measure of the degree of convergence to the stationary distribution taking into account a number of realizations of the stochastic process. The idea underlying the method is to compare the total sample variance of the Markov chain (within and between realizations) with the variance between realizations. Assuming m realizations $\left\{x_j^{(t)}\right\}_{t=1}^{n}$ of the Markov chain, let us define first the following estimators:

- $\hat{\mu}_j = \frac{1}{n}\sum_{t=1}^{n} x_j^{(t)}$ as the sample mean of the j-th realization,
- $\hat{\mu} = \frac{1}{m}\sum_{j=1}^{m} \hat{\mu}_j$ as the overall sample mean,
- $\hat{\gamma}_{0i}^{W} = \frac{1}{n-1}\sum_{t=1}^{n}\left(x_i^{(t)} - \hat{\mu}_i\right)^2$ as the sample variance *within* j-th realization,
- $\hat{\gamma}_0^W = \frac{1}{m}\sum_{j=1}^{m}\hat{\gamma}_{0j}^{W}$ as the mean of intra-variances,
- $\hat{\gamma}_0^B = \frac{n}{m-1}\sum_{j=1}^{m}\left(\hat{\mu}_j - \hat{\mu}\right)^2$ as the sample variance *between* realizations.

Then, given the sample variance of the "stationary distribution" by

$$\hat{\gamma}_0 = \left(1 - \frac{1}{n}\right)\hat{\gamma}_0^W + \frac{1}{n}\hat{\gamma}_0^B \tag{3.44}$$

and corrected, taking into account the sampling variability of $\hat{\mu}$, by

$$\hat{V} = \hat{\gamma}_0 + \frac{1}{mn}\hat{\gamma}_0^B \tag{3.45}$$

, the expression for the corrected Gelman and Rubin statistic[3], is

$$\hat{\rho}_c = \frac{d+3}{d+1}\frac{\hat{V}}{\hat{\gamma}_0^W} \tag{3.46}$$

[3]Named also as *corrected scale reduction factor* (CSRF) or *shrink factor*.

Journal	No	IF	T	P	Q
Monthly Notices of the Royal Astronomical Society	5	5.521	56	9	Q1
Physical Review D	1	4.691	56	14	Q2
Annals of Applied Statistics	1	2.237	117	10	Q1
The Astrophysical Journal	4	6.733	56	6	Q1
Statistical Analysis and Data Mining	1	-	-	-	-
Astronomy and Astrophysics	1	5.084	56	11	Q1
Statistical Methodology[4]	1	-	-	-	-

Table 3.1: Impact factor for the scientific journals in which the articles discussed in Section 3.5 have been published. Meaning of Abbreviations: No = number of analyzed papers published in the journal, IF = ISI Impact Factor (2012), T = Total of journals in the category (56 for astronomy & astrophysics and 117 for statistics & probability) , Q = Quartile for the journal.

, where $d = {}^{2\hat{V}^2}/_{\hat{\text{var}}(\hat{V})}$ are the degrees of freedom, estimated by the method of moments, of a Student $\mathtt{t}\left(\hat{\mu}, \hat{V}, d\right)$ used to construct a Bayesian credible interval for the target distribution, which is assumed to be Gaussian.

3.5 Hierarchical Bayesian Models in Astrostatistics

In last few years there has been a marked increase on interest in the use of the Bayesian graphical modeling framework in the area of astrostatistics (Loredo, 2013). This fact is reflected by the increasing number of publications about this topic in journals with impact factor such as those listed in Table 3.1. We have focused our revision mainly to the two last years trying to classify the papers according to the astrophysical problem to which they address.

Celestial mechanics Hogg et al. (2010) proposes a two level hierarchical model to infer the parameters of the eccentricity distribution of a population of binary stars (or exoplanet) affected by a single companion. The evidence is given by a number of radial velocity measurements for each observed object. The likelihood for a determinate radial velocity measure is Gaussian parametrized by the sum between the overall binary system velocity, common for all measures for a given star, the radial velocity equation for the star, applied to the measure, and a noise component. The radial velocity equation for each star is made to depend deterministically on its velocity amplitude, period, orbital phase, eccentricity

and longitude of perihelion and the noise variance is decomposed into two terms, one for the measure and an overall term for the star. In such a way the likelihood component for the set of measures for a determinate star depends on seven bottom level parameters. In turn, the authors propose two alternative top level parametrizations for the distribution of eccentricities with their corresponding priors.

Distance estimation/interstellar extinction maps Distance is a fundamental problem in astronomy whose resolution is necessary to understand the structure and evolution of stellar populations. There exist multiple factors with affect the correct determination of distances. One of them is the interstellar extinction. Sale (2012) proposes a two level hierarchical model which combines distance-extinction relationship (top level), stellar parameters (bottom level) and multiband photometry observations for the star population associated to a determinate sightline of our Galaxy.

Stellar formation and evolution Barentsen et al. (2013) use a Bayesian network to infer masses, ages and mass accretion rates for stars in a star-forming region. The observations are spectral energy distributions (SEDs) in 2 broad bands and 2 narrow band filters collected form two different photometric surveys. Kelly et al. (2012) develop a hierarchical Bayesian model to infer the parameters of IR (infrared) SEDs of dust emission from observed flux densities. Stein et al. (2013) propose a Bayesian graphical model for inferring the relationship between the initial mass of a Sun-like star and its final mass as a white dwarf.

Black holes and active galactic nuclei Hemberger et al. (2013) employ the Bayesian graphical modeling framework to infer fitting formulae for the final spin and gravitational energy radiated by a blank hole as a function of its initial spin. They use a simulated sample of black holes. Kelly et al. (2013) employ a two-level model to estimate the high-frequency X-ray power spectral density (PSD) for an active galactic nucleus (AGN) given a dataset consisting in two time series of photon counts (a count for the source and the other for the background) for a determinate AGN.

Solar physics Asensio Ramos and Arregui (2013) propose a hierarchical model to infer the parameters of solar coronal loops. In this model the data are constituted by a set

of temporal series giving the motion of the apex for each observed coronal loop. The continuous curves corresponding to those time series have an oscillatory component which is parametrized by means of its amplitude, the density contrast between the tube and the environment, the transverse inhomogeneity length scale, etc. (1st level). The prior distributions of the latter physical parameters are made to depend, in turn, on a set of hyperparameters (2rd level) whose posterior distributions are the object of inference.

Ultra-high energy cosmic rays Soiaporn et al. (2013) propose a three level hierarchical Bayesian framework to assess the association between ultra-high energy cosmic rays (UHECRs) and candidate source populations. The levels of parameters (distributed over a space of association graphs) include, from top to down, source properties, cosmic ray production and cosmic ray propagation; plus an traditional level of observables (detection and measurement).

Gravitational wave astronomy Adams et al. (2012) propose a hierarchical model to infer the spatial distribution and the chirp mass distribution of the white dwarf binaries population in the Milky Way. They use data provided by a space based gravitational wave detector, namely by the future LISA mission. The datum for a single source includes a waveform plus a realization of the LISA instrument noise. The likelihood for the white dwarf binary signal is Gaussian parameterized by the frequency of the source, its distance, its mass, the angles of inclination, polarization and phase and sky location parameters (bottom level). In turn, for the galaxy shape, authors employ a bulge plus disk model consisting in a mixture of two distributions with four top level parameters. The model is completed with three additional hyperparameters for the chirp mass distribution.

Cosmological parameters Brewer et al. (2014) propose a two level hierarchical model to test the hypothesis that the stellar initial mass function (IMF) may vary within and between galaxies. March et al. (2011) and Weyant et al. (2013) present Bayesian hierarchical models to infer cosmological parameters from Type Ia supernovae data. Martinez (2013) proposes a two level hierarchical model to infer the overall distribution of masses for a dwarf spheroidal galaxies population, belonging to our Local Group, in the context of the Λ-CDM cosmological paradigm. In this model data are constituted by a set of triples, one for each galaxy, of the observed values for its half-light radius, the mass enclosed within the half-light

radius (or equivalently, the velocity dispersion), its total luminosity and the corresponding errors. Likelihood component for each dwarf galaxy is parameterized by a bottom level of parameters which are its maximum circular velocity, the radius corresponding to this velocity and its real luminosity. For these parameters are assumed linear relationships which are parameterized taking into account four theoretical models for the underlying dark matter density profiles.

Chapter 4

A Bayesian Graphical Model for Frequency Recovery

This chapter covers the three former stages of the general methodology for constructing our BGM, proposed in Section 1.3. It is structured in five sections as follows. Section 4.1 describes our methodological approach. Section 4.2 is devoted to the domain analysis. Sections 4.3 is devoted to the BGM construction. Section 4.4 specifies the MCMC algorithm used for inference. Finally, Section 4.5 treats some relevant aspects of our implementation in the BUGS language.

4.1 Methodology

As we saw in Chapter 3 the construction of a BGM involves the inclusion of both qualitative and quantitative information. The quantitative information corresponds to priors being the rest of this type of information gathered by inference. Inference consists in determining the posterior distributions for parameters of interest given the data. It is carried out by means of a determinate algorithm that we must specify once constructed the graphical model. Taking into account these considerations we can summarize our methodology in the following four stages:

1. **Domain Analysis**. We select, with the aid of the domain expert, a set of relevant variables to the problem that we want to solve. This task is done from a simulated data base composed of real attributes of a population of periodic variable stars, and from

information related to the Gaia mission design, and the CU7 software that process the raw Gaia data. We seek dependence relationships between the selected variables. We also try to identify, in our role of external observer of the domain, the presence of implicit variables associated to the data, e.g. categories linked to determinate values of other variables. The objective of this first stage is to acquire the knowledge, analytical whenever possible or otherwise experimental, needed for the completion of the model we are constructing.[1]

2. **Definition of the structure, priors and factorization of the model**. From the knowledge gathered in the first stage we proceed to construct the graphical model. This construction begins with a qualitative stage in which we specify the nodes, the probabilistic dependencies (arcs) of each node given its parents, the functional forms of these dependencies and the type of distribution of each (non-orphan) node given its parents. It is followed by a quantitative phase in which we specify our prior knowledge (or lack thereof) about the numerical parameters of the distributions of the orphan nodes in the graph and other fixed parameters associated to non-orphan nodes. Finally, we specify the factorization of the joint probability density function associated with the model.

3. **Inference Algorithm**.We specify an MCMC algorithm which consist in a Gibbs sampling scheme. This scheme samples, in each iteration, the marginal posterior distribution of parameters that are the focus of inference given the data.

4. **Implementation**. We implement the model constructed in the second stage using the declarative BUGS language.

[1] (a) It is important to clarify that although we can inject the generic form of the distribution of a real attribute in the model, e.g. a mixture of Gaussian distributions, we can not inject the exact values of the parameters of this distribution. The distribution of these parameters is the inference target and their values can only be used for evaluation purposes as we will see in the next chapter.

4.2 Domain Analysis

4.2.1 Database of Simulated Cepheids

Our starting point is a database with 36 attributes and N=36688 instances corresponding to a simulated sample of the classical LMC cepheids population. Three CU7 working groups have been involved in the generation of the database. These groups have tried to reproduce a process that begins with the collection of the light emitted by each periodic variable source by means of the CCDs in the Gaia astrometric field and ends with the extraction of a set of attributes of the sources from the gathered light. We have been forced to work with data generated with a simulated process because during the development of this thesis the Gaia mission was not yet operational . At knowledge level and according to the role played in the recovery process, we can establish a first classification of the database attributes in the following three categories:

Real attributes. These attributes correspond mainly to the parameters of the light curve of a periodic variable star observed by Gaia at a given position of the sky. They can in turn be classified as *direct measurements* or *derived measure*ments. Direct measurements are the *equatorial coordinates* (α, δ) of the source and its ecliptic latitude β. Derived measurements are the trigonometric parallax π, the *mean apparent magnitude* $\overline{m}_G$ of the light curve $m_G(t)$ in the G Band, the *peak-to-peak amplitude* A of the light curve, its *periodicity* P in days (equivalent to its *frequency* $\nu = \frac{1}{P}[d^{-1}]$), its distance r, the interstellar extinction E_V and the colour index $m_V - m_{I_C}$ [2]. The values of all these attributes have been generated artificially by the QA working group (García Sedano, 2012) from the knowledge existing in the astrophysical literature specialized in the study of the LMC Cepheid population. At this input level[3], although not included explicitly in the database, it is customary to take into account the existence of a complete model of the source light curve $m_G(t)$ involved in the simulation process (see Section 2.3.2).

[2] The Gaia colour index is obtained by subtracting the magnitudes in its two photometric filters as $m_{GB} - m_{GP}$, but we have not been provided with simulations of this parameter.

[3] Keeping in mind the suitability of analyzing an information processing system as a set of transformations between representation spaces (see e.g. Mira and Delgado, 2009, Section 2), we will use interchangeably the terms "real attribute" and "input attribute" and we will do the same with the terms "recovered attribute" and "output attribute".

Satellite attributes. These attributes are related to the way in which Gaia collects the input attributes and include the *number of observations* or *transits* o of the source, the *photometric error* $\sigma_{\overline{m}_G}$ associated to its mean apparent magnitude $\overline{m}_G$ and the parallax error σ_π. The photometric measurement error, simulated by the QA group based on Jordi et al. (2009), constitutes only a first source of error in the recovery process. As we saw in Section 2.2.1, the telescope only take a finite number of measurements according to the so-called *scanning law*. These brightness measurements represent a discrete time series $\{m_G(t_k)\}_{k=1}^{o}$ sampled from the continuous light curve $m_G(t)$. This sampling procedure constitutes a potential source for systematic biases due to the sampling regularities. To analyse these potential biases, the BE group has simulated a sequence of observation epochs $\{t_k\}_{k=1}^{o}$ for every source given its celestial coordinates and the scanning law and generated the corresponding magnitude measurements. Unfortunately only the parameter o has been retained in the database which constitutes a crude approximation of the way in which the satellite scans the sky.

Recovered attributes. These database attributes correspond to the information (the evidence) recovered from the simulated astronomical time series $\{m_G(t_k)\}_{k=1}^{o}$ and include: the *recovered mean apparent magnitude* $\overline{m}_{G,\mathrm{rec}}$ of the light curve of the source, the *recovered peak-to-peak amplitude* A_{rec} of its variability, the *recovered period* P_{rec} (or, equivalently, the *recovered frequency* $\nu_{\mathrm{rec}} = \frac{1}{P_{\mathrm{rec}}}$) and the results of a polynomial fit of the time series consisting in the first terms of a Fourier series as triplets in the form amplitude-phase-order $\{(A_{\mathrm{rec},j}, \phi_{\mathrm{rec},j}, j)\}$ with indication of the residual. It is important to remark that the only telescope output is the photometric time series (not included in the database used in this work) and that the output attributes listed therein are the product of a post processing stage applied by the Characterization software package (see Section 2.3.2).

4.2.2 Analytical Relations between Attributes

The domain expert proposed the development of a model to infer the real distributions of several attributes, namely, the peak to peak amplitude, the (decimal) logarithm of the frequency and the apparent G-magnitude, from the values recovered by Gaia. For each of these attributes we have a pair $(V_{\mathrm{input}}, V_{\mathrm{rec}})$ constituted by the input (real) attribute and its corresponding recovered (output) attribute. We show the empirical distributions

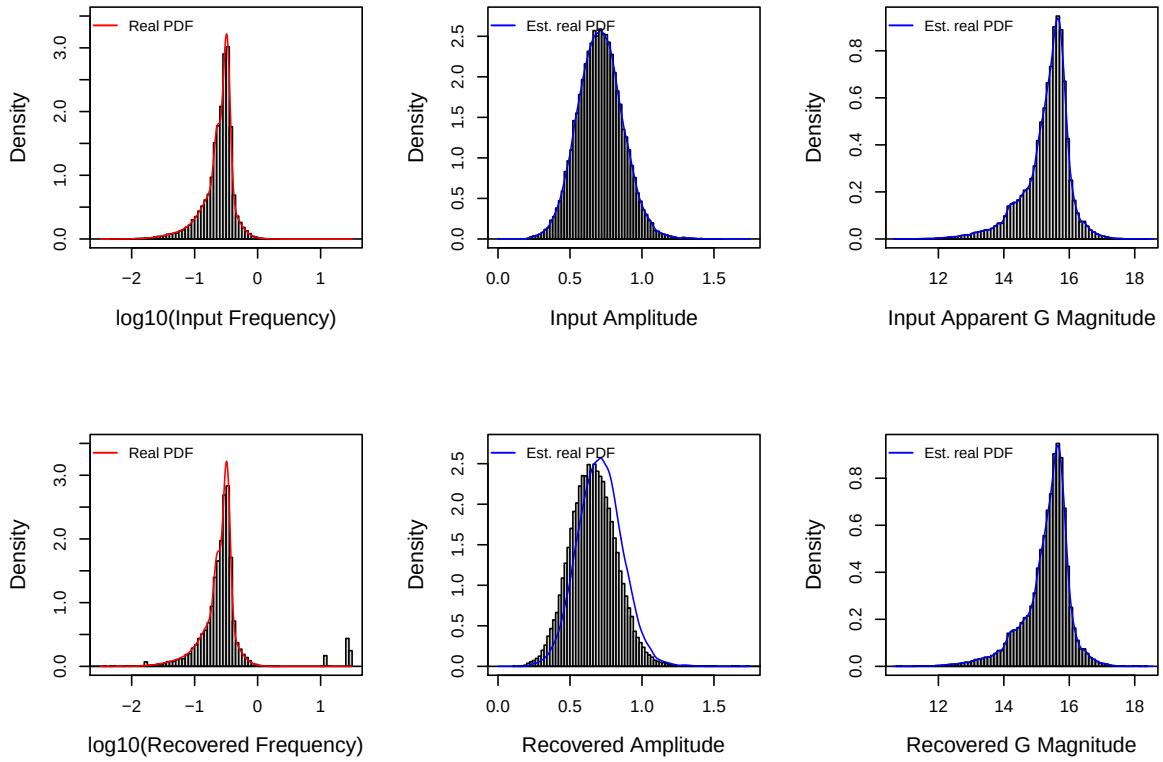


Figure 4.1: Real and recovered marginal distributions of amplitudes, apparent G magnitudes and frequencies. Solid lines depicted in blue represent the PDF of real (input) amplitude, and apparent G magnitude distributions estimated from the corresponding input histogram.

of these pairs in Figure 4.1. We can see that: i) the recovered amplitude distribution becomes slightly shifted to the left as compared to the input one; ii) the distributions in the pair $(\overline{m}_{G,\text{input}}, \overline{m}_{G,\text{rec}})$ are quite similar; and iii) three sub distributions (peaks) of output frequencies appear which were not found in the input distribution. In principle we know the analytical PDF family for the V_{input} attributes because we know how their samples have been generated. However, problems arise when the input distribution depends on other variables not explicitly included in the model.

Input frequency distribution This marginal distribution has been sampled from a mixture of five normal distributions parametrized by the QA group from Antonello et al. (2002) by

$$f(\log(\nu)) \approx 0.12 \cdot \mathsf{N}(-1, 0.35) + 0.11 \cdot \mathsf{N}(-0.64, 0.05) + \tag{4.1}$$
$$0.42 \cdot \mathsf{N}(-0.62, 0.22) + 0.10 \cdot \mathsf{N}(-0.54, 0.05) + 0.25 \cdot \mathsf{N}(-0.48, 0.05)$$

This PDF is depicted with red lines in the left column of Fig. 4.1.

Input amplitude distribution The input sample is shown in the top center of Fig. 4.1 and the marginal distribution represented with a solid blue line. This distribution has been simulated from the frequency using the OGLE III catalog[4]. This dependence on the frequency has been modelled analytically as

$$f(A \mid \log(\nu)) = \begin{cases} \mathsf{N}(-0.5 \cdot \log(\nu) + 0.2, 0.15) & \log(\nu) < -1 \\ \mathsf{N}(0.7, 0.15) & \log(\nu) > -1 \end{cases} \tag{4.2}$$

Input apparent G-magnitude distribution The apparent G-magnitudes have been generated by García Sedano (2012) from the apparent magnitudes m_V and the colour index $m_V - m_{I_C}$ based on (Jordi et al., 2010). The two latter variables depend in turn, respectively, on the interstellar extinction E_V, the distance r and the absolute magnitude M_V, and from the decimal logarithm of the frequency $\log(\nu)$. Finally. both the colour index $m_V - m_{I_C}$ and the absolute magnitudes M_V depend on $\log(\nu)$ (Sandage et al., 2004). The relations

[4] http://ogledb.astrouw.edu.pl/~ogle/CVS/ceph_query.html

between all these variables are given by the following Equations:

$$m_G = m_V - 0.0257 - 0.0924 \cdot (m_V - m_{I_C}) - 0.1623 \cdot (m_V - m_{I_C})^2 + 0.0090 \cdot (m_V - m_{I_C})^3 \tag{4.3}$$

$$m_V = E_V + M_V + 5\,(log\,(r) - 1) \tag{4.4}$$

$$f\,(m_V - m_{I_C} \mid \log\,(\nu)) = \begin{cases} \mathsf{N}\,(-0.315 \cdot \log\,(\nu) + 0.380, 0.1) & \log\,(\nu) < -1 \\ \mathsf{N}\,(-0.160 \cdot \log\,(\nu) + 0.501, 0.1) & \log\,(\nu) > -1 \end{cases} \tag{4.5}$$

$$f\,(M_V \mid \log\,(\nu)) = \begin{cases} \mathsf{N}\,(2.567 \cdot \log\,(\nu) - 1.634, 0.1) & \log\,(\nu) < -1 \\ \mathsf{N}\,(2.963 \cdot \log\,(\nu) - 1.335, 0.1) & \log\,(\nu) > -1 \end{cases} \tag{4.6}$$

Although there is no available information about the dependence $M_G \mid \log\,(\nu)$ in the literature, we can assume that its PDF family is the same as the family of the PDF expressed by Eq. 4.6. Otherwise if we consider Eq. 4.4 but applied to the G-magnitude and leaving out the extinction we have

$$m_G = M_G + 5\,(log\,(r) - 1) \tag{4.7}$$

So if we put together the two precedent hypotheses we have a way to model a dependence $m_G \mid \log\,(\nu)\,, \log\,(r)$.

Distance distribution This distribution has been generated by means of a successive number of deterministic relations that take into account the Geocentric equatorial coordinates $(\alpha_0, \delta_0, r_0)$ of the LMC center and an random spatial model of the galaxy as an exponential disk. We have designed and implemented an extended BGM in BUGS which includes a complete parameterization for distances. Nevertheless, the final version of our BGM presented in this thesis does not include such a parameterization. This is due to the fact that the extended model does not converge. A possible cause could be that we assume that the ecliptic latitudes β are constants (see Equation 4.23). But β is related with the Cartesian Geocentric equatorial coordinates (x'_i, y'_i, z'_i) of the stars, which are random variables in the extended model. We refer readers interested in this extension to Appendix A, after reading Section 4.3.

Number of transits and ecliptic latitude Let us now analyze the role of the number of transits o, a satellite attribute according to the classification of Section 4.2.1. This variable is the size of the time series $\{m_G(t_k)\}_{k=1}^{o}$, i.e the size of the sub-sample taken from each astronomical source and depends on (*is an effect of*) its equatorial coordinates (α, δ). It is beyond the scope of this thesis to model and parametrize this dependence, which would involve modelling the generation of the complete epoch sequence $\{t_k\}_{k=1}^{o}$ for each source taking into account the Gaia scanning law, process which is entrusted to BE. Anyway it is necessary to select some attribute to model *the cause* of systematic biases due to the sampling regularities. An option is to include the coordinates (α, δ) taken as constants in the model. The second option is to include the ecliptic latitudes β (also as constants). The influence of β over the rate of correct detection of periodic signal by Gaia has been studied by (Eyer and Mignard, 2005). In that article we see that, for high values of β, like of the LMC sources, the relation between the rate of correct detentions and β is approximately linear with a negative slope. Therefore, will adopt the latter strategy and study the influence of β over the output attributes through experiments.

Recovered G-magnitude distribution The recovered mean apparent G-magnitudes has been generated by QA from the input G-mag by

$$f\left(\overline{m}_{G,\mathrm{rec}} \mid \overline{m}_G\right) = \mathsf{N}\left(\overline{m}_G, \sigma_{G(\mathrm{rec})}\right) \tag{4.8}$$

with

$$\sigma_{G(\mathrm{rec})} = \sigma^{\mathrm{tr}}_{G(\mathrm{rec})} \cdot \frac{1}{\sqrt{n_{\mathrm{obs}}}} \tag{4.9}$$

, where $\sigma^{\mathrm{tr}}_{G(\mathrm{rec})}$ is the magnitude error or standard deviation for one transit and n_{obs} is the mean total number of transits (≈ 70) at the end of the mission (see Section 2.2.2). The error model expressed by Equations 4.8 and 4.9 assumes homokedasticity within each time series. This is only an approximation because the real time series are naturally heteroskedastic. But given the lack of detailed information about the series, we can only incorporate average apparent magnitudes and uncertainties per source to our models. The models are only heteroskedastic in the sense that these averages are per time series, and not for the entire set of sources.

Finally, for both the output amplitudes and frequencies we do not know the analyti-

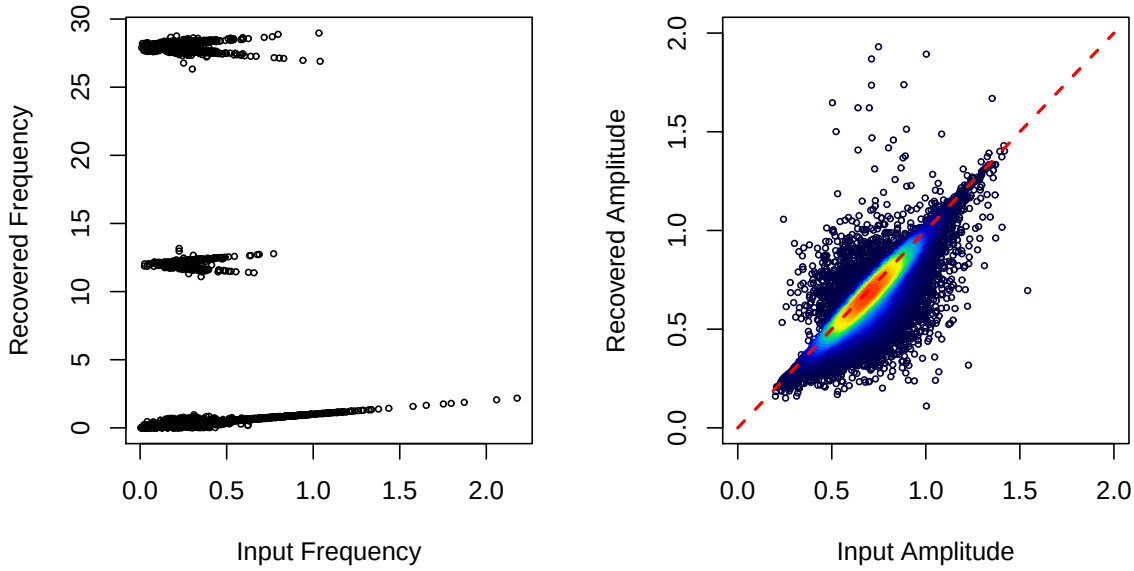


Figure 4.2: Dispersion graphs comparing input and recovered values for frequency and amplitude.

cal expression of their conditional distributions given the input variables and the ecliptic latitude. Therefore we must infer them from the analysis of their empirical distributions. experimental analysis from their empirical distributions. We will do it in the following sub-sections.

4.2.3 Global Analysis for Recovered Parameters

Description Our objective is to obtain a first approximation of the relationship between the input and recovered values for frequency and amplitude from the visual inspection of the dispersion graphs of these two pairs of variables.

Results and commentary Figure 4.2 shows the scatter plots of these two parameters. The most complex pattern is shown by the relation between the recovered and input frequency, depicted in the sub-figure to the left of Fig. 4.2, where we can see that pairs (ν, ν_{rec}) mostly fall in three zones: near the identity straight line $\nu_{\text{rec}} = \nu$ and near straight lines $\nu_{\text{rec}} = \pm\nu + b$ with b approximately equal to 12 or 28. We will assess the details of the distribution over these lines in Section 4.2.4.

For the relation between recovered and input amplitudes we found that 36580 of the 36645 instances of the LMC database fall in the 2x2 square to the right of Fig. 4.2[5]. We see that the distribution $A_{\text{rec}} \mid A$ is skewed in the sense that, if we fix a value for the input amplitude, the recovered amplitude falls over the semi plane $A_{\text{rec}} < A$ with higher frequency (density) than over the semiplane $A_{\text{rec}} > A$. This bias is coherent with the fact that the peak to peak amplitude extracted from a discrete and noiseless time series associated to each astronomical object is necessarily lower than the amplitude of its (continuous) light curve. We will get more insight on the concrete form for this bias in Section 4.2.5.

4.2.4 Detailed Analysis of Recovered Frequencies

4.2.4.1 A taxonomy for recovered frequencies

Description In this sub-section we have two objectives: **i)** to obtain a more detailed description, also by means of visual inspection, of the relationship between the input and recovered values of the frequency, focusing in the neighbours of the straight lines $\nu_{\text{rec}} = \nu$ and $\nu_{\text{rec}} = \pm\nu + b$ with b near 12 or 28 and **ii)** in case of encountering a pattern which allows us to classify the pairs (ν, ν_{rec}) in categories, do this classification within some predefined error margin, account the proportions of each class/category and represent the classes. For that:

1. We make dispersion graphs for pairs (ν, ν_{rec}) for each of the three zones discovered in the experimental analysis performed in Section 4.2.3.

2. We postulate the hypothesis that most of the recovered frequencies either are correctly recovered frequencies that fall in the identity straight line $\nu_{\text{rec}} = \nu$, either are aliased frequencies that fall in any of the straight lines of expression $\nu_{\text{rec}} = \pm\nu + b$, with $b \neq 0$ and not necessarily exactly equal to 12 or 28. Then we try to find concrete expressions for b from our knowledge about some regularities in the Gaia scanning law. For the latter purpose we generate a prototypical spectral window corresponding to an hypothetical source with ecliptic latitude in the LMC range and depict it with two levels of detail.

3. It is considered that a frequency ν_{rec} is recovered in a straight line $\nu_{\text{rec}} = \pm\nu + b$ if the

[5] We do not include in the graph the remaining 65 pairs (A, A_{rec}) which correspond to recovered amplitudes in the range $(2, 2412035]$.

slope	$\pm k_2$	No	%
1	$+2$	1248	3.41
1	-2	1218	3.32
1	$+7$	76	0.21
1	-7	103	0.28
1	$+9$	180	0.49
1	-9	187	0.51
± 1	Rest up ± 19	275	0.75
	Total	3287	8.97

(a) Substructure for $k_1 = 0$.

slope	$\pm k_2$	No	%
1	$+1$	49	0.13
-1	$+1$	51	0.14
1	-1	76	0.21
-1	-1	64	0.17
± 1	Rest up ± 19	69	0.19
	Total	309	0.84

(b) Substructure for $+k_1 = +3$.

slope	$\pm k_2$	No	%
1	$+1$	108	0.29
-1	$+1$	113	0.31
1	-1	60	0.16
-1	-1	78	0.21
1	$+3$	105	0.29
-1	$+3$	79	0.22
1	-3	190	0.52
-1	-3	164	0.45
± 1	Rest up ± 19	330	0.90
	Total	1227	3.35

(c) Substructure for $+k_1 = +7$.

Category	No	%
Perfect recovery	31677	86.44
Aliases in Sub-Tables a to c	4823	13.16
Not classified	145	0.40
Total	36645	100

(d) Totals.

Table 4.1: Statistics of correctly recovered and incorrectly recovered (spurious) frequencies.

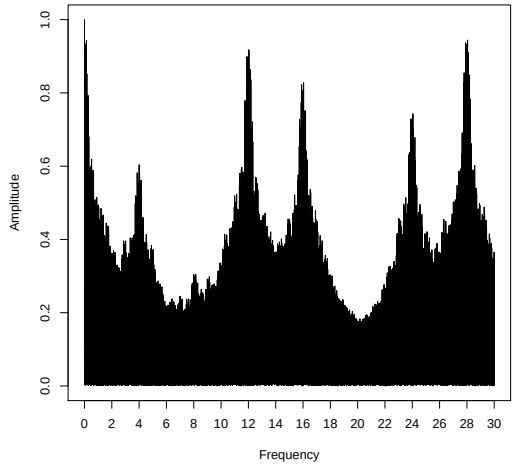
inequality $|\pm\nu + b - \nu_{\text{rec}}| < 0.005$ holds, i.e. when the absolute error is lower than a 0.5%.

Results and commentary The spectral window to use for comparison purposes is depicted in Figure 4.3 and the results of the experiment are presented in Figure 4.4 and Table 4.1.

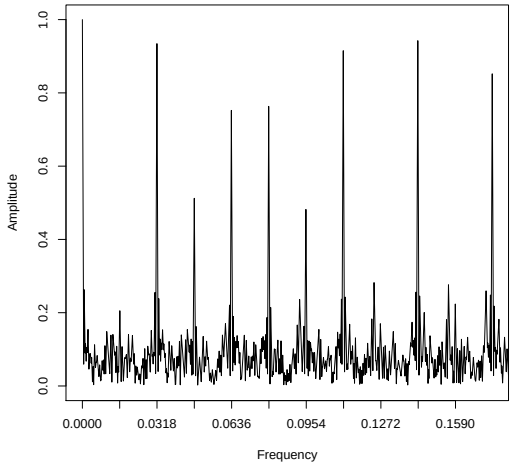
We find that most of the pairs $(\nu_{\text{input}}, \nu_{\text{rec}})$ fall on straights lines of the form:

$$\nu_{\text{rec}} = \pm\nu_{\text{input}} \pm k_1\nu_s \pm k_2\nu_p \tag{4.10}$$

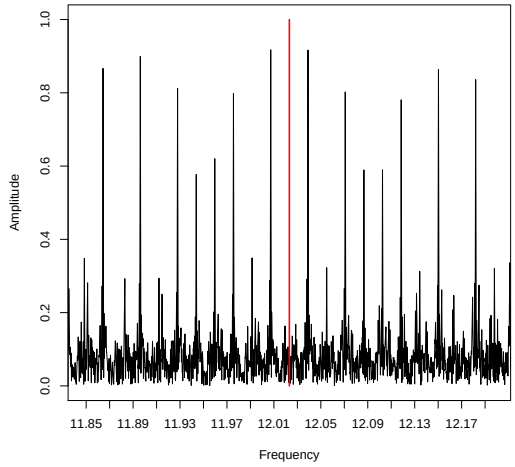
, where $k_1 \in \{0, 3, 7\}$, $k_2 \in \{0, ..., 19\}$, $\nu_s \approx {}^1/_{0.25} = 4\text{d}^{-1}$ is the rotational frequency of the satellite and $\nu_p = {}^1/_{63}\text{d}^{-1}$ is its precessional frequency. For taxonomic purposes we will refer to each of these lines as a *locus/category* of recovered frequencies. Moreover, if we fix k_1, we find a local *loci substructure* centered around the line $\nu_{\text{rec}} = \pm\nu_{\text{input}} \pm k_1\nu_s$ by



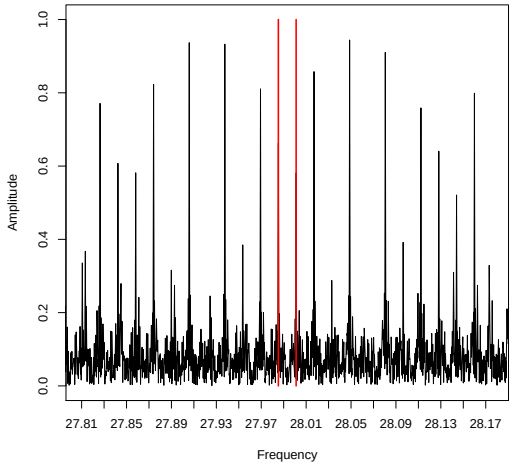
(a) Global view.



(b) Substructure with symmetry axis at $\nu = 0$.



(c) Substructure with symmetry axis at $\nu = 12.0232$.



(d) Substructure with central axis at $\nu' = 27.9853$ and $\nu'' = 28.0011$.

Figure 4.3: Prototypical spectral window for a LMC astronomical source. Sub-Figure a) depict a global perspective of the windows showing the existence of substructures of sampling frequencies for multiples of the Gaia's rotational frequency approximately equal to 4d^{-1}. Sub-Figures b) to d) depict each substructure in detail showing its symmetry and the separation between peaks equal to a precessional frequency of $^1/_{63}\text{d}^{-1}$.

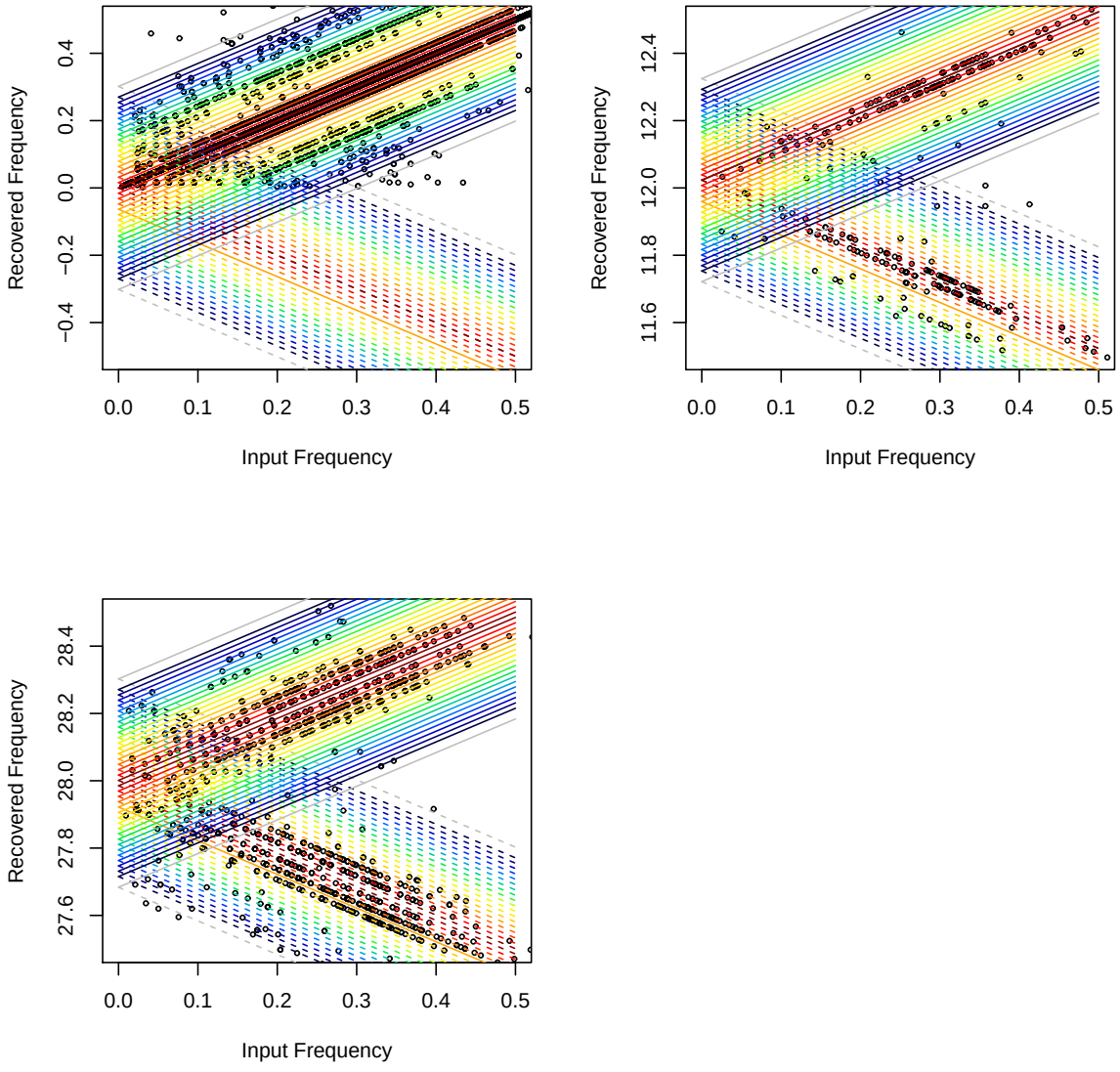


Figure 4.4: Detailed dispersion graphs comparing input and recovered values of frequencies and classes. Classification of correctly recovered and incorrectly recovered (spurious) frequencies.

varying k_2. Therefore, with the exception of the straight line with equation $\nu_{\rm rec} = \nu_{\rm input}$ all these *loci* are of spurious (aliased) frequencies[6].

Let us now to examine in detail the three scenarios found. The **first scenario**, depicted in the top row to the left of Fig. 4.4 with the associated reference spectral window zoomed in Fig. 4.3b, corresponds to the local substructure around the identity line $\nu_{\rm rec} = \nu_{\rm input}$ ($k_1 = 0$). Here, as we collect in Table 4.1c and 4.1a, we find that a 86.44% of the total number of frequencies in the database are "perfectly" recovered in that line, and a 8.22% are recovered in some alias with expression $\nu_{\rm rec} = +\nu_{\rm input} \pm k_2\nu_p$ with $\pm k_2 \in \{\pm 2, \pm 7, \pm 9\}$. In the sub-figure of Fig. 4.4 the identity line is depicted in dark red and the aliases are depicted in degraded colour. Note that the distance $\nu_P = {}^1\!/_{63}\mathrm{d}^{-1}$ between parallel straight lines is equal to the spacing between peaks in the zoomed spectral window. Also, the most frequent aliases correspond to the peaks with highest amplitudes in $\nu = 2 \cdot \nu_p$ and $\nu = 9 \cdot \nu_p$.

The **second scenario**, depicted in the top row to the right of Fig. 4.4 with the associated reference spectral window zoomed in Fig. 4.3c, corresponds to the local substructure around the straight lines $\nu_{\rm rec} = \pm\nu_{\rm input} + 3\nu_s$. In this case we observe a local symmetry in the spectral window around $\nu = 3\nu_s = 12.0232$ with $\nu_s = 4.0077$ and the same separation between peaks $\nu_p = {}^1\!/_{63}\mathrm{d}^{-1}$ that in the first scenario. This symmetry axis is depicted as a dark red line in Fig. 4.3c and in the corresponding sub figure of Fig. 4.4 (for positive and negative slopes). We find that this second scenario is one order of magnitude less probable that the first, being in it only a 0.84% of the total number of recovered frequencies, mainly for $\pm k_2 \in \{\pm 1\}$ (see Table 4.1b).

Finally, the **third scenario** is depicted in the bottom row of Fig. 4.4 with the associated reference spectral window zoomed in Fig. 4.3d, In this case we appreciate in the spectral window two central axis (depicted in dark red): i) $\nu' = 7\nu'_s = 27.9853$ with $\nu'_s = 3.9979$ and ii) $\nu'' = 7\nu''_s = 28.0011$ with $\nu_s = 4.0002$, separated by $\nu_p = {}^1\!/_{63}\mathrm{d}^{-1}$ being the window peaks also spaced by ν_p. In the corresponding sub-figure of Fig. 4.4 we observe that the straight lines with negative intercept are distributed as $\nu_{\rm rec} = \pm\nu_{\rm input} + 7\nu'_s - k_2\nu_p$ while the straight lines with positive intercept are distributed as $\nu_{\rm rec} = \pm\nu_{\rm input} + 7\nu''_s + k_2\nu_p$. In

[6]Note that some of the lines in Eq. 4.10 actually could not to appear in the dataset for the complete range of $\nu_{\rm input}$, because only are recovered positive frequencies; e.g., the line $\nu_{\rm rec} = \nu_{\rm input} + k_1\nu_s - k_2\nu_p$ can only appear when the condition $\nu_{\rm input} > -k_1\nu_s + k_2\nu_p$ holds, the line $\nu_{rec} = -\nu_{\rm input} - k_1\nu_s - k_2\nu_p$ never can appear and the line $\nu_{rec} = +\nu_{\rm input} + k_1\nu_s + k_2\nu_p$ always appears. Note otherwise that the recovered frequencies of the form $\nu_{\rm rec} = +\nu_{\rm input} \pm ...$ and $\nu_{\rm rec} = -\nu_{\rm input} \pm ...$, with with nonzero intercept, are aliased of $\nu_{\rm input}$and $-\nu_{\rm input}$ respectively.

this scenario are recovered as aliases, with $\pm k_2 \in \{\pm 1, \pm 3\}$, a 3.35% of the total number of recovered frequencies (see Table 4.1c).

4.2.4.2 Dependence on the ecliptic latitude

Description The experiment objective is to analyze the proportions (percentages) of frequencies recovered in each *locus* $\nu_{\text{rec}} = \pm\nu_{\text{input}} \pm k_1\nu_s \pm k_2\nu_p$ when we fix the values of the ecliptic latitude β. For that:

1. We consider *loci* in the two more frequent scenarios according to the results of the experiment of sub-section 4.2.4.1, namely, the substructures for $k_1 \in \{0, 7\}$.

2. The values of β are binned in ten intervals $[\beta_1^i, \beta_2^i]$ $(i \in \{1, ..., 10\})$ with the same number of instances $f_i\,(\nu_{\text{rec}})$ in each.

3. We compute the proportions of each *locus* for the i-th bin as the quotient between the number of frequencies in the j-th *locus* $f_i^j\,(\nu_{\text{rec}})$ (for the interval) and the total number of frequencies (instances) $f_i\,(\nu_{\text{rec}})$ in the interval. Then we make dispersion graphs of the pairs $\left(\frac{\beta_1^i+\beta_2^i}{2}, \frac{f_i^j(\nu_{\text{rec}})}{f_i(\nu_{\text{rec}})}\right)$ for each *locus*. Therefore, considering the more frequent *loci* and fixed the i-th bin we have $\sum_j f_i^j(\nu_{\text{rec}})/f_i(\nu_{\text{rec}}) \approx 1$.

4. It is considered that the frequency ν_{rec} of a star is recovered in the *locus* $\nu_{\text{rec}} = \pm\nu_{\text{input}} \pm k_1\nu_s \pm k_2\nu_p$ if the inequality $|\pm\nu_{\text{input}} \pm k_1\nu_s \pm k_2\nu_p - \nu_{\text{rec}}| < 0.005$ holds, i.e. when the absolute error is lower than a 0.5%.

Results and Commentary The results of the experiment are presented in Figure 4.5. We find that for a perfect recovery, depicted in the sub-figure on the bottom row to the left, the proportion of recovered frequencies increases linearly with β. Otherwise, fixed a substructure k_1 of aliased frequencies proportion of recovered frequencies decreases linearly with β with a slope more pronounced as $|k_2|$ increases. This suggest us to model the probability of recovery of an input frequency in a determinate *locus* (category) by means of a logistic regression submodel with β as covariate.

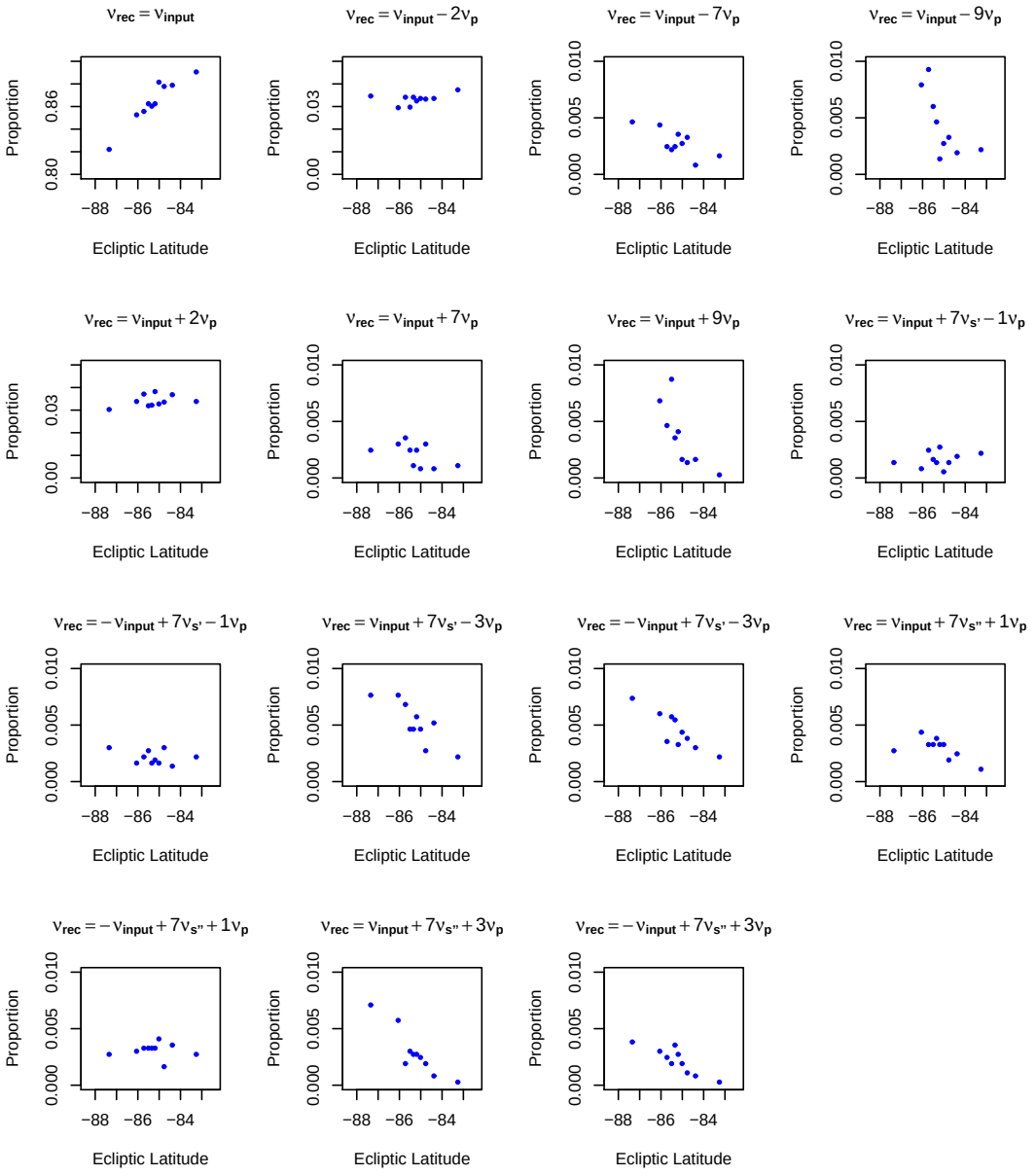


Figure 4.5: Proportions of recovered frequencies versus ecliptic latitude.

4.2.5 Detailed Analysis of Recovered Amplitudes

4.2.5.1 *Loci* and amplitude relationship

Description The experiment objective is to get insight in the form of the conditional distribution for the recovered amplitude given an input amplitude. Moreover, we try to check the hypothesis that recovered amplitudes are also subject, like recovered frequencies, to the influence of the ecliptic latitude. Therefore this suggests us to analyze the relationship between loci of frequencies and pairs $(A_{\text{input}}, A_{\text{rec}})$. For that we make dispersion graphs of these pairs for the most frequent *loci* in Table 4.1.

Results and Commentary The results of the experiment are presented in Figure 4.6. We find that for a perfect recovery, depicted in the sub-figure on top row to the left, the distribution $A_{\text{rec}} \mid A$ is skewed in the sense that we conclude in the experiment of Section 4.2.3 with a central parameter approximately equal to the input frequency, although we can also appreciate in this first scenario a superimposed horizontal fringe whose ordinates do not seem to depend on the input amplitude. Otherwise, for loci of aliased frequencies we observe that skewness increases as the input frequency does according to a certain slope to be determined.

4.2.5.2 Parameter estimation for dependence on real amplitudes

In light of the results of the experiment of sub-section 4.2.5.1 we try now to obtain concrete values of parameters for the dependence $A_{\text{rec}} \mid A_{\text{input}}$ given a determinate *locus*. For this purpose we assume the homoscedasticity hypothesis for errors and fit two linear regression models

$$A_{rec,i} = \beta_1^j A_{rec,i} + \beta_0^j + \epsilon_i^j,\ j = 1, 2$$

, the former for the identity locus and the latter for *loci* $\nu_{\text{rec}} = \pm\nu_{\text{in}} + 7\nu_s - 3\nu_p$ [7]. For the identity locus we assume for the error component a skewed Student t distribution (Azzalini and Genton, 2008) with one degree of freedom, that is, a skewed Cauchy distribution $\epsilon_i^1 \sim \mathsf{st}\,(0, \omega, \alpha, 1)$ where ω and α denote respectively the shape and scale parameters. We fit that regression model by using the SN package (Azzalini, 2013). Otherwise, for the *locus*

[7] We only select one locus of aliased frequencies (and its symmetric) bearing in mind the implementation and evaluation of our BGM model.

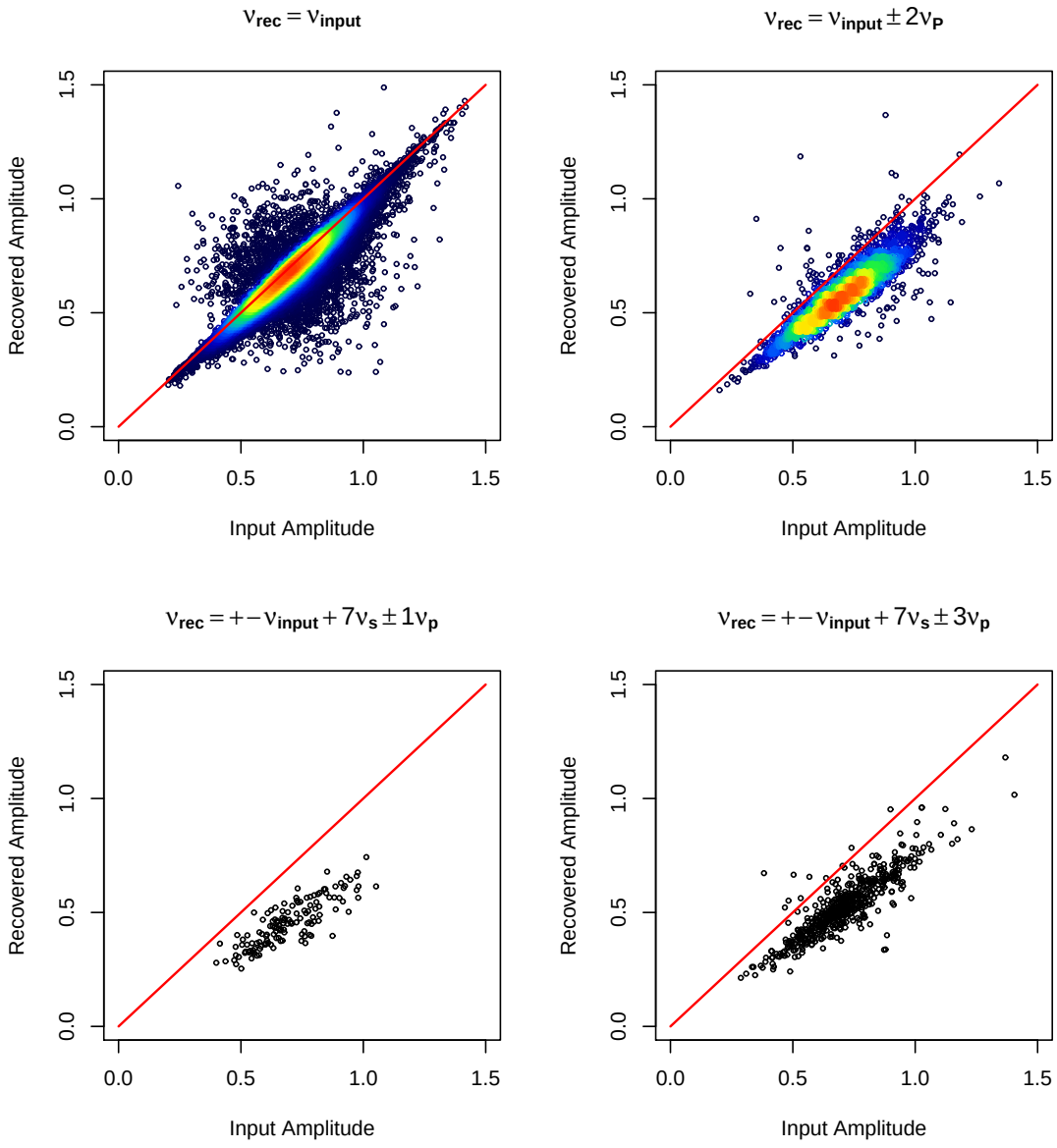


Figure 4.6: Loci and amplitude relationship.

Locus	β_0	β_1	ω	α
$\nu_{\text{rec}} = \nu_{\text{in}}$	0.0037	0.9993	0.0204	-2.3953
$\nu_{\text{rec}} = \pm\nu_{\text{in}} + 7\nu_s - 3\nu_p$	-0.008440	0.749224	0.0266	0

(a) Parameter estimations.

Locus	β_0	β_1	ω	α
$\nu_{\text{rec}} = \nu_{\text{input}}$	0.0004	0.0006	0.0002	0.0418
$\nu_{\text{rec}} = \pm\nu_{\text{in}} + 7\nu_s - 3\nu_p$	0.009162	0.012588	-	-

(b) Errors.

Table 4.2: Results of fitting Cauchy regression models for recovered amplitudes.

$\nu_{\text{rec}} = \pm\nu_{\text{in}} + 7\nu_s - 3\nu_p$ we assume that $\epsilon_i^2 \sim \mathsf{t}\,(0, \omega, 1)$ and use the HETT package (Taylor, 2012) for fitting.

Results and Commentary The results of the experiment are presented in and Table 4.2. We will use these data below to parameterize the conditional $A_{\text{rec}} \mid A_{\text{input}}$.

4.3 BGM Construction

In the present section we develop the qualitative phase of our BGM construction in consonance with the suggestions of the domain expert and the experimental results obtained in Sections 4.2.3 to 4.2.5. Simultaneously we include the quantitative information corresponding to our prior knowledge about fixed parameters in the model.

4.3.1 Graph Structure

The graph structure of our model is depicted in Figure 4.7. In the figure we mainly include the nodes that are random variables not computed deterministically given their parents. This is done to facilitate the ulterior task of factorizing the joint probability distribution associated to the graph. We summarize in Table 4.3 a description of the meaning of the nodes/variables/parameters in the graph and their type of distribution (see Section 4.3.2).

We group the nodes in the graph into subvectors following an approximate hierarchy. This hierarchy distinguish between the evidential nodes, the rest of nodes (attributes) which replicate with the rectangle in Fig. 4.7 and the nodes outside the rectangle. In our case this is better than perform a strictly hierarchical grouping because e.g. we have orphan nodes located in different graph levels. We also distinguish, within the hierarchy, four classes of

nodes by their level/type of knowledge seen from the perspective of the external observer. According to these criteria we have:

1. A bottom level constituted by the *evidential nodes* which represent the observed variables in the sample:

$$\mathcal{D} = (\nu_{\text{rec},i}, A_{\text{rec},i}, m_{G_{\text{rec}},i}) \tag{4.11}$$

   These nodes are denoted with double circles. They are the output/recovered frequency $\nu_{\text{rec},i}$, amplitude $A_{\text{rec},i}$ and apparent G-magnitude $m_{G_{\text{rec}},i}$ for the i-th star. They are enclosed by a rectangle (a *plate*) which is replicated as many times (N) as there are stars in the sample.

2. A first level of random parameters hierarchy:

$$\boldsymbol{\theta}_1 = \left(\log(\nu_i), A_i, m_{G,i}, T_{\nu_{\text{rec},i}}, T_{\nu_i}\right) \tag{4.12}$$

   We have two classes of nodes at this level. The set of *input nodes* is constituted by the real frequency $\log(\nu_i)$ (in logarithmic scale), the real peak-to-peak amplitude A_i and the real apparent G-magnitude $m_{G,i}$ of the i-th star. The *categorical nodes* $T_{\nu_{\text{rec},i}}$ and T_{ν_i} indicate the distribution to which a node belongs when the node is modeled by a mixture of distributions. T_{ν_i} is associated with the logarithm of the real frequency for the i-th star and $T_{\nu_{\text{rec},i}}$ is associated with its recovered frequency and recovered amplitude. All the nodes at the first level of random parameters hierarchy replicate with the plate. They depend on (but not only on) non informative orphan nodes which are outside the plate.

3. A top level of of random parameters hierarchy (hyperparameters):

$$\boldsymbol{\theta}_2 = (a_A, b_A, \mu_A, \tau_A, \mathbf{a}_G, \mathbf{b}_G, \tau_G, \mu_\nu, \boldsymbol{\theta}_\nu, \tau_\nu, \boldsymbol{\omega}_\nu, \boldsymbol{w}_\nu, \Lambda) \tag{4.13}$$

   This class is composed by of all the orphan nodes in the graph. We only have a vague (or non informative) prior knowledge about the distributions of these orphan nodes. The nodes denoted by a and b represent slopes and intercepts for the conditional distributions of the real amplitude and apparent G-magnitude given the logarithm of the frequency. The nodes denoted by τ represent precisions (the inverse squares of

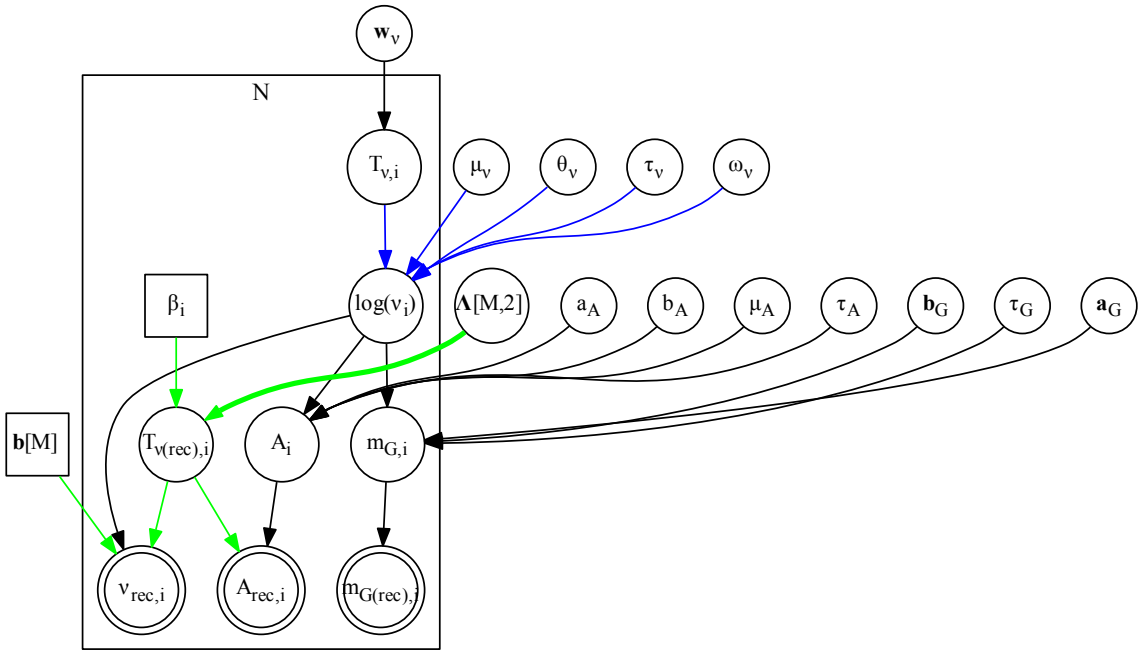


Figure 4.7: Graph structure of our proposed BGM. The arcs depicted in green correspond to a submodel which discriminate (classify) each recovered Frequency and Amplitude according to the ecliptic latitude of the corresponding astronomical source. The rest of arcs correspond to a hierarchical model by means of which the observed values are generated from the real ones. Note the basic structure enclosed in a rectangle which is repeated (replicated) N times (using the *plate* notation) to account the complete set of observations. Fixed parameters are not included in the graph, with the exception of the ecliptic latitude (β_i) and the intercepts vector $\mathbf{b}$ for categories of aliased frequencies. See the text and Table 4.3 for node descriptions.

standard deviations). The nodes denoted by μ represent means. The nodes denoted by Λ represent coefficients of a multinomial logistic regression submodel with the ecliptic latitude β as a predictor. The rest of nodes are associated with the parameterization of the real frequency distribution followed in the work.

4.3.2 Distributions, Parameterizations and Priors

4.3.2.1 Input frequencies

Taking into account that we know the analytical PDF form for $\log(\nu_i)$ (Section 4.2.1, Eq. 4.1), we parametrize the (decimal) logarithm of the input frequencies by means of a mixture

Node	Description	Type of distribution
τ_G	Precision	Gamma NI prior
$\mathbf{a}_G$	Slopes	Gaussian NI prior
$\boldsymbol{b}_G$	Intercepts	Gaussian NI prior
$m_{G,i}$	Input apparent G magnitude	Gaussian
$m_{G,\mathrm{rec},i}$	Recovered apparent G magnitude	Gaussian
μ_A	Mean	Gaussian NI prior
τ_A	Precision	Gamma NI prior
a_A	Slope	Gaussian NI prior
b_A	Intercept	Gaussian NI prior
A_i	Input amplitude	Gaussian
$A_{\mathrm{rec},i}$	Recovered amplitude	Mixture of skewed Cauchy
$\boldsymbol{w}_\nu$	Mixing proportions.	Gamma NI prior
T_{ν_i}	Category of $\log(\nu_i)$	Categorical
μ_ν	Mean	Non informative
$\boldsymbol{\theta}_\nu$	Mean Perturbations	Gaussian NI prior
τ_ν	Precision	Non informative
$\boldsymbol{\omega}_\nu$	Precision Perturbations	Uniform prior
$\log(\nu_i)$	Input frequency $[d^{-1}]$.	Mixture of Gaussian
Λ	Logistic R. coefficients	Student t prior
$T_{\nu_{\mathrm{rec},i}}$	Category of $\nu_{\mathrm{rec},i}$	Categorical
$\nu_{\mathrm{rec},i}$	Recovered frequency	Mixture of Gaussian

Table 4.3: Description of parameters. Meaning of Abbreviations: NI = non informative

of Gaussian distributions, but considering only three components

$$
\begin{aligned}
f\left(\log\left(\nu_i\right) \mid T_{\nu_i}, \mu_\nu, \boldsymbol{\theta}_\nu, \tau_\nu, \boldsymbol{\omega}_\nu\right) = \\
\delta^1_{T_{\nu_i}} \mathsf{N}\left(\mu_\nu, \tau_\nu\right) + \delta^2_{T_{\nu_i}} \mathsf{N}\left(\mu_\nu + \sqrt{\tau_\nu^{-1}}\theta_{\nu 1}, \tau_\nu \omega_{\nu 1}^{-2}\right) + \\
\delta^3_{T_{\nu_i}} \mathsf{N}\left(\mu_\nu + \sqrt{\tau_\nu^{-1}}\theta_{\nu 2} + \sqrt{\tau_\nu^{-1}}\omega_{\nu 1}\theta_{\nu 2}, \tau_\nu \omega_{\nu 1}^{-2}\omega_{\nu 2}^{-2}\right)
\end{aligned}
\tag{4.14}
$$

where:

- The three Kronecker deltas $\delta^j_{T_{\nu_i}}$ indicate the Gaussian component to which $\log\left(\nu_i\right)$ belongs according to the value of the parameter T_{ν_i} distributed as

$$
p\left(T_{\nu_i}\right) = \mathsf{Cat}\left(3, w_{\nu 1}, w_{\nu 2}, w_{\nu 3}\right) \tag{4.15}
$$

  Therefore $\delta^j_{T_{\nu_i}} = 1$ denotes that the categorical variable T_{ν_i} takes the value "j", i.e. $\log\left(\nu_i\right)$ belongs to the j-th component, with probability (mixing proportion) $w_{\nu j}$.

- Parameters μ_ν and τ_ν denote, respectively, the mean (location parameter) and the precision (inverse square of the scale parameter) of the first component of $\log\left(\nu_i\right)$.

- Parameters $(\theta_{\nu 1}, \theta_{\nu 2})$ and $(\omega_{\nu 1}, \omega_{\nu 2})$ denote, respectively, perturbation parameters which affect the mean and the scale parameter of a given component to obtain the mean and scale parameter of the following component (Robert and Mengersen, 1999).

- For parameters $\boldsymbol{w}_\nu$, μ_ν , $\boldsymbol{\theta}_\nu$, τ_ν and $\boldsymbol{\omega}_\nu$ we take the following non informative priors:

$$
\begin{aligned}
p\left(\boldsymbol{w}_\nu\right) &= \mathsf{Dir}\left(1, 1, 1\right) \\
p\left(\mu_\nu\right) &= \mathsf{N}\left(0, 0.001\right) \\
p\left(\theta_{\nu j}\right) &= \mathsf{N}\left(0, 0.01\right) \\
p\left(\tau_\nu\right) &= \mathsf{Gamma}\left(0.001, 0.001\right) \\
p\left(\omega_{\nu j}\right) &= \mathsf{U}\left(0, 1\right)
\end{aligned}
\tag{4.16}
$$

  The use of the three former priors guarantees that the full conditional distribution of the node is available in closed form (see Section 4.4). The Dirichlet prior $p\left(\boldsymbol{w}_\nu\right)$ with parameters $(1, 1, 1)$ is equivalent to a uniform distribution over the constrained

hiperparameter space of T_{ν_i}. The Gaussian priors $p(\mu_\nu)$ and $p(\theta_{\nu j})$ are proper priors which approximate a flat prior over $\mathbb{R}$. The Gamma prior $p(\tau_\nu)$ with the parameters shape and rate both equal to 0.001 is a weakly informative prior which verifies that $\mathrm{E}(\tau_\nu) = 1$ and $\mathrm{Var}(\tau_\nu) = 10^{-3}/10^{-6} = 1000$. Finally, the uniform prior $p(\omega_{\nu j})$ represents our lack of knowledge about the residual variance of a Gaussian component relative to the variance of the preceding component.

4.3.2.2 Input amplitudes

Taking into account that we also know the analytical PDF form for $A_i \mid \log(\nu_i)$ (Eq. 4.2), we parametrize the input amplitude as

$$f(A_i \mid \log(\nu_i), a_A, b_A, \mu_A, \tau_A) = \mathbf{1}_{\{\log(\nu_i)<-1\}}\mathsf{N}(a_A \cdot \log(\nu_i) + b_A, \tau_A) + \mathbf{1}_{\{\log(\nu_i)>-1\}}\mathsf{N}(\mu_A, \tau_A) \tag{4.17}$$

where:

- $\mathbf{1}_S$ denotes the indicator function of the subset S, i.e. $\mathbf{1}_S(x) = \begin{cases} 1 & x \in S \\ 0 & x \notin S \end{cases}$.
- Parameters a_A and b_A are, respectively, the slope and the intercept of the regression line of A on $\log(\nu)$ when $\log(\nu) < -1$.
- Parameter μ_A denotes the mean of the amplitude when $\log(\nu) > -1$.
- Parameter τ_A denotes the precision, which we take equal for both distributions (conditional and unconditional) .
- For parameters a_A , b_A , λ_A and τ_A we take the following non informative priors:

$$\begin{aligned} p(a_A) &= \mathsf{N}(0, 0.001) \\ p(b_A) &= \mathsf{N}(0, 0.001) \\ p(\mu_A) &= \mathsf{N}(0, 0.01) \\ p(\tau_A) &= \mathsf{Gamma}(0.001, 0.001) \end{aligned} \tag{4.18}$$

For all these priors the full conditional distribution of the node is available in closed form. The Gaussian priors are proper priors which approximate a flat prior over $\mathbb{R}$. The Gamma prior $p(\tau_A)$ has been selected taking into account the same considerations that for $p(\tau_\nu)$.

4.3.2.3 Apparent magnitudes in the G Band

Taking into account Eq. 4.7 but discarding the distance r we parametrize this node as

$$f(m_{G,i} \mid \log(\nu_i), a_{G1}, b_{G1}, a_{G2}, b_{G2}, \tau_G) = \tag{4.19}$$
$$\mathbf{1}_{\{\log(\nu_i)<-1\}}\mathsf{N}(a_{G1}\cdot\log(\nu_i)+b_{G1},\tau_G)+\mathbf{1}_{\{\log(\nu_i)>-1\}}\mathsf{N}(a_{G2}\cdot\log(\nu_i)+b_{G2},\tau_G)$$

with the prior distributions

$$\begin{aligned} p(a_{G1}) &= \mathsf{N}(0, 0.001) \\ p(a_{G2}) &= \mathsf{N}(0, 0.001) \\ p(b_{G1}) &= \mathsf{N}(0, 0.001) \\ p(b_{G2}) &= \mathsf{N}(0, 0.001) \\ p(\tau_G) &= \mathsf{Gamma}(0.001, 0.001) \end{aligned} \tag{4.20}$$

For all these priors the node's full conditional distribution is available in closed form.

4.3.2.4 Recovered frequencies

Taking into account the experimental results of sub-section 4.2.4.1 we parametrize each recovered frequency as a mixture of Gaussian distributions where the mean of each component represent the straight line (*locus*) in which the input frequency has been recovered, i.e. the identity locus or some locus of spurious (aliased) frequencies. For each recovered frequency we assign a categorical variable for each existing locus. In the present paragraph we present the parametrization $\nu_{\mathrm{rec},i} \mid \log(\nu_i), T_{\nu_{\mathrm{rec},i}}$ (red arcs in Fig. 4.7) leaving for the next epigraph the parameterization for the categories $T_{\nu_{\mathrm{rec},i}}$ whose parameters correspond

to the mixing proportions of the mixture. The former parametrization is done as

$$f\left(\nu_{\text{rec},i} \mid \log\left(\nu_i\right), T_{\nu_{\text{rec},i}}\right) = \tag{4.21}$$
$$\delta^1_{T_{\nu_{\text{rec},i}}} \mathsf{N}\left(10^{\log(\nu_i)}, \tau_{\nu_{\text{rec}}}\right) + \sum_{j=2}^{M} \delta^j_{T_{\nu_{\text{rec},i}}} \mathsf{N}\left((-1)^{j-1}\, 10^{\log(\nu_i)} + b_j, \tau_{\nu_{\text{rec}}}\right)$$

where:

- $\mathsf{N}\left(a_j 10^{\log(\nu_i)} + b_j, \tau_{\nu_{\text{rec}}}\right)$, with $a_j = (-1)^{j-1}$ for $j \in \{1, ..., M\}$, is a Gaussian component corresponding to the conditional distribution $\nu_{\text{rec},i} \mid \nu_i$ of the recovered i-th frequency *given* the i th real frequency *when* the frequency is recovered over the j th *locus* $\nu_{\text{rec}} = a_j \nu + b_j$.

- We assume a fixed number of components: the identity *locus* and $M-1$ *loci* of spurious components with predefined slopes and intercepts. The sub index $j = 1$ correspond to the identity locus (the first term of the second member of Eq. 4.21) with slope $a_1 = 1$ and intercept $b_1 = 0$. Otherwise the spurious components, $j \in \{2, ..., M\}$, are defined with alternating slopes -1 o +1 depending on the exponent in $a_j = (-1)^{j-1}$ for accounting the symmetry of the power spectrum. Each pair of symmetric spurious components shares the same intercept, i.e. $b_{j+1} = b_j$ when $j\text{mod}2 = 0$, which is modeled as a constant.

- The Kronecker deltas $\delta^j_{T_{\nu_{\text{rec},i}}}$ indicate the Gaussian component to which $\nu_{\text{rec},i}$ belongs according to the value of the parameter (categorical variable) $T_{\nu_{\text{rec},i}}$. Therefore $\delta^j_{T_{\nu_{\text{rec},i}}} = 1$ denotes that the categorical variable $T_{\nu_{\text{rec},i}}$ takes the value "j", i.e. that $\nu_{\text{rec},i}$ belongs to the j-th component with some probability (mixing proportion) $\pi_{ij} = w^j_{\nu_{\text{rec},i}}$ which is different (in principle) for each recovered frequency (see the next paragraph below).

- The precision $\tau_{\nu_{\text{rec}}} = 10000$ is assumed to be constant (and equal) for all components.

4.3.2.5 Categories of recovered frequencies

A determinate value $j \in \{1, .., M\}$ of the categorical node $T_{\nu_{\text{rec},i}}$ indicates that the the i-th frequency has been recovered in the j-th locus (Gaussian component), which is done with a certain probability π_{ij}. Taking into account the experimental results of sub-section 4.2.4.2

we make depend π_{ij} on the ecliptic latitude β_i and parametrize this dependence by means of a multinomial logistic regression submodel using the *softmax* function as transfer function. Thus, we model the conditional distribution for $T_{\nu_{\mathrm{rec},i}}$ as

$$p\left(T_{\nu_{\mathrm{rec},i}} \mid \{\boldsymbol{\lambda}_j\}_{j=2}^{M}\right) = \mathsf{Cat}\left(M, \{\pi_{ij}\left(\beta_i', \boldsymbol{\lambda}_j\right)\}_{j=1}^{M}\right) \tag{4.22}$$

with

$$\pi_{ij}\left(\beta_i', \boldsymbol{\lambda}_j\right) = \frac{e^{\boldsymbol{\lambda}_j^T \cdot \left(1, \beta_i'\right)}}{\sum_{l=1}^{M} e^{\boldsymbol{\lambda}_l^T \cdot \left(1, \beta_i'\right)}} \tag{4.23}$$

where

- We rescale the predictor of the logistic regression submodel by substracting the mean and dividing by two times the standard deviation, i.e.$\beta_i' = \frac{\beta_i - \overline{\beta}}{2 \cdot \mathrm{sd}(\beta)}$ where β_i are the ecliptic latitude values. This guaranties that the mean and the standard deviation of β_i' are respectively 0 and 0.5.

- The hyperparameter vectors $\boldsymbol{\lambda}_j = (\lambda_{0j}, \lambda_{1j})$ for $j \in \{2, ..., M\}$ contain the coefficients of the logistic regression submodel. Determining the distributions of these coefficients is one of the goals of the inference. We assign them the weakly informative priors $p\left(\lambda_{kj}\right) = \mathsf{t}\left(0, {}^{1}/_{2.5^2}, 7\right)$, $k \in \{0, 1\}$. This election provide a minimal prior information to constrain the range of coefficients λ_{kj} once the covariate β_i has been rescaled (Gelman et al., 2008). This approximation is used to enhance the convergence rate of our model.

- We assume $\boldsymbol{\lambda}_1 = \mathbf{0}$, which implies considering the identity *locus* as the *reference category*.

It is customary to express Eq. 4.23 in the alternative form

$$\ln\left(\frac{\pi_{ij}}{\pi_{i1}}\right) = \boldsymbol{\lambda}_j \cdot (1, \beta_i) = \lambda_{0j} + \lambda_{1j}\beta_i \tag{4.24}$$

to better understand the role of the identity locus in the model. The logarithms to the left of the expression are so-called the *logits*. They express the relation (in logarithmic scale) between the probability of recovery the i-th frequency in the j-th spurious component and the probability of recovery that frequency in the identity locus. So, the *logit* variation is

determined by the variation of the predictors modulated by the model coefficients (Gelman et al. (2004), Chapter 16, Section 7, p. 430).

4.3.2.6 Recovered amplitudes

Taking into account the experimental results of sub-section 4.2.5.2, we model the conditional distribution for the recovered amplitude $A_{\mathrm{rec},i}$ by means of a mixture of two skewed Student t distributions (Azzalini and Genton, 2008) with location parameters $\xi_1 = A_i$ and $\xi_j = 0.749 \cdot A_i, \forall j = 2, .., M$ and scale ω, shape α and degrees of freedom ν parameters taking as constants as follows

$$f\left(A_{\mathrm{rec},i} \mid A_i, T_{\nu_{\mathrm{rec},i}}\right) = \tag{4.25}$$
$$\delta^1_{T_{\nu_{\mathrm{rec},i}}} \mathsf{ST}\left(A_i, 0.020, -2.395, 1\right) + \sum_{j=2}^{M} \delta^j_{T_{\nu_{\mathrm{rec},i}}} \mathsf{ST}\left(0.749 \cdot A_i, 0.0266, 0, 1\right)$$

Note that priors included here are not quite realistic and really slope, intercepts, and the rest of fixed parameters of Student t should be included as random parameters in future versions of our model.

4.3.2.7 Recovered apparent magnitudes

We parameterize the distribution of the i-th recovered apparent G magnitude by means of a Gaussian distribution with mean $m_{G,i}$ and standard deviation $\sigma_{G(rec),i} = f\left(G_i\right)$ computed deterministically using Equation 4.9 (Jordi et al., 2007)[8]:

$$f\left(m_{G_{rec},i} \mid m_{G,i}, r_i\right) = \mathsf{N}\left(m_{G,i}, \sigma_{G(rec),i}\right) \tag{4.26}$$

4.3.3 Factorization

In this subsection we present the factorization of the joint probability density function associated to the graph of Fig. 4.7. According to the classification of the nodes described in Section 4.3.1, now we can factorize the joint PDF in the following three factors:

[8] In the final implementation in BUGS presented in the thesis we have temporally deactivated this functionality and assumed an homoscedasticity hypothesis for errors.

1. The conditional distribution of the data given their parents[9]

$$p\left(\mathcal{D} \mid \boldsymbol{\theta}_1\right) = \prod_{i=1}^{N} f_1\left(\nu_{\mathrm{rec},i} \mid \log\left(\nu_i\right), T_{\nu_{\mathrm{rec},i}}\right) \cdot \tag{4.27}$$
$$f_2\left(A_{\mathrm{rec},i} \mid A_i, T_{\nu_{\mathrm{rec},i}}\right) \cdot f_3\left(m_{G_{\mathrm{rec}},i} \mid m_{G,i}\right)$$

   This is the *likelihood of the data under the model* which, fixed the data, is a function $\mathcal{L}\left(\boldsymbol{\theta}_1\right)$ of the parameters.

2. The conditional distribution of the first level of random parameters given the parameters of the top level

$$p\left(\boldsymbol{\theta}_1 \mid \boldsymbol{\theta}_2\right) = \prod_{i=1}^{N} g_1\left(T_{\nu_{rec,i}} \mid \{\boldsymbol{\lambda}_j\}_{j=2}^{M}\right) \cdot \tag{4.28}$$
$$g_2\left(A_i \mid \log\left(\nu_i\right), a_A, b_A, \mu_A, \tau_A\right) \cdot g_3\left(m_{G,i} \mid \log\left(\nu_i\right), \mathbf{a}_G, \mathbf{b}_G, \tau_G\right) \cdot$$
$$g_4\left(\log\left(\nu_i\right) \mid T_{\nu_i}, \lambda_\nu, \boldsymbol{\theta}_\nu, \tau_\nu, \boldsymbol{\omega}_\upsilon\right) \cdot g_5\left(T_{\nu_i} \mid \boldsymbol{w}_\nu\right)$$

   which corresponds to a first level of prior distributions.

3. The distribution of the top level parameters (hyperparameters, orphan nodes)

$$p\left(\boldsymbol{\theta}_2\right) = h_1\left(a_A\right) \cdot h_2\left(b_A\right) \cdot h_3\left(\mu_A\right) \cdot h_4\left(\tau_A\right) \cdot h_5\left(\mathbf{a}_G\right) \cdot h_6\left(\mathbf{b}_G\right) \cdot \tag{4.29}$$
$$h_7\left(\tau_G\right) \cdot h_8\left(\boldsymbol{w}_\nu\right) \cdot h_9\left(\mu_\nu\right) \cdot h_{10}\left(\boldsymbol{\theta}_\nu\right) \cdot h_{11}\left(\tau_\nu\right) \cdot h_{12}\left(\boldsymbol{\omega}_\nu\right) \cdot h_{13}\left(\Lambda\right)$$

   This factor corresponds to a top level of prior distributions (hyperpriors). Parameters for hyperpriors are fixed and not explicitly included in the graph.

Finally the complete PDF factorization is given by

$$p\left(\boldsymbol{\theta}, \mathcal{D}\right) = p\left(\mathcal{D} \mid \boldsymbol{\theta}\right) \cdot p\left(\boldsymbol{\theta}\right) = \tag{4.30}$$
$$p\left(\mathcal{D} \mid \boldsymbol{\theta}_1\right) \cdot p\left(\boldsymbol{\theta}_1 \mid \boldsymbol{\theta}_2\right) \cdot p\left(\boldsymbol{\theta}_2\right)$$

[9] Note that although T_{ν_i} is not included in the set of parents, the three evidential nodes $\mathcal{D}_i = (\nu_{\mathrm{rec},i}, A_{\mathrm{rec},i}, m_{G_{\mathrm{rec}},i})$ are conditionally independent of T_{ν_i} given their parents. Therefore it holds that $p\left(\mathcal{D}_i \mid \mathrm{pa}\left(\mathcal{D}_i\right)\right) = p\left(\mathcal{D}_i \mid \boldsymbol{\theta}_1\right)$.

Node	Sampler	Node	Sampler
τ_G	Conjugate gamma	$\boldsymbol{w}_\nu$	Conjugate Dirichlet
$\mathbf{a}_G$	Conjugate Gaussian	T_{ν_i}	Discrete slice
$\boldsymbol{b}_G$	Conjugate Gaussian	μ_ν	Conjugate Gaussian
$m_{G,i}$	Conjugate Gaussian	$\boldsymbol{\theta}_\nu$	Conjugate Gaussian
μ_A	Conjugate Gaussian	τ_ν	Slice
τ_A	Conjugate gamma	$\boldsymbol{\omega}_\nu$	Slice
a_A	Conjugate Gaussian	$\log(\nu_i)$	Metropolis 1D
b_A	Conjugate Gaussian	Λ	Metropolis 1D
A_i	Metropolis 1D	$T_{\nu_{\text{rec},i}}$	Categorical

Table 4.4: Samplers used to estimate the full conditional distribution for parameters of the BGM proposed in this thesis. With the exception of $\boldsymbol{w}_\nu$, for parameters which are random vectors the sampler is applied independently to each component.

4.4 Inference Algorithm

Given the model proposed in Section 4.3 we have to sample the joint posterior distribution for its $22 + 5N$ parameters, namely

$$\pi^*(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta} \mid \mathcal{D}) \propto \mathcal{L}(\boldsymbol{\theta}_1) \cdot p(\boldsymbol{\theta}_1 \mid \boldsymbol{\theta}_2) \cdot p(\boldsymbol{\theta}_2) \tag{4.31}$$

, where N is the sample size of the data $\mathcal{D}$, and then marginalize over the nuisance parameters to obtain the posterior distribution for the parameters of interest. Our focus mainly are in the hyperparameters of real frequencies, amplitudes and apparent G-magnitudes inside the vector $\boldsymbol{\theta}_2$ of the top level of random parameters hierarchy. Therefore we are interested in the marginal *a posteriori* $\pi^*(\boldsymbol{\theta}_2)$. The marginalization to obtain samples form this latter distribution can be accomplished by the general MCMC procedure depicted in Section 3.4.2. This procedure retains only the values of $\boldsymbol{\theta}_2$ once a sample for the joint posterior has been obtained and discards the rest[10].

As we saw in Section 3.4.5, the joint posterior distribution in Eq. 4.31 can be efficiently sampled by means of a Gibbs sampling scheme. In each iteration, this algorithm traverses the graph in topological order and sample the full conditional distribution of each node, given the data and the most recent values of the rest of nodes, using its Markov Blanket. In BUGS we only have, in principle, to declare the graph and let the software to determine the best sampling algorithm for each node[11]. We include in Table 4.4 the complete list of

[10] In the OpenBUGS graphical interface this is easily accomplished by the option `Inference->Samples->Sample Monitor Tool`.

[11] Note that in BUGS it is not necessary to implement the inference mechanism because it is incorporated

sampling algorithm used to compute the full conditional distribution for parameters of the proposed BGM. The list corresponds to the default samplers taken by OpenBUGS, with the exception of the categorical sampler for $T_{\nu_{\text{rec},i}}$ forced by us. We exemplify only the full conditional distribution sampling in two scenarios, one for node τ_G and the other for the set of nodes $\{\log(\nu_i)\}$. For that, we assume the following topological ordering for the graph depicted in Fig. 4.7:

$$\begin{aligned}\boldsymbol{\omega}_\nu \succ \tau_\nu \succ \tau_G \succ \tau_A \succ \boldsymbol{\theta}_\nu \succ \mu_A \succ \mu_\nu \succ \boldsymbol{b}_G \succ b_A \succ \boldsymbol{a}_G \succ a_A \succ \boldsymbol{w}_\nu \\ \underset{i}{\succ} T_{\nu_i} \succ \Lambda \underset{i}{\succ} T_{\nu_{rec,i}} \underset{i}{\succ} \log(\nu_i) \underset{i}{\succ} m_{G,i} \underset{i}{\succ} A_i\end{aligned} \tag{4.32}$$

and assume too that we are it the t-th iteration of the sampling algorithm. Then, in the third step of this iteration we have to sample $\tau_G^{(t)}$ according to

$$\tau_G^{(t)} \sim \pi(\tau_G \mid \text{bl}(\tau_G)) \propto h_7(\tau_G) \cdot \prod_{i=1}^{N} g_3\left(m_{G,i}^{(t-1)} \mid \log(\nu_i)^{(t-1)}, \boldsymbol{a}_G^{(t-1)}, \boldsymbol{b}_G^{(t-1)}, \tau_G\right)$$

, which can be done from a closed form for the posterior PDF, namely a conjugate Gamma. Otherwise, for the same t-th iteration, the 16-th step is developed by cycling across the set $\{\log(\nu_i)\}$ and sampling each node by

$$\begin{gathered}\log\left(\nu_i^{(t)}\right) \sim \pi(\log(\nu_i) \mid \text{bl}(\log(\nu_i))) \propto \\ g_4\left(\log(\nu_i) \mid T_{\nu_i}^{(t)}, \mu_\nu^{(t)}, \boldsymbol{\theta}_\nu^{(t)}, \tau_\nu^{(t)}, \boldsymbol{\omega}_\nu^{(t)}\right) \cdot g_2\left(A_i \mid \log(\nu_i), a_A^{(t)}, b_A^{(t)}, \mu_A^{(t)}, \tau_A^{(t)}\right) \cdot \\ g_3\left(m_{G,i}^{(t-1)} \mid \log(\nu_i), \mathbf{a}_G^{(t)}, \mathbf{b}_G^{(t)}, \tau_G^{(t)}\right) \cdot f_1\left(\nu_{\text{rec},i} \mid \log(\nu_i), T_{\nu_{\text{rec},i}}^{(t)}\right)\end{gathered}$$

, which can be done by the 1D Metripolis Hasting algorithm given that a closed form for the posterior PDF is unknown.

4.5 Implementation

4.25, 4.3

Given that BUGS is a declarative language, implementation of equations in Section 4.3.2 should be, in principle, quite straightforward. Nevertheless, we have been faced to

in the software tool. Nevertheless it is important to deepen some degree in the third stage of the methodology proposed in Section 4.1 taking in mind future implementations on R or Java.

some problems. The main difficulty has been to code the mixtures of distributions which constitute the cornerstone of our work and, in particular, the sampling mixtures of distributions which are not included in the list of standard distributions in the open-source version of the language OpenBUGS[12] v. 3.2.2 employed in the thesis. To model the likelihood for such a sampling distribution we have used the so-called Serguei Smirnov's zeroes trick (Smirnov, 2001). The trick takes into account that the likelihood of observing a zero for an exponential distribution with rate parameter λ is given by $f(0) = \lambda$. So, we can introduce an array of zeroes as dummy data and make λ to be the likelihood of the distribution by means of which we generate our real data. For example, by Equation 4.25 we model the conditional distribution of the recovered amplitude $A_{\text{rec},i} \mid A_i, T_{\nu_{\text{rec},i}}$ as a mixture of three skewed Cauchy distributions whose mixing proportions are given by the probability of each locus $T^j_{\nu_{\text{rec},i}}$ of frequency. According to (Azzalini and Genton, 2008, expr. (4), page 109), the PDF for a skewed Cauchy can be derived from the univariate standard Student t PDF and its distribution function[13], which have allow us to develop the following code segment:

```
1  # #Serguei Smirnov's zeroes trick
2  dummy[i] <- 0
3  dummy[i] ~ dexp(likelihood.recA[i])
4  norm.recA[i]<-(recA[i]-A[i])/scale.recA[i]
5  # CDF for a skewed Cauchy distribution
6  arg.cdf[i]<-shape.recA[i]*norm.recA[i]*sqrt(2/(1+pow(norm.recA[i],2)))
7  # Likelihood for a skewed Cauchy distribution
8  likelihood.recA[i]<-(2/scale.recA[i])*1/(pi*(1+pow(norm.recA[i],2)))
       *0.5*(1+arg.cdf[i]/sqrt(2+pow(arg.cdf[i],2)))
9  # scale and shape parameters for the i-th recovered amplitude
10 scale.recA[i] <- cscale.recA[T.recNu[i]]
```

[12] Available in http://www.openbugs.net .
[13] Taking the degrees of fredom to one.

Chapter 5

Model Evaluation

This chapter is entirely devoted to develop the fourth stage of the general methodology, for constructing our BGM, proposed in Section 1.3, namely, the model evaluation. It is structured in six sections as follows. Section 5.1 summarizes the methodological stages we follow. Section 5.2 describes the preparation of a sample of observed variables to train the model. Section 5.3 is devoted to the model training. In Section 5.4 we analyze the model convergence. Section 5.5 is devoted to the estimation and analysis of the posterior distributions for the parameters of interest. Finally, in Section 5.6 we compare the parameters inferred by our model with the real ones.

5.1 Methodology

We summarize our methodology in the following five stages:

1. **Preparation of the Training Set**. We prepare a training set from the complete sample of observed variables bearing in mind that we want to evaluate the capacity of our model to infer the real distributions of frequencies and amplitudes in front of an extreme scenario of systematic biases in the recovered data.

2. **Model Training**. We train[1] the model with the OpenBUGS MCMC engine in two

[1]Some remarks about terminology. The use of the term "model training" must be understood here in the context of the Bayesian statistical learning where we aim to learn the posterior values for a number of parameters in the model given the evidence. The evaluation of the model in the thesis is done by comparing those posterior values with the real ones which in our case are available. In future implementations, by using as data the real values provided by Gaia, evaluation mechanism should be changed to the use of posterior predictive distributions.

stages, the former for convergence analysis and the latter to obtain a set of samples for parameters of interest.

3. **Convergence Diagnosis**. We use two criteria to evaluate the convergence: the autocorrelation function (ACR) and the corrected GR statistic.

4. **Estimation of Posterior Distributions**. We analyze the samples and estimated densities for posterior distributions of parameters of interest.

5. **Comparisons with Real Parameters**. We compute summary statistics for estimated posteriors distributions, construct infered PDFs for input distributions and compare them with real input distributions whenever possible.

5.2 Preparation of the Training Set

We have prepared the training set, i.e. a subsample $\mathcal{T} \subsetneq \mathcal{D}$ of recovered parameters, taking into account the following strategies:

1. **To use a reduced number *M* of aliased *loci* components**. First, if we include in the model several *loci* from the same substructure, the model does not work well. This happens, in particular, for the substructure around the identity line. Second, the size of *internal* data structures generated when the model is compiled in OpenBUGS grows considerably when we increase the number M of spurious components and the software tends to be unstable[2].

2. **To increase the proportions of the aliased *loci***. The proportions of most *loci* of aliased frequencies, even when the complete database (36645 instances) is considered, are negligible (see Tab. 4.1). The model is able to infer the real distributions of frequencies, amplitudes and apparent G-magnitudes if we maintain these proportions in the training set[3]. Nevertheless, we are interested in evaluating the effectiveness of our model to infer those real distributions in front of an extreme scenario of systematic biases in the recovered data. This alternative scenario could occur for other stellar

[2]

(a) At least in the platform on which we have installed the software (see Section 5.3 bellow)

[3]We have confirmed this fact by performing additional experiments not included in this work.

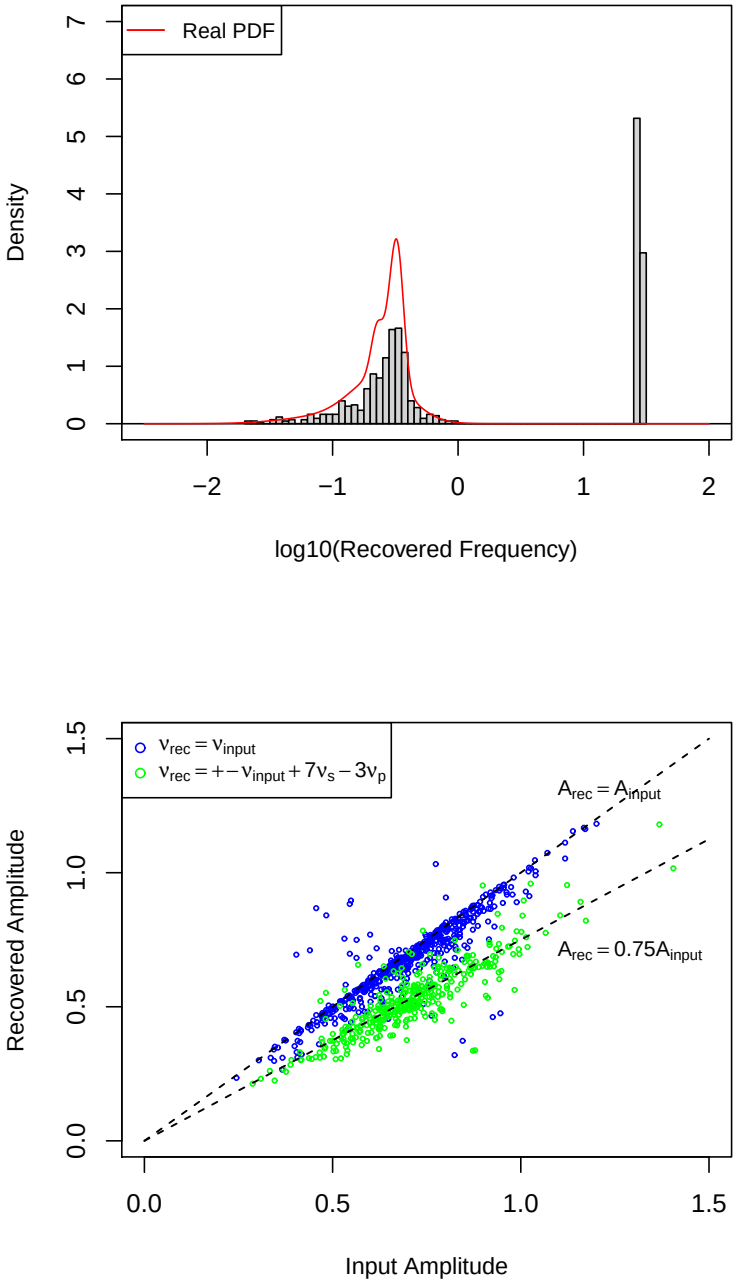


Figure 5.1: Biases in the Training Set.

populations to which our model would be applicable, although that is not the case for the analyzed Cepheids population.

3. **To employ a relatively small sample size N**. First, if we want to increase the proportions of aliased loci (2nd strategy) in the sample we must decrease the number of samples in the identity line. Second, the size of data structures generated when the model is compiled in OpenBUGS also grows considerably when we increase the sample size.

Based on the considerations of the previous paragraphs we have constructed a dataset $\mathcal{T} = \{(A_{\text{rec},i}, \nu_{\text{rec},i}, m_{G,\text{rec},i})\}_1^{854} \subsetneq \mathcal{D}$ composed of 500 randomly (without replacement) selected instances from the *locus* $\nu_{\text{rec}} = \nu_{\text{in}}$ and all instances (354) from the *locus* $\nu_{\text{rec}} = \pm\nu_{\text{in}} + 7\nu_S - 3\nu_p$. Figure 5.1 shows the systematic biases for the empirical frequency distribution (histogram) vs the real one (its PDF) and for the empirical conditional distributions of the recovered amplitude given the input amplitude for the three locus, identity and $\nu_{\text{rec}} = \pm\nu_{\text{in}} + 7\nu_S - 3\nu_p$, whose observed parameters are included in the training set.

5.3 Model Training

We have trained the model from $\mathcal{T}$ using the OpenBUGS MCMC engine configuring the samplers as was indicated in Table 4.4. We have divided the training in two stages and generated three Markov chains (more properly realizations) in each, with a total of 30000 iterations. We have used the first stage, consisting of 20000 iterations, as a *burn-in* phase, being the corresponding realizations discarded after used for convergence evaluation. Therefore, we obtain 10000 samples from the second stage of each realization (30000 in total). We will assume that these samples were drawn from the posterior distribution of the parameters of interest[4].

We have executed OpenBUGS v. 3.2.2 over a Windows 7 OS in an Intel Core i7 machine at 2.67 GHz with 6.00 Gb of RAM employing a single core[5]. The total running time of the simulation process inverted by our desktop computer has been of 68.7 min. with a total of

[4]We monitor the mixing proportions, means and standard deviations of each Gaussian component of the logarithm of the frequency distribution $\log(\nu)$. Obtaining these parameters from those in Equation 4.14 by deterministic relationships is straightforward.

[5]Currently OpenBUGS for Windows does not allow to be directly executed simultaneously in several cores preventing a way to parallelize MCMC chains.

N	Chains	Mem. (Mb)	Time (sec.)
1354	1	20.989	38
1354	3	21.399	115
2708	1	46.452	79
2708	3	46.867	241
5416	1	76.439	210
5416	3	74.788	618

Table 5.1: Empirical study of scalability for different sample size N and number of simulated chains after the 500 first iterations.

20.71 Mb of allocated memory for data structures generated by OpenBUGS once the model has been compiled.

Although a theoretical analysis of scalability of the proposed solution is beyond to the scope of this thesis, we can not fail to make an empirical study, however modest, which is included in this Section given its brevity. Table 5.1 shows the allocated memory just after model compilation and the consumed time after 500 first iterations for three different size samples N generating one and three chains in each. From the table becomes apparent that the temporal complexity order is linear both in N and the number of simulated chains.

5.4 Convergence Diagnosis

To evaluate the convergence within and between the three chains we respectively use the sample autocorrelation function (ACR) and the corrected GR statistic (see Section 3.4.6) applied both to the first 20000 iterations of the algorithm. Both analysis are done with the aid of the CODA package once chains has been imported in R , after being exported by OpenBUGS. For convergence within chains we compute and plot the ACR up to 200 lags. For convergence diagnosis between chains we compute and plot the evolution of the shrink factor (corrected GR statistic) and do the same with the upper bound of a credible interval (at 95%) for it. Given that we are mainly interested in the posterior distributions for parameters of input frequencies, amplitudes and apparent G-magnitudes and and for the sake of conciseness we only present the graphs for these eighteen parameters.

The results of the analysis are depicted in Figures 5.2 to 5.5. Since the ACR function should decrease to zero as the lag increases and the upper bound for corrected scale reduction factor (CSRF) should approach unity if the chain is reaching its stationary distribution, we conclude that the worst scenario (high autocorrelation) is depicted by chains for parameters

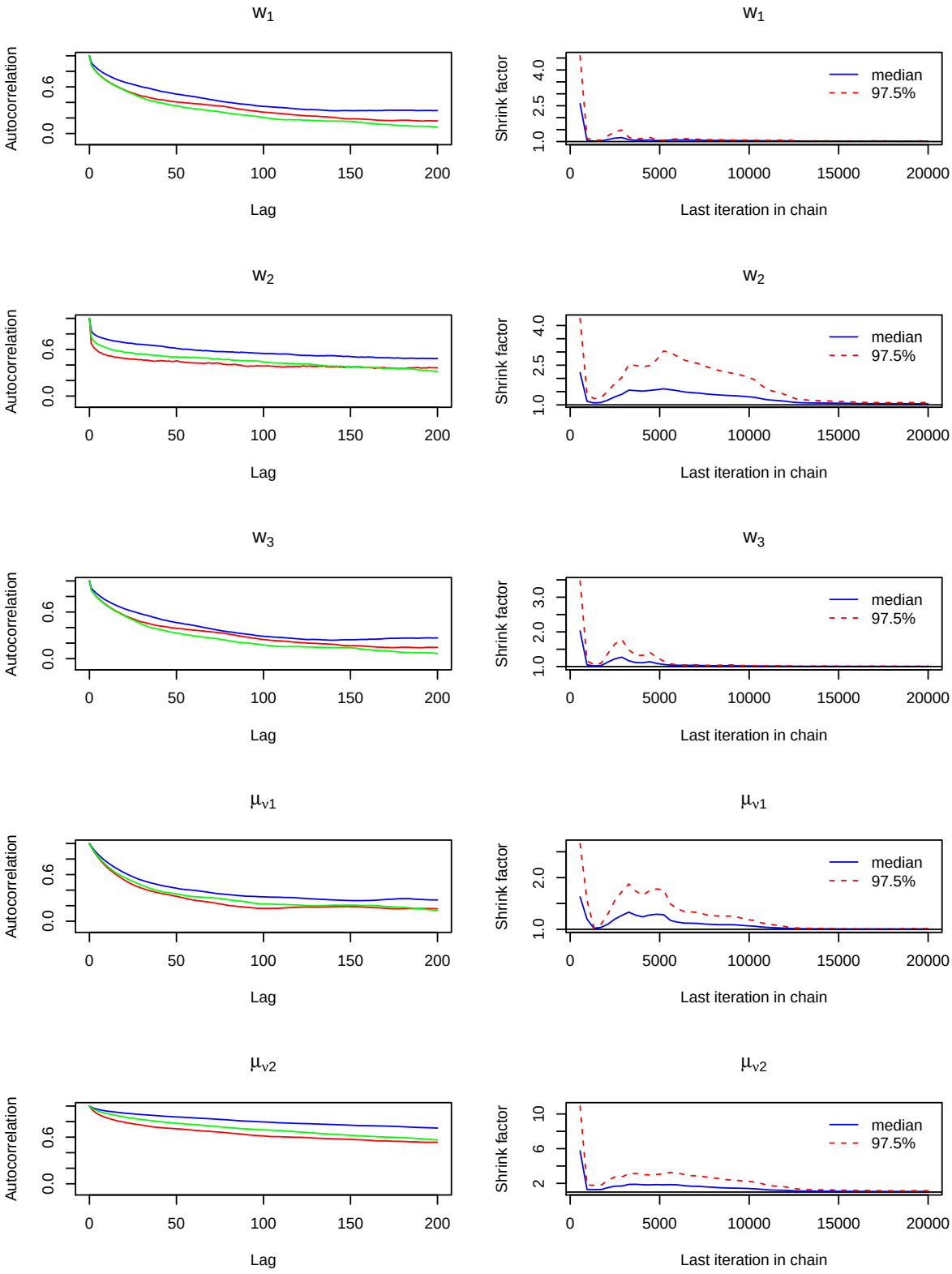


Figure 5.2: Autocorrelation plots (left) and posterior distributions (right) for parameters of $\log(\nu)$ (i).

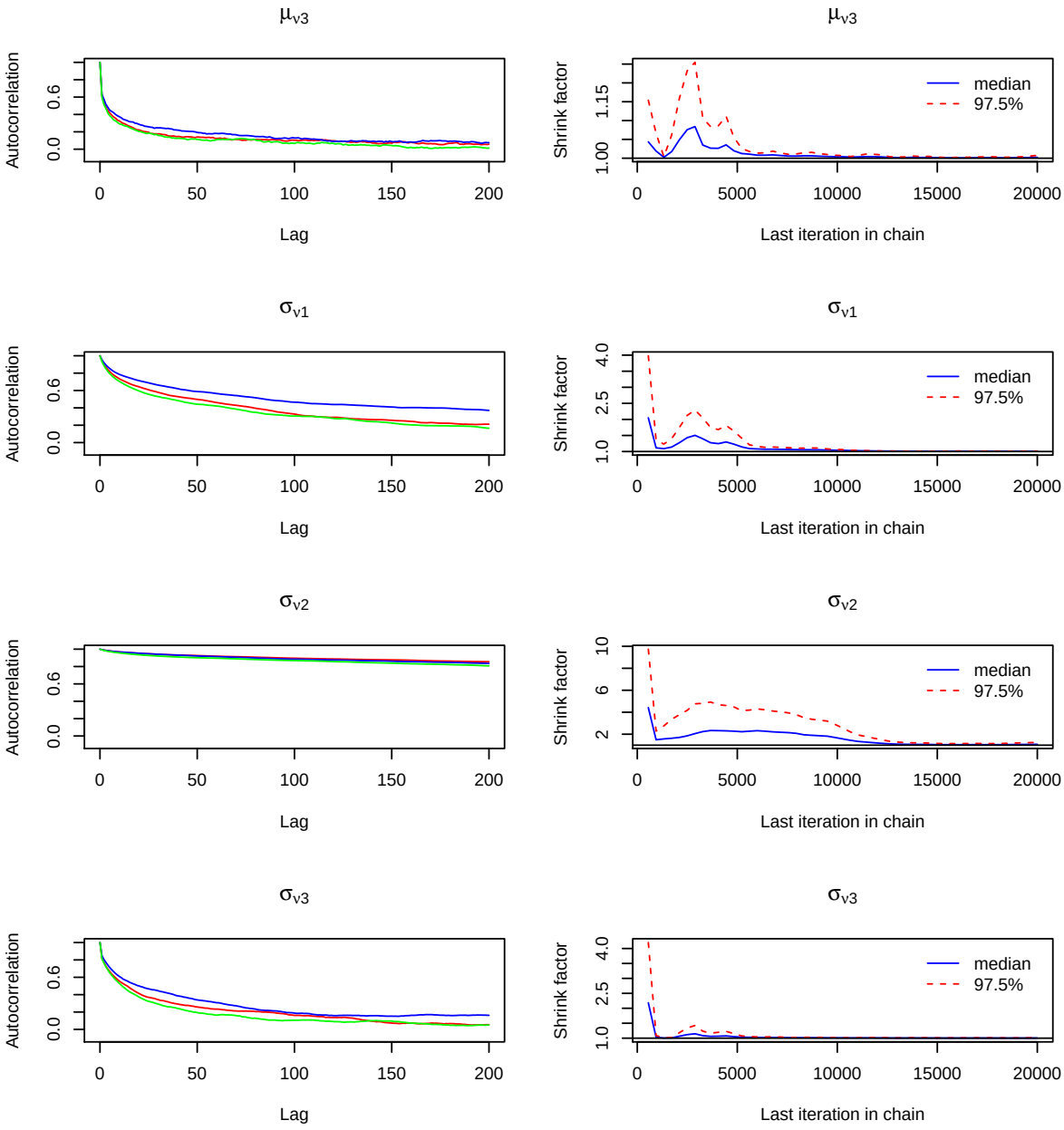


Figure 5.3: Autocorrelation plots (left) and posterior distributions (right) for parameters of $\log(\nu)$ (ii).

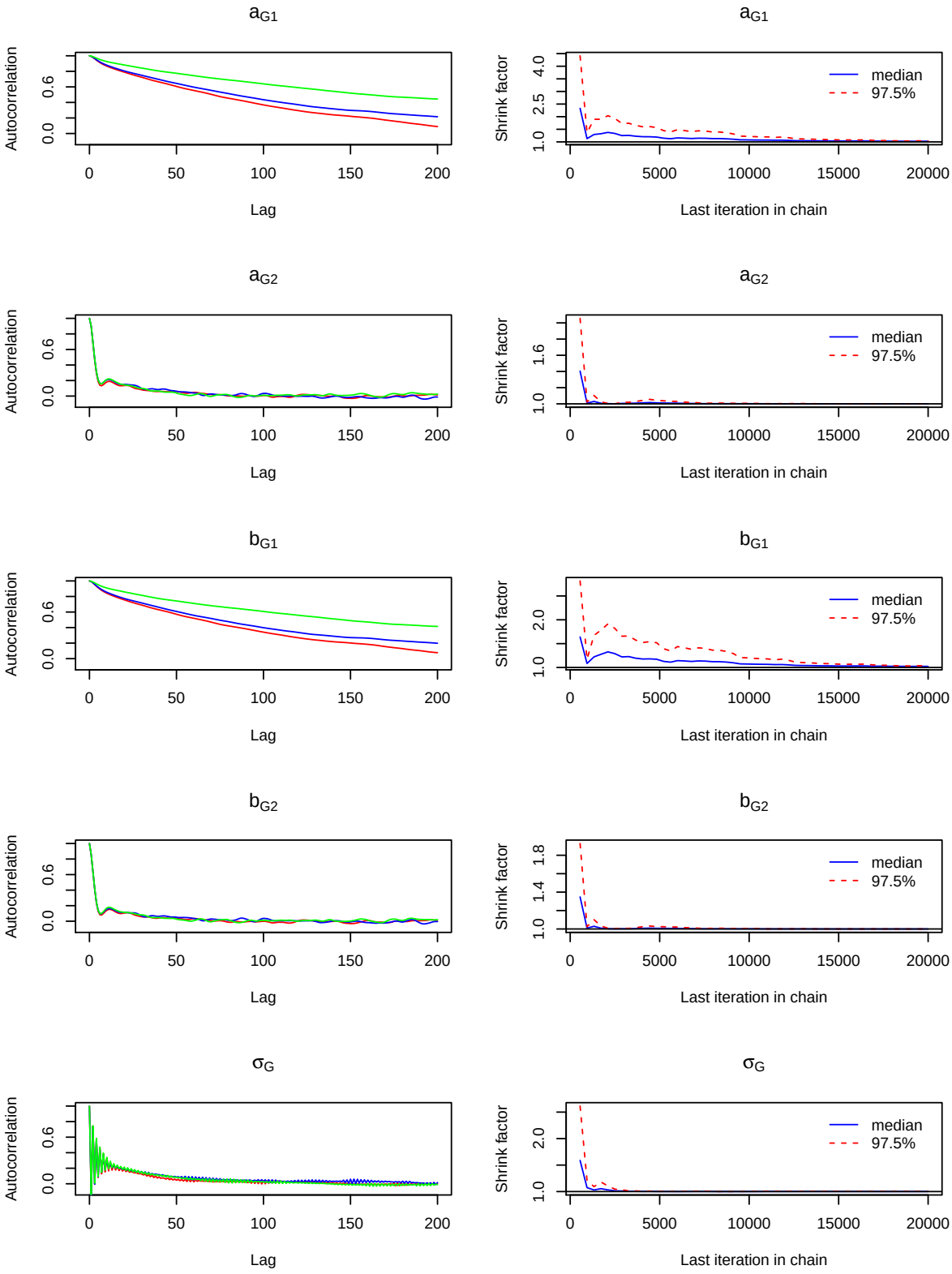


Figure 5.4: Autocorrelation plots (left) and posterior distributions (right) for parameters of apparent G magnitude.

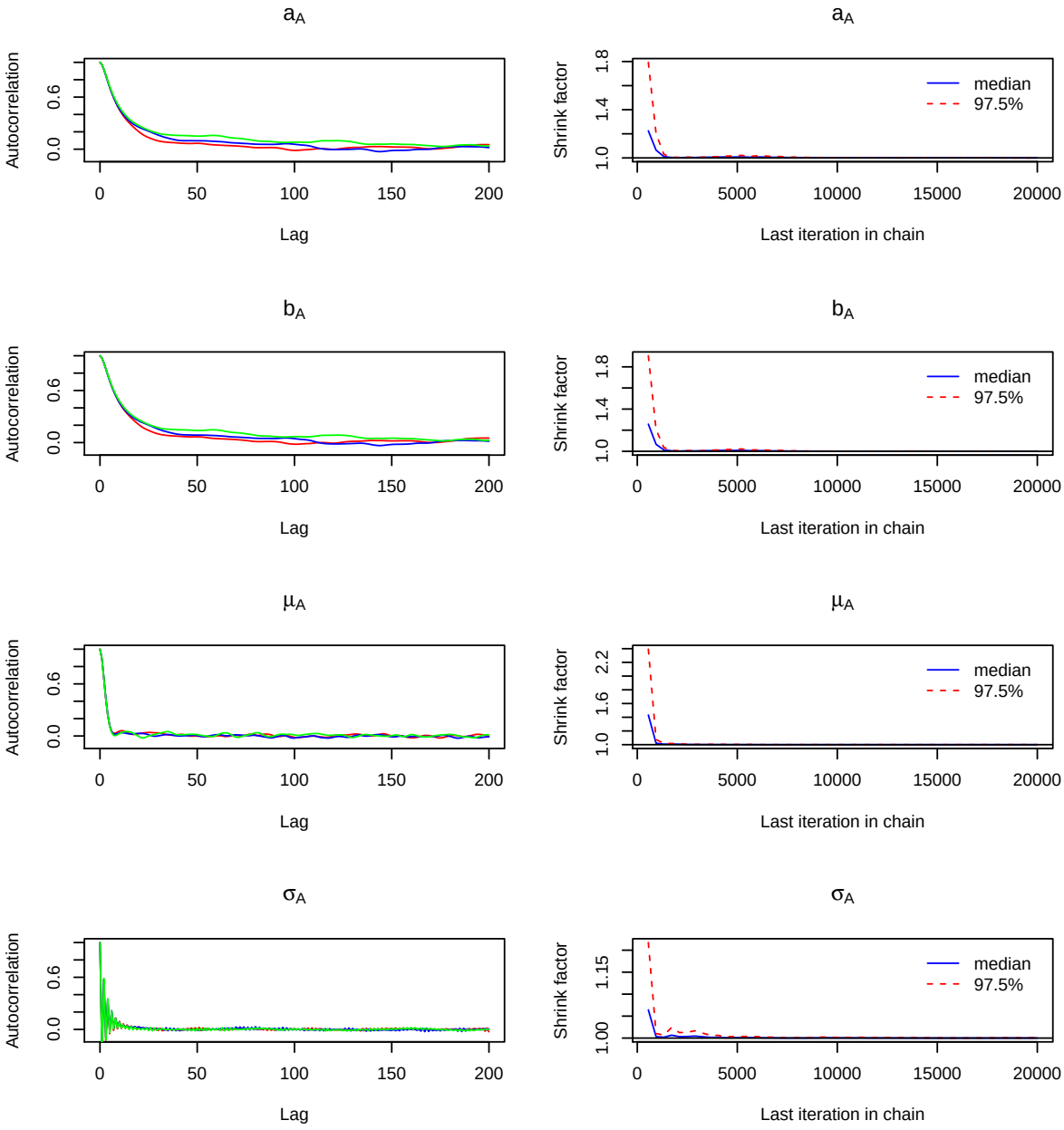


Figure 5.5: Autocorrelation plots (left) and posterior distributions (right) for parameters of amplitude.

of the second Gaussian component of $\log(\nu)$, namely the mixing proportion $w_{\nu 2}$, the mean $\mu_{\nu 2}$ (second and bottom rows of Fig. 5.2) and the standard deviation $\sigma_{\nu 2}$ (third row of Fig. 5.3). In particular, chains for $\sigma_{\nu 2}$ depict the worst behavior with a minimum value for the ACR after 200 lags of about 0.9 and a value for the CSRF of 1.25 after the 20000 iterations. In contrast, the best scenario is depicted by chains for parameters of the conditional distributions of apparent G-magnitude and amplitude given the frequency, showed respectively in Figs. 5.4 and 5.5, in particular for the slope a_{G2}, the intercept b_{G2} and the mean μ_A. For these three latter parameters, which correspond to $\log(\nu) > -1$, the ACR value is nearly zero after lags greater than 50. In the best scenario, values of the CSRF are at the end of simulation all bounded by 1.03. Finally, the behavior of chains for standard deviations σ_G and σ_A is also satisfactory.

5.5 Posterior Distributions for Parameters of Interest

In this Section we present and analyze the samples of posteriors distributions for parameters of interest and estimations of their PDFs[6]. Results are depicted in Figures 5.6 and 5.7, for parameters of logarithm of the frequency, and in Figures 5.8 and 5.9 for parameters of apparent G-magnitude and amplitude, respectively. In the left column of each figure we depict the trace plots for the last 10000 iterations of the three simulated chains. In the right columns we depict the corresponding estimated densities from the total of 30000 samples for each parameter.

Let now start by analyzing first the best scenario. This is depicted in Figs. 5.8 and 5.9 by samples and posterior densities for parameters $(a_{G2}, b_{G2}, \sigma_G)$ and (μ_A, σ_A) of the respective conditional distributions $m_G \mid \log(\nu)$ and $A \mid \log(\nu)$ when $\log(\nu) > -1$. We see that each individual sample (chain) is well mixed, that is, they traverse quickly the corresponding posterior parameter space, and this space also seems to be the same for three chains. Moreover, there is no evidence of trends in the three time series for these five parameters. All these fact are consistent with our analysis of convergence summarized in the previous Section. The corresponding posterior densities are well centered Gaussian distributions.

At the opposite end, the worst scenario is depicted by samples and estimated posterior

[6]For the sake of conciseness we postpone the presentation of associated statistics to Tables 5.2 and 5.3 in Section 5.6.

densities for parameters of the first and second Gaussian component of $\log(\nu)$. As we can see in corresponding rows of Fig. 5.6 and Fig. 5.7, individual chains for mixing proportions, means and standard deviations of these components did not mix well and present evident trends that correspond to the high autocorrelation detected in the previous Section. An extreme case is presented for $\sigma_{\nu 2}$, where chains become separated from each other and the samples we have obtained neither are independent nor belong all to the same parameter space. The corresponding densities clearly show a significant skewness to the right in the mixing proportion and mean densities of the second component, and the presence of local extrema in the posterior density of $\sigma_{\nu 2}$.

5.6 Comparison with Real Parameters

The main objective of this Section is to evaluate the ability of the model developed in Chapter 4 to retrieve the real distributions of frequencies, amplitudes and apparent G-magnitudes of the simulated Cepheids population given their recovered values in the training set $\mathcal{T} = \{(A_{\text{rec}i}, \nu_{\text{rec},i}, m_{G,\text{rec},i})\}_1^{854} \subsetneq \mathcal{D}$. Recall from Section 4.2.2 that real theoretical distributions of $\log(\nu_{\text{in}})$ and $A_{\text{in}} \mid \log(\nu_{\text{in}})$ are, respectively, a mixture of five Gaussians and a Gaussian distribution with mean independent of $\log(\nu_{\text{in}})$ if $\log(\nu_{\text{in}}) > -1$ or a linear combination of $\log(\nu_{\text{in}})$ elsewhere. Also, we do not have analytical expression for the real PDF of $m_{G_{\text{in}}} \mid \log(\nu_{\text{in}})$. To achieve this objective:

1. We compute summary statistics, namely the mean and the 2.5%-97.5% percentiles, from the samples of posterior distributions for the parameters of interest inferred by our model.

2. We compare the former posterior means with the parameters of the the real theoretical distributions.

3. We construct and depict theoretical distributions using the posterior means and compare them with: i) the empirical distribution in the set $\mathcal{I} = \left\{(\nu_i, A_i, m_{G,i})_{\text{input}}\right\}_1^{854}$ and ii) the real theoretical distributions.

The results of our analysis are shown in Tables 5.2 to 5.3 and Figure 5.10. For the decimal logarithm of the frequency $\log(\nu)$ we have tried to aid comparison in Table 5.3 by presenting posterior parameters in increasing order by means. Even so, it is difficult to make a

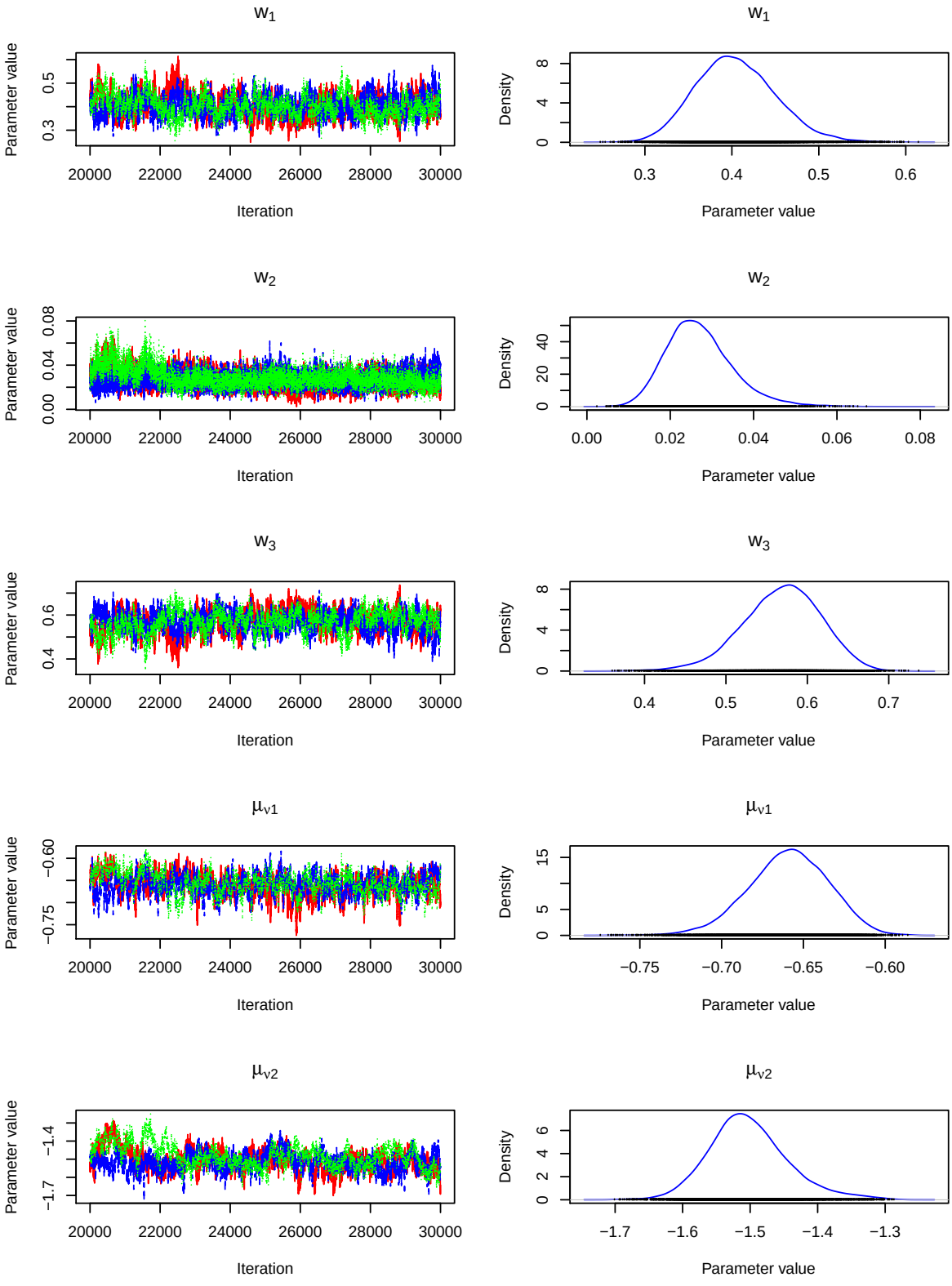


Figure 5.6: Trace plots and posterior distributions for parameters of $\log(\nu)$ (i).

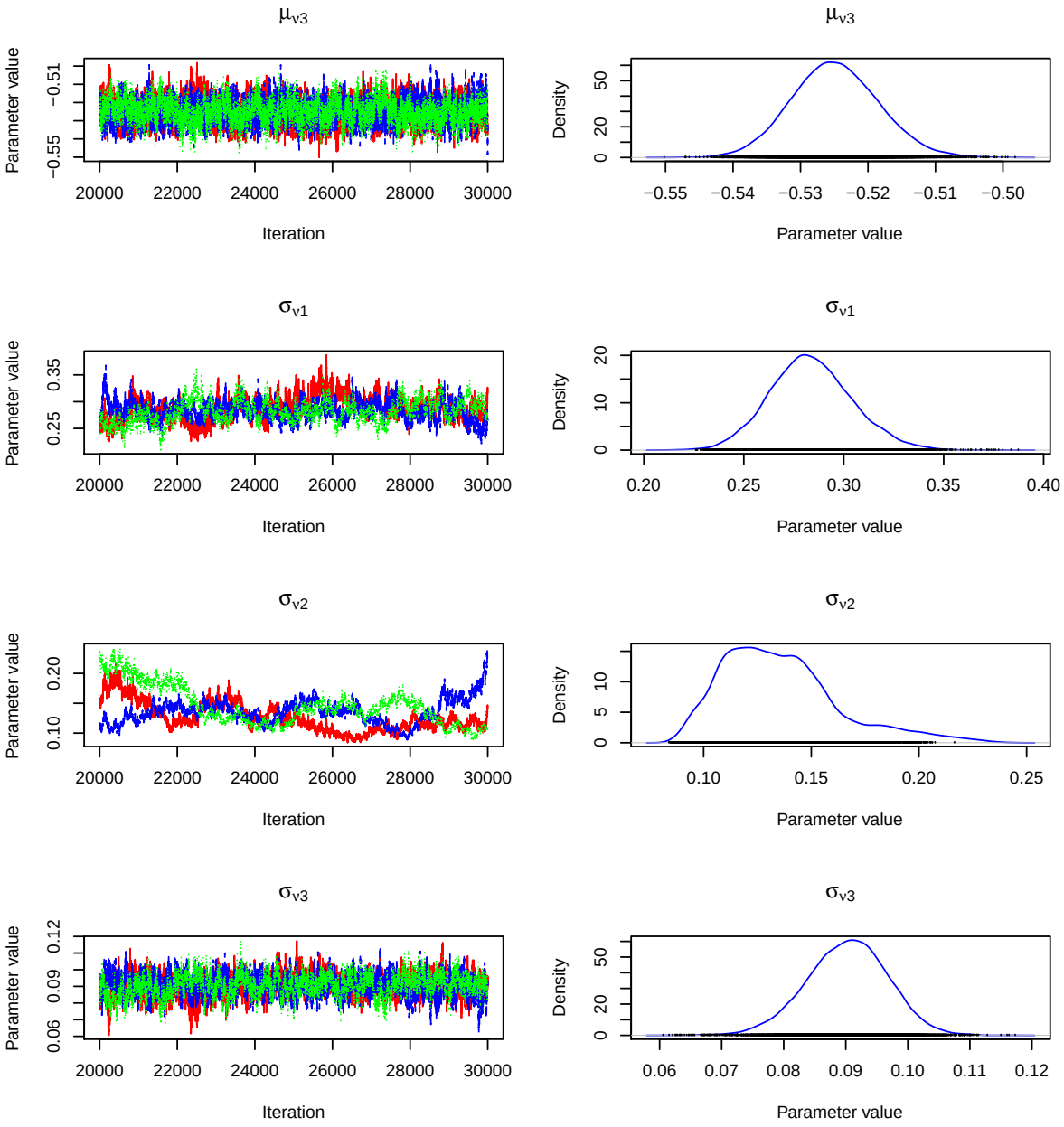


Figure 5.7: Trace plots and posterior distributions for parameters of $\log(\nu)$ (and ii).

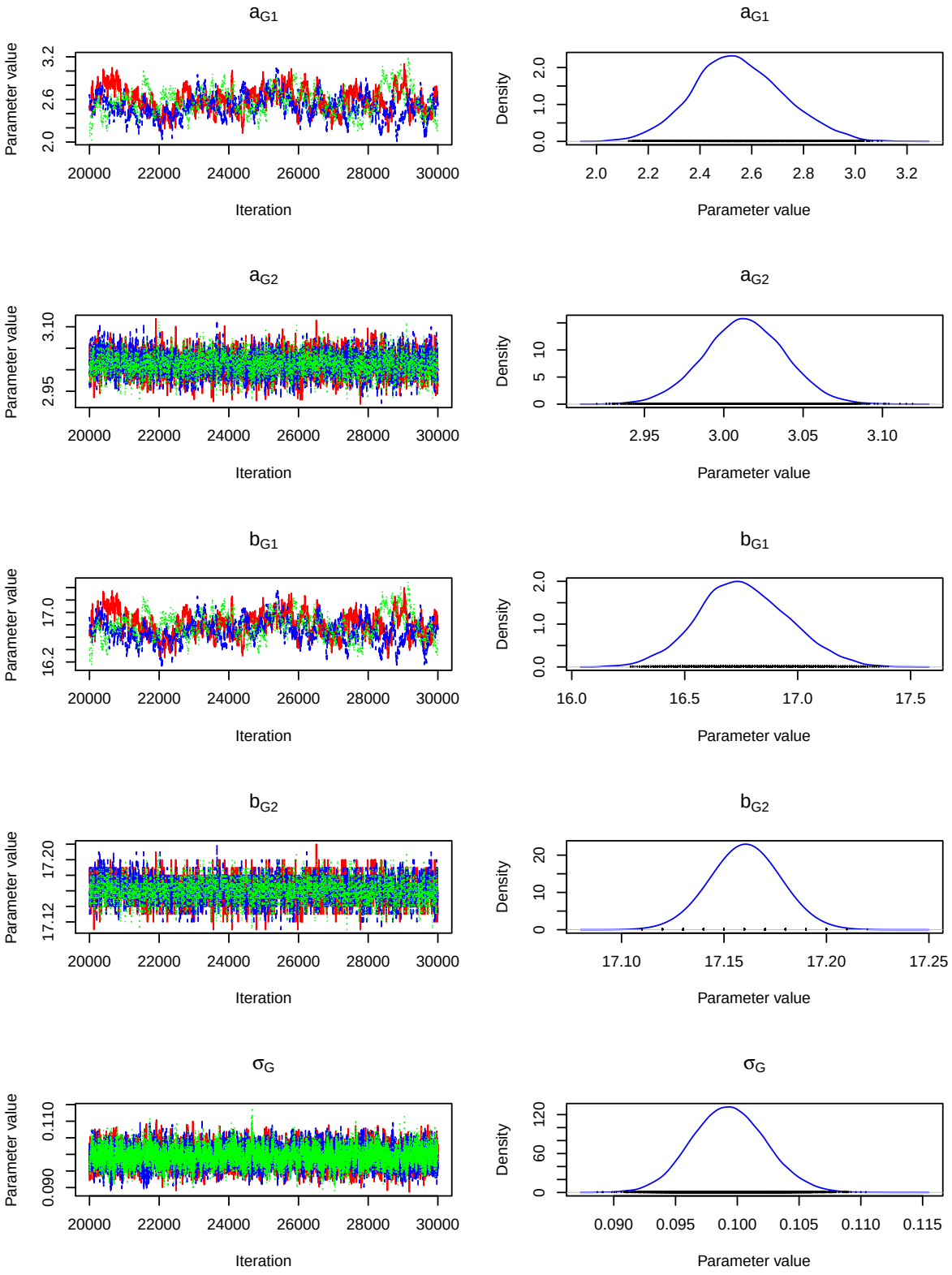


Figure 5.8: Trace plots and posterior distributions for parameters of the apparent G magnitude.

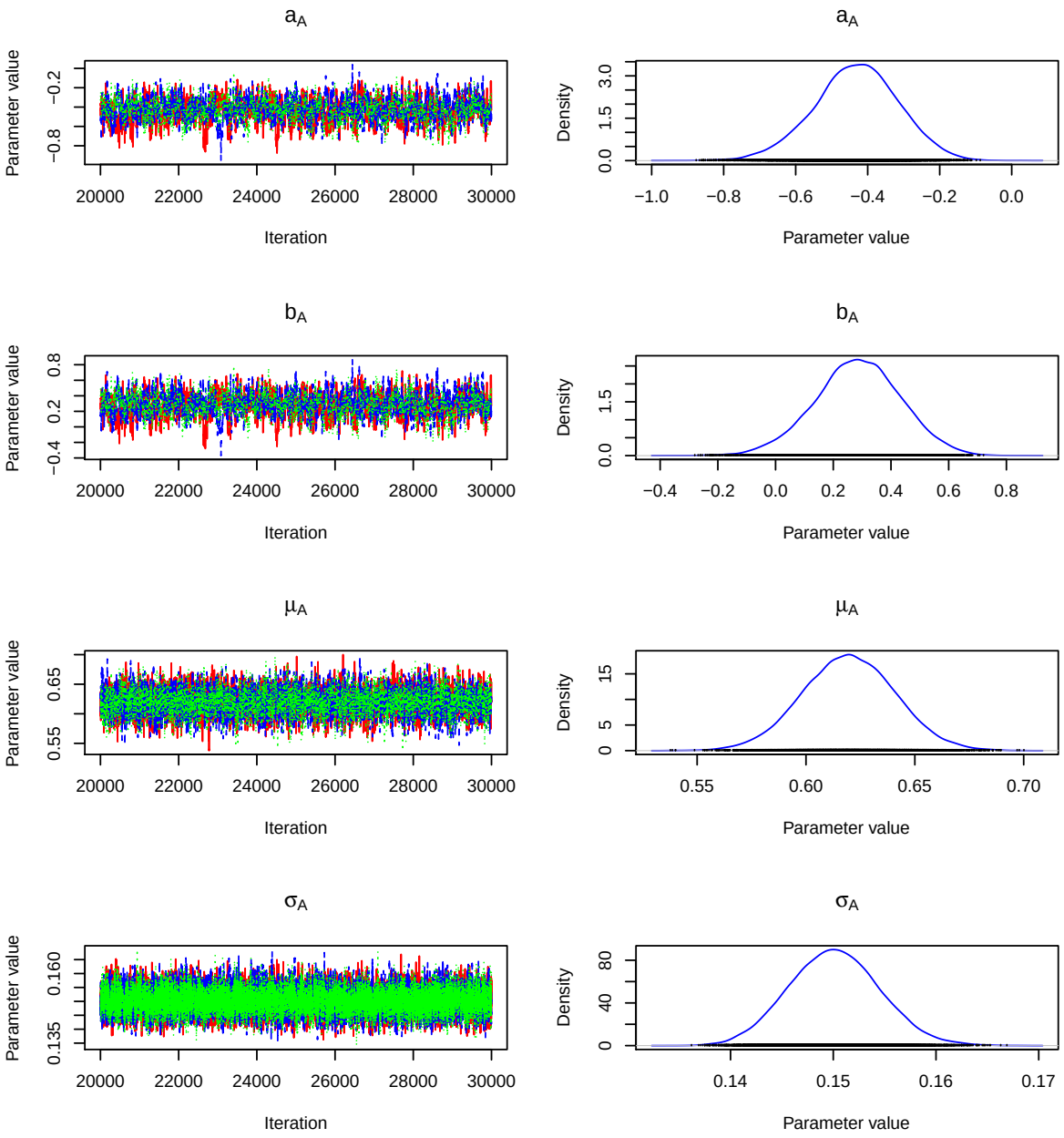


Figure 5.9: Trace plots and posterior distributions for parameters of the amplitude.

correspondence with the real parameters given that the real frequency has five Gaussian components. So, it is better to examine directly the comparison graph in the top row to the left of Figure 5.10. There we can see that fitting of the PDF for $\log(\nu)$ with only three components, depicted with the dotted blue line, is quite satisfactory and reconstruct the input PDF (solid red line) successfully.

With respect to the posterior parameters of the conditional distribution $m_{G_{\text{in}}} \mid \log(\nu_{\text{in}})$ we only dispose of the empirical distribution in $\mathcal{I}$ for comparison. The top rows of Table 5.3 shows that the interquartile range for each parameter is satisfactorily narrow. The bottom row of Fig. 5.10 show the two regression lines

$$m_{G,i} = \begin{cases} \overline{a}_{G1} \cdot \log(\nu_i) + \overline{b}_{G1} & \log(\nu) < -1 \\ \overline{a}_{G2} \cdot \log(\nu_i) + \overline{b}_{G2} & \log(\nu) > -1 \end{cases} \tag{5.1}$$

, where the two error bars indicate a one (plus/minus) inferred standard deviation $\overline{\sigma}_G$ for the corresponding conditional density of $m_{G,i}$. We see that the fitting to the empirical distribution in $\mathcal{I}$ is successful.

For parameters of the conditional distribution $A_{\text{in}} \mid \log(\nu_{\text{in}})$ we have constructed the two regression lines

$$A_i = \begin{cases} \overline{a}_A \cdot \log(\nu) + \overline{b}_A & \log(\nu) < -1 \\ \overline{\mu}_A & \log(\nu) > -1 \end{cases} \tag{5.2}$$

The central rows of Tab. 5.2 and the top row to the right of Fig. 5.10 show that the system underestimates the true value of the mean μ_A when $\log(\nu_{\text{in}}) > -1$.

The last objective of this Section is to evaluate the ability of the multinomial logistic regression submodel given by Equation 4.23 to constrain the mixing proportions of the mixtures of distributions which model recovered frequencies and amplitudes (Equations 4.21 and 4.25). Recall that parameters $\lambda_{\beta j}$, $j \in \{1, 2\}$, represent the slopes for the standardized ecliptic latitude β' for each aliased component and λ_{0j} represent the intercepts. We see in the bottom rows of Tab. 5.3 that the interquartile range for $\lambda_{\beta j}$ is too wide. Nevertheless, the decrements of the corresponding *logits* are significant. For example, for a unitary increment in β' the $\ln\left(\frac{\pi_{i2}}{\pi_{i1}}\right)$ decreases on average by -0.766, which corresponds to a new value of the (π_{i2}/π_{i1}) ratio equal to 46.5% of the initial ratio.

Inf. Parameter	**Posterior Mean**	**2.5%-97.5% Percentiles**	**Real value**	**Real parameter**
$w_{\nu 2}$	0.027	0.014,0.046	0.126	$w_{\nu 1}$
-	-	-	0.109	$w_{\nu 2}$
$w_{\nu 1}$	0.405	0.322,0.502	0.419	$w_{\nu 3}$
$w_{\nu 3}$	0.568	0.465 ,0.656	0.104	$w_{\nu 4}$
-	-	-	0.247	$w_{\nu 5}$
$\mu_{\nu 2}$	-1.502	-1.607 ,-1.365	-0.989	$\mu_{\nu 1}$
-	-	-	-0.643	$\mu_{\nu 2}$
$\mu_{\nu 1}$	-0.659	-0.710 ,-0.614	-0.618	$\mu_{\nu 3}$
$\mu_{\nu 3}$	-0.525	-0.537 ,-0.512	-0.536	$\mu_{\nu 4}$
-	-	-	-0.476	$\mu_{\nu 5}$
$\sigma_{\nu 2}$	0.136	0.095 ,0.204	0.355	$\sigma_{\nu 1}$
-	-	-	0.049	$\sigma_{\nu 2}$
$\sigma_{\nu 1}$	0.284	0.245 ,0.326	0.225	$\sigma_{\nu 3}$
$\sigma_{\nu 3}$	0.090	0.077,0.103	0.046	$\sigma_{\nu 4}$
-	-	-	0.048	$\sigma_{\nu 5}$

Table 5.2: Summary statistics of posterior distributions for parameters of the decimal logarithm of the frequency $\log(\nu)$ and comparison with its real parameters. w, μ and σ denote, respectively, mixing proportions, means and standard deviation of each Gaussian component.

Parameter	**Posterior Mean**	**2.5%-97.5% Percentiles**	**Real value**
a_{G1}	2.551	2.221 ,2.909	-
b_{G1}	16.762	16.38 ,17.17	-
a_{G2}	3.013	2.962 ,3.063	-
b_{G2}	17.161	17.130 ,17.190	-
σ_G	0.099	0.094,0.105	-
a_A	-0.4317	-0.671 ,-0.209	-0.5
b_A	0.2848	-0.018 ,0.572	0.2
μ_A	0.6197	0.578,0.661	0.7
σ_A	0.1501	0.142,0.159	0.15
λ_{02}	-1.132	-1.314 ,-0.955	-
λ_{03}	-0.981	-1.156 ,-0.816	-
$\lambda_{\beta 2}$	-0.766	-1.140 ,-0.395	-
$\lambda_{\beta 3}$	-0.743	-1.091 ,-0.385	-

Table 5.3: Summary statistics of posterior distributions for the rest of parameters of interest and comparison with its real parameters when proceed. a, b and σ denote, respectively, slopes, intercepts and standard deviations for conditional distributions of apparent G-magnitude G and amplitude A given the decimal logarithm of the frequency; and λ denote coefficients of the logistic regression submodel with covariate the rescaled ecliptic latitude β'.

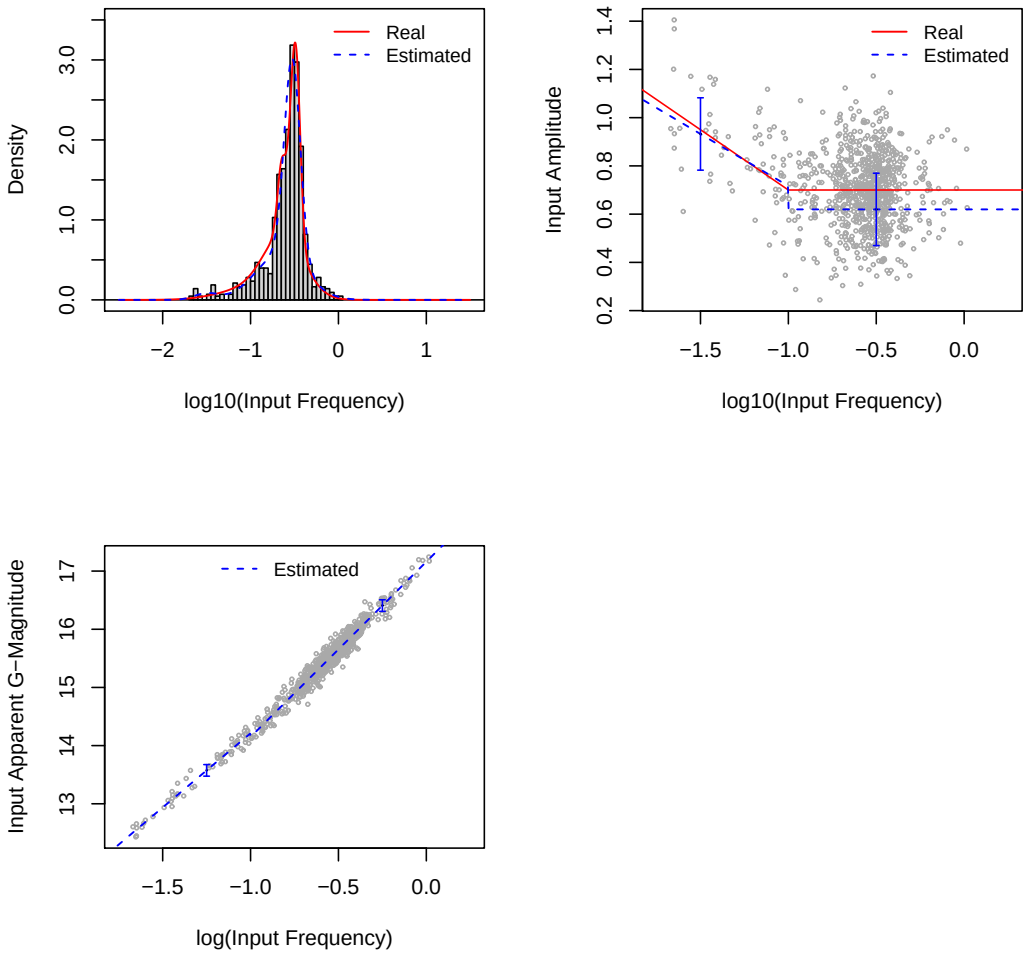


Figure 5.10: Real versus estimated marginal distributions for frequencies and conditional distributions of apparent G-magnitude and amplitude given the the frequency. The estimated regression lines for A and m_G given $\log(\nu)$ are constructed from posterior means in Table 5.3. Their cut-point is at $\log(\nu) = -1$. The error bars indicate a one (plus/minus) inferred standard deviation ($\overline{\sigma}_A$ or $\overline{\sigma}_G$) for the corresponding Gaussian conditional distribution.

Chapter 6

Summary and Conclusions

6.1 Conclusions

In this thesis we have presented a two-level BGM to infer the true distributions of the physical parameters amplitude, frequency and apparent G-magnitude of the LMC classical Cepheid population from their observed values by the Gaia satellite (after processed by the CU7 software of the DPAC). We have modeled the real frequencies (in logarithmic scale) by means of a mixture of Gaussian distributions with a fixed number of components. We have used piecewise linear models (with a fixed knot value depending on the frequency) to model the dependency of the real amplitudes and G-magnitudes on the logarithm of the real frequency. We have modeled the observed amplitudes and frequencies by mean of mixtures of distributions with some fixed parameters. We have considered in our BGM the photometric measurement error in the Gaia broad G-band and tackled the aliasing problem in the frequencies recovery which arises as a product of the Gaia scanning law. We have modeled the recovery probabilities of aliased frequencies by means of a logistic regression submodel using the ecliptic latitude as a predictor. The model has not addressed completely the aliasing problem because we have only used some predefined configurations of aliased data and discarded the rest. Furthermore, we have restricted to a very narrow range of ecliptic latitudes in which the relationship between the recovery probability of aliased frequencies and the ecliptic latitude is monotone.

We have used and algorithm based in MCMC simulation techniques for the inference, due to the complexity of the problem parameter space. We have performed the evaluation of the model with simulated data given that the Gaia mission was not yet completely

operational. We have obtained a successful approximation for the distribution of the real frequencies once the model has been trained. We have obtained also good results in the estimation of the conditional distributions of real G-magnitudes. Unfortunately, our results have also shown a significant bias in the estimation of the conditional amplitude distribution for one of the ranges of the piecewise linear model.

6.2 Future Work

We summarize some possible extension for our model and future investigation lines as follows:

- To include a parameterization for distances (see Section 4.2.2 and Appendix A). For that, we should consider that the coordinates of the stars are observed variables (data) and are not constants as we have assumed for the ecliptic latitude in the model presented in this thesis. At present, we are interested in an investigation line which proposes us the construction of a simpler BGM which includes such a parameterization for a RR Lyrae star population gathered from the OGLE III catalog.

- To model a number of parameters considered fixed in our model as random parameters. This includes the hyperparameters for the skewed Student t distributions in the mixture for the recovered amplitude of Equation 4.25 and the intercepts for the spurious components in the mixture for the recovered frequencies of Equation 4.21. It also includes the number of Gaussian components for the distributions of the real and recovered frequencies. Finally, it includes the cut-point $\log(\nu) = -1$ of the piecewise linear models for the real amplitudes and G-magnitudes.

- To analyze other Cepheids populations, e.g. those of our Galaxy. This implies also to revise the logistic regression submodel used to model the dependency of the recovery probability of aliased frequencies on the ecliptic latitude.

- To change and enhance the evaluation methodology of the model. The key point here is that we have used the real (but simulated) values of the parameters of interest for comparisons. These parameters actually are the focus of the inference and will not be available when we work with a sample provided by Gaia. An option is to use

the posterior predictive distribution to check if the inferred parameters are discrepant with the observations in systematic ways.

- To search for an alternative parameterization for the probabilities of the *loci* of recovered frequencies. The logistic regression submodel, with the ecliptic latitude as a covariate, imposes severe constraints on those probabilities. This fact influences in turn on the model of recovered frequencies as a mixture of Gaussians. Perhaps it would be better to use non-informative priors for the recovery probabilities of each star.

- To integrate the model into the software of CU7. This line implies to implement the model in the Java language.

References

Adams, M. R.; Cornish, N. J. and Littenberg, T. B. (2012). «Astrophysical Model Selection in Gravitational Wave Astronomy». *Physical Review D*, **86(12)**, p. 124032.

Aerts, C. (2007). «Lecture notes on Asteroseismology». *Technical Report*, Universities of Leuven and Nijmegen.

Antonello, E; Fugazza, D and Mantegazza, L (2002). «Variable stars in nearby galaxies. VI. Frequency-period distribution of Cepheids in IC 1613 and other galaxies of the Local Group». *Astronomy and Astrophysics*, **388**, pp. 477–482.

Asensio Ramos, A. and Arregui, I. (2013). «Coronal Loop Physical Parameters from the Analysis of Multiple Observed Transverse Oscillations». *Astronomy and Astrophysics*, **554**, p. A7.

Ashdown, I. and Eng, P. (2002). «Photometry and radiometry. A tour guide for computer graphics enthusiasts». *Technical Report*, Ledalite Architectural Products, Inc.

Azzalini, A. (2013). *R package* `sn`*: The skew-normal and skew-t distributions (version 0.4-18)*. Università di Padova, Italia.
`http://azzalini.stat.unipd.it/SN`

Azzalini, A. and Genton, M. G. (2008). «Robust Likelihood Methods Based on the Skew-t and Related Distributions». *International Statistical Review*, **76(1)**, pp. 106–129.

Barentsen, G.; Vink, J. S.; Drew, J. E. and Sale, S. E (2013). «Bayesian Inference of T Tauri Star Properties Using Multi-Wavelength Survey Photometry». *Monthly Notices of the Royal Astronomical Society*, **429(3)**, pp. 1981–2000.

Bessell, M. S. (1990). «UBVRI passbands». *Astronomical Society of the Pacific*, pp. 1181–1199.

Bessell, MS; Castelli, F and Plez, B (1998). «Model atmospheres broad-band colors, bolometric corrections and temperature calibrations for O-M stars». *Astronomy and Astrophysics*, **333**, pp. 231–250.

Bono, Giuseppe; Castellani, Vittorio and Marconi, Marcella (2002). «Theoretical Models for Bump Cepheids». *The Astrophysical Journal*, **565**, pp. L83–L86.

Brewer, B. J.; Marshall, P. J.; Auger, M. W.; Treu, T.; Dutton, A. A. and Barnabè, M. (2014). «The SWELLS Survey - VI. Hierarchical Inference of the Initial Mass Functions of Bulges and Discs». *Monthly Notices of the Royal Astronomical Society*, **437(2)**, pp. 1950–1961.

Brooks, S. P. and Gelman, A. (1998). «General Methods for Monitoring Convergence of Iterative Simulations». *Journal of computational and graphical statistics*, **7(4)**, pp. 434–455.

Chib, S. and Greenberg, E. (1995). «Understanding the Metropolis-Hastings Algorithm». *The American Statistician*, **49(4)**, pp. 327–335.

Cooper, G.F. and Herskovits, E. (1992). «A Bayesian method for the induction of probabilistic networks from data». *Machine learning*, **9(4)**, pp. 309–347.

Cuypers, J. (2013). *Period Search Software Requirement Specification GAIA-C7-SP-ROB-JCU-002-4*. DPAC.

Cuypers, J. and Guy, L. (2011). *Variability Characterisation Software Requirement Specification GAIA-C7-SP-ROB-JCU-007*. DPAC.

Dawid, A. P. (1979). «Conditional Independence in Statistical Theory». *Journal of the Royal Statistical Society. Series B (Methodological)*, **41(1)**, pp. 1–31.

Dawid, A. P. (1980). «Conditional Independence for Statistical Operations». *The Annals of Statistics*, **8(3)**, pp. 598–617.

De Bruijne, J. (2003). «Stellar fluxes: transformations and calibrations». *Technical Report GAIA-JdB-005*, DPAC.

De Meyer, F. (2003). «Deconvolution of the Fourier spectrum». *Technical Report 33*, Koninklijk Meteorologisch Instituut van België.

De Ridder, J.; Guy, L.; Lecoeur, I.; Soldan, J. and Aerts, C. (2009). *Time Series modelling Software Requirements Specification GAIA-C7-SP-IVS-JDR-004*. DPAC.

Deeming, T.J. (1975). «Fourier analysis with unequally-spaced data». *Astrophysics and Space Science*, **36(1)**, pp. 137–158.

Díez, F.J. (2010). *Introducción a los Modelos Gráficos Probabilistas*. UNED, Madrid. `http://www.ia.uned.es/~fjdiez/libros/intro-mgp.html`

Djorgovski, S.G.; Brunner, R.; Mahabal, A.; Williams, R.; Granat, R. and Stolorz, P. (2003). «Challenges for Cluster Analysis in a Virtual Observatory». In: *Statistical Challenges in Astronomy*, pp. 127–141. Springer.

Eyer, L. and Mignard, F. (2005). «Rate of Correct Detection of Periodic Signal with the Gaia Satellite». *Monthly Notices of the Royal Astronomical Society*, **361(4)**, pp. 1136–1144.

Eyer, L.; Mowlavi, N.; Varadi, M.; Spano, M.; Lecoeur-Taibi, I. and Clementini, G. (2009). «The GAIA Mission and Variable Stars». In: *Proceedings of the Annual meeting of the French Society of Astronomy and Astrophysics*, .

Feigelson, E.D. and Babu, G.J. (2003). *Statistical Challenges in Astronomy*. Springer.

García Sedano, F. (2012). *Diseño de un Prototipo de Sistema Experto para la generación de los Informes de Calidad Relativos a los Objetos Variables detectados por la Misión Espacial GAIA*. Master's thesis, ETSI Informática, UNED.

Gelman, A.; Carlin, J. B.; Stern, H. S. and Rubin, D. B. (2004). *Bayesian*. Chapman & Hall/CRC.

Gelman, Andrew; Jakulin, Aleks; Pittau, Maria Grazia and Su, Yu-Sung (2008). «A weakly informative default prior distribution for logistic and other regression models». *The Annals of Applied Statistics*, pp. 1360–1383.

Gelman, Andrew and Shalizi, Cosma Rohilla (2013). «Philosophy and the practice of Bayesian statistics». *British Journal of Mathematical and Statistical Psychology*.

Gentle, James E; Härdle, Wolfgang Karl and Mori, Yuichi (2012). *Handbook of computational statistics: concepts and methods*. Springer.

Heckerman, D. (1996). «A Tutorial on Learning With Bayesian Networks». *Technical Report MSR-TR-95-06*, Microsoft Research.

Hemberger, D. A.; Lovelace, G.; Loredo, T. J.; Kidder, L. E.; Scheel, M. A.; Szilágyi, B.; Taylor, N. W. and Teukolsky, S. A. (2013). «Final Spin and Radiated Energy in Numerical Simulations of Binary Black Holes with Equal Masses and Equal, Aligned or Antialigned Spins». *Physical Review D*, **88(6)**, p. 064014.

Hobson, M.P.; Jaffe, A.H.; Liddle, A.R.; Mukherjee, P. and Parkinson, D. (2010). *Bayesian Methods in Cosmology*. Bayesian Methods in Cosmology. Cambridge University Press.

Hogg, D. W.; Myers, A. D. and Bovy, J. (2010). «Inferring the Eccentricity Distribution». *The Astrophysical Journal*, **725(2)**, p. 2166.

Højsgaard, S.; Edwards, D. and Lauritzen, S. (2012). *Graphical models with R*. Springer.

Jordi, C. (2012). «Photometric relationships between Gaia photometry and existing photometric systems». *Technical Report GAIA-C5-TN-UB-CJ-041-7*, DPAC.

Jordi, C.; Fabricius, C.; Figueras, F.; Voss, H. and Carrasco, J.M. (2007). «Proposal for the internal calibration of G and integrated G-BP and G-RP fluxes». *Technical Report GAIA-C5-TN-UB-CJ-042-001*, DPAC.

Jordi, C.; Fabricius, C.; Figueras, F.; Voss, H. and Carrasco, J.M. (2009). «Error model for photometry and spectrophotometry». *Technical Report GAIA-C5-TN-UB-CJ-047*, DPAC.

Jordi, C; Gebran, M; Carrasco, JM; de Bruijne, J; Voss, H; Fabricius, C; Knude, J; Vallenari, A; Kohley, R and Mora, A (2010). «Gaia broad band photometry». *Astronomy and Astrophysics*, **48**, pp. 1–14.

Jordi, C.; Høg, E.; Brown, A. G. A.; Lindegren, L.; Bailer-Jones, C. A. L. and

Carrasco, J. M. (2006). «The design and performance of the Gaia photometric system». *Monthly Notices of the Royal Astronomical Society*, **367(1)**, pp. 290–314.

Karttunen, H.; Kröger, P.; Oja, H. and Poutanen, M. (2007). «Photometric Concepts and Magnitudes». In: Hannu Karttunen et al. (Ed.), *Modern Astronomy*, chapter 4, pp. 83–93. Springer-Verlag Berlin Heidelberg, 5th edition.

Kelly, B. C.; Shetty, R.; Stutz, A. M.; Kauffmann, J.; Goodman, A. A. and Launhardt, R. (2012). «Dust Spectral Energy Distributions in the Era of Herschel and Planck: A Hierarchical Bayesian-fitting Technique». *The Astrophysical Journal*, **752(1)**, p. 55.

Kelly, B. C.; Treu, T.; Malkan, M.; Pancoast, A. and Woo, J. (2013). «Active Galactic Nucleus Black Hole Mass Estimates in the Era of Time Domain Astronomy». *The Astrophysical Journal*, **779(2)**, p. 187.

Koen, C. and Eyer, L. (2002). «New Periodic Variables from the Hipparcos Epoch Photometry». *Monthly Notices of the Royal Astronomical Society*, **331(1)**, pp. 45–59.

Korb, K.B. and Nicholson, A.E. (2003). *Bayesian Artificial Intelligence*. CRC Computer Science & Data Analysis. Chapman & Hall.

Kuhn, T.S. (2012). *The structure of scientific revolutions*. University of Chicago press, Chicago London, secondth edition.

Lauritzen, S.L. (1996). *Graphical Models*. Oxford University Press.

Leavitt, H. S. and Pickering, E. C. (1912). «Periods of 25 Variable Stars in the Small Magellanic Cloud». *Harvard College Observatory Circular*, **173**, pp. 1–3.

Lindegren, L; Babusiaux, C; Bailer-Jones, C; Bastian, U; Brown, A.G.A.; Cropper, M; Høg, E; Jordi, C; Katz, D; van Leeuwen, F; ; Luri, X; Mignard, F; de Bruijne, J.H.J. and Prusti, D (2008). «The Gaia Mission: Science, Organization and Present Status». In: *Proceedings of the International Astronomical Union*, volume 3, S248, pp. 217–223. Cambridge Univ Press.

Léna, P.; Rouan, D.; Lebrun, F.; Mignard, F. and Pelat, D. (2012). *Observational Astrophysics*. Springer-Verlag, Berlin Heidelberg.

Loredo, T. J. (2013). «Bayesian Astrostatistics: A Backward Look to the Future». In: *Astrostatistical Challenges for the New Astronomy*, volume 1, pp. 15–40. Springer.

López Del Fresno, M.; Solano Márrquez, E. and Sarro Baro, L.M. (2011). «Data mining in the Spanish Virtual Observatory. Applications to Corot and Gaia». In: *Highlights of Spanish Astrophysics VI*, volume 1, pp. 721–726.

Lunn, D.; Jackson, C.; Best, N.; Thomas, A. and Spiegelhalter, D. (2012). *The BUGS Book: A Practical Introduction to Bayesian Analysis*. CRC Texts in Statistical Science. Chapman & Hall.

Lunn, David; Spiegelhalter, David; Thomas, Andrew and Best, Nicky (2009). «The BUGS project: Evolution, critique and future directions». *Statistics in medicine*, **28(25)**, pp. 3049–3067.

March, M. C.; Trotta, R.; Berkes, P.; Starkman, G. D. and Vaudrevange, P. M. (2011). «Improved Constraints on Cosmological Parameters from Type Ia Supernova Data». *Monthly Notices of the Royal Astronomical Society*, **418(4)**, pp. 2308–2329.

Marr, D.; Ullman, S. and Poggio, T. (2010). *Vision: A Computational Investigation Into the Human Representation and Processing of Visual Information*. MIT Press.

Martinez, G. D. (2013). «A Robust Determination of Milky Way Satellite Properties using Hierarchical Mass Modeling». *pre-print*. ArXiv:1309.2641 [astro-ph.GA].

Mignard, F. (2005). «About the Nyquist Frequency». *Technical Report GAIA-FM-022*, Observatoire de la Côte d'Azur.

Mignard, F; Bailer-Jones, C; Bastian, U; Drimmel, R; Eyer, L; Katz, D; van Leeuwen, F; Luri, X; O'Mullane, W; Passot, X; Pourbaix, D and Prusti, D (2008). «Gaia: Organisation and Challenges for the Data Processing». In: *Proceedings of the International Astronomical Union*, volume 3, S248, pp. 224–230. Cambridge Univ Press.

Mira, J. and Delgado, A. E. (2009). «Sensory representation spaces in neuroscience and computation». *Neurocomputing*, **72(4)**, pp. 793–805.

Mira, J. and Delgado, A.E. (2001). «Aspectos Metodológicos en IA». In: *Aspectos básicos de la Inteligencia Artificial*, pp. 53–87. Sanz y Torres.

Moitinho, A.; Miranda, B.; Gomes, M. and Ribeiro, R. (2011). *Global Variability Studies: Bias Estimation (GWP-S-721-04000) Software Requirement Specification GAIA-C7-SP-SIM-AM-001*. DPAC.

Mowlavi, N; Wyrzykowski, L and Varadi, M (2011). *Light curve simulation of variable objects*. DPAC.

Neal, R. M. (1993). «Probabilistic Inference Using Markov Chain Monte Carlo Methods». *Technical Report CRG-TR-93-1*, Department of Computer Science, University of Toronto.

Neal, R. M. (2003). «Slice Sampling». *Annals of Statistics*, **31(3)**, pp. 705–767.

Neapolitan, R.E. (2004). *Learning bayesian networks*. Prentice Hall, NJ.

Newell, Allen (1982). «The Knowledge Level». *Artificial Intelligence*, **18**, pp. 87–127.

Pearl, J. (1985). «Bayesian Networks: A Model of Self-Activated Memory for Evidential Reasoning». *Technical Report CSD-850021*, Cognitive Science Departament, UCA.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausble Inference*. Morgan Kaufmann Pub.

Plummer, M.; Best, N.; Cowles, K. and Vines, K. (2006). «CODA: Convergence Diagnosis and Output Analysis for MCMC». *R news*, **6(1)**, pp. 7–11.
`http://cran.r-project.org/web/packages/coda/index.html`

Proakis, J. G. and Manolakis, D. G. (1996). *Digital Signal Processing (3rd Ed.): Principles, Algorithms, and Applications*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA. ISBN 0-13-373762-4.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
`http://www.R-project.org/`

Robert, Christian P and Mengersen, Kerrie L (1999). «Reparameterisation issues in mixture modelling and their bearing on MCMC algorithms». *Computational Statistics & Data Analysis*, **29(3)**, pp. 325–343.

Robert, C.P. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer, New York, 2th edition.

Roberts, G. O. and Rosenthal, J. S. (2004). «General State Space Markov Chains and MCMC Algorithms». *Probability Surveys*, **1**, pp. 20–71.

Sale, S. E. (2012). «3D Extinction Mapping Using Hierarchical Bayesian Models». *Monthly Notices of the Royal Astronomical Society*, **427(3)**, pp. 2119–2131.

Sandage, A; Tammann, GA and Reindl, B (2004). «New Period-Luminosity and Period-Color Relations of Classical Cepheids II. Cepheids in LMC». *Astronomy and Astrophysics*, **424**, pp. 43–71.

Smirnov, S. (2001). «New improved ones' trick». Website. Last checked: 02.11.2014. `https://www.jiscmail.ac.uk/cgi-bin/webadmin?A2=BUGS;6a80f387.01`

Soiaporn, K.; Chernoff, D.; Loredo, T.; Ruppert, D. and Wasserman, I. (2013). «Multilevel Bayesian Framework for Modeling the Production, Propagation and Detection of Ultra-High Energy Cosmic Rays». *Annals of Applied Statistics*, **7(3)**, pp. 1249–1285.

Solano, E. (2006). «The Virtual Observatory». *Lecture Notes and Essays in Astrophysics*, **2**, pp. 71–90.

Spiegelhalter, D.J. and Lauritzen, S.L. (1990). «Sequential Updating of Conditional Probabilities on Directed Graphical Structures». *Networks*, **20(5)**, pp. 579–605.

Spirtes, P.; Glymour, C.N. and Scheines, R. (2000). *Causation, prediction, and search*. The MIT Press.

Stein, N. M.; Van Dyk, D. A.; von Hippel, T.; DeGennaro, S.; Jeffery, E. J. and Jefferys, W. H. (2013). «Combining Computer Models to Account for Mass Loss in Stellar Evolution». *Statistical Analysis and Data Mining*, **6(1)**, pp. 34–52.

Studenỳ, M. (2005). *On Probabilistic Conditional Independence Structures*. Springer.

Taylor, Julian (2012). *hett: Heteroscedastic t-regression*. R package version 0.3-1. `http://CRAN.R-project.org/package=hett`

Weyant, A.; Schafer, C. and Wood-Vasey, W. M. (2013). «Likelihood-free Cosmological Inference with Type Ia Supernovae: Approximate Bayesian Computation for a Complete Treatment of Uncertainty». *The Astrophysical Journal*, **764(2)**, p. 116.

Appendix A

A Modified BGM with Parameterization for Distances

Recall from Section 4.2.2 that the starting point for this parameterization is the deterministic relation

$$m_G = M_G + 5\,(log\,(r) - 1) \tag{A.1}$$

Absolute Magnitudes in G Band Taking into account Eq. 4.4 applied to the absolute G-magnitude and Eq. 4.7 we parametrize this node as

$$f\,(M_{G,i} \mid \log\,(\nu_i)\,, a_{G1}, b_{G1}, a_{G2}, b_{G2}, \tau_G) = \tag{A.2}$$
$$\mathbf{1}_{\{\log(\nu_i)<-1\}}\mathsf{N}\,(a_{G1}\cdot\log\,(\nu_i) + b_{G1}, \tau_G) + \mathbf{1}_{\{\log(\nu_i)>-1\}}\mathsf{N}\,(a_{G2}\cdot\log\,(\nu_i) + b_{G2}, \tau_G)$$

with the prior distribution

$$\begin{aligned} p\,(a_{G1}) &= \mathsf{N}\,(0, 0.001) \\ p\,(a_{G2}) &= \mathsf{N}\,(0, 0.001) \\ p\,(b_{G1}) &= \mathsf{N}\,(0, 0.001) \\ p\,(b_{G2}) &= \mathsf{N}\,(0, 0.001) \\ p\,(\tau_G) &= \mathsf{Gamma}\,(0.001, 0.001) \end{aligned} \tag{A.3}$$

Distance We compute the decimal logarithm $\log\,(r_i)$ of the distance deterministically from the distribution of Cartesian Geocentric equatorial coordinates of the LMC sources,

distribution that in turn depends (also deterministically) on a model of the galaxy as a (random) exponential disk whose parameters are our inference focus. In the present paragraph we present the parametrization for $\log(r_i)$ leaving for the next epigraph the parameterization for the disk exponential model. The logarithm of the distance $\log(r_i)$ is computed as

$$\log(r_i) = \log\left(\left\|(x_i', y_i', z_i')\right\|_2\right) \tag{A.4}$$

where

- (x_i', y_i', z_i') are the Cartesian Geocentric equatorial coordinates of the LMC sources computed as

$$(x_i', y_i', z_i') = (x_i, y_i, z_i) \cdot T^t + (x_0', y_0', z_0') \tag{A.5}$$

- $(x_0', y_0', z_0') = (3.29, 17.59, -46.69)$ are the Cartesian Geocentric equatorial coordinates of the disk center equivalent to the spherical equatorial coordinates $(\alpha_0, \delta_0, r_0) = (1.39, -1.20, 50)$.[1]

- The transformation matrix T correspond to the composition of rotations[2]

$$\mathrm{rot}_{z''}^{+}\left(\alpha_0 - \frac{\pi}{2}\right) \circ \mathrm{rot}_{x'}^{-}\left(\delta_0 + \frac{\pi}{2}\right) \circ \mathrm{rot}_{z'}^{+}(\theta) \circ \mathrm{rot}_{x}^{-}(i) \tag{A.6}$$

, with $\theta = 0.51$rad (position angle) and $i = 0.54$rad (inclination angle), and is given by

$$T = \begin{pmatrix} 0.94 & -0.25 & -0.22 \\ 0.28 & 0.95 & 0.15 \\ 0.17 & -0.21 & 0.96 \end{pmatrix} \tag{A.7}$$

- (x, y, z) are the Cartesian coordinates of the disk exponential model.

Radial distance and height To parametrize the proper Cartesian coordinates (x_i, y_i, z_i) of the disk we first compute the third coordinate z_i deterministically as

$$z_i = (-1)^{B_i} \cdot h_i \tag{A.8}$$

[1] Angles and distances are expressed respectively in rad and kpc.
[2] The super-indexes '+' and '-' denote, respectively, clockwise and counterclockwise rotations.

where:

- $B_i = \mathsf{Bern}\,(0.5)$ is a semi-cylinder indicator, with values 0 or 1 for a point above or below the disk plane, respectively.
- h_i is the the height above (or below) the plane of the disk (always positive). For this variable we assign the exponential distribution

$$p\,(h_i \mid h_z) = \mathsf{Exp}\,(1/h_z) \tag{A.9}$$

- h_z is the *scale height* of the disk. For this vertical scale factor we take the non informative prior

$$p\,(1/h_z) = \mathsf{Gamma}\,(0.001, 0.001) \tag{A.10}$$

Secondly we compute deterministically the two first Cartesian coordinates as

$$\begin{aligned} x_i &= R_i \cdot \cos\,(\varphi_i) \\ y_i &= R_i \cdot \sin\,(\varphi_i) \end{aligned} \tag{A.11}$$

where:

- R_i is the radial distance from the disk center (measured in the plane of the disk). For this variable we assign the exponential distribution

$$p\,(R_i \mid h_R) = \mathsf{Exp}\,(1/h_R) \tag{A.12}$$

- h_R is the *radial scale length*. For this radial scale factor we take the non informative prior

$$p\,(1/h_R) = \mathsf{Gamma}\,(0.001, 0.001) \tag{A.13}$$

- φ_i is the position angle drawn in the plane of the disk, for which we assign the prior

$$p\,(\varphi_i) = \mathsf{U}\,(0, 2\pi)$$