Astronomy and Computing 5 (2014) 45–56

Contents lists available at ScienceDirect

Astronomy and Computing

journal homepage: www.elsevier.com/locate/ascom

Full length article

How to combine correlated data sets—A Bayesian hyperparameter matrix method

Y.-Z. Ma [a,b,*], A. Berndsen [a]

[a] Department of Physics and Astronomy, University of British Columbia, Vancouver, V6T 1Z1, BC, Canada

[b] Canadian Institute for Theoretical Astrophysics, Toronto, Canada

ARTICLE INFO

Article history:
Received 11 November 2013
Accepted 24 April 2014

Keywords:
Bayesian analysis
Data analysis
Statistical method
Observational cosmology

ABSTRACT

We construct a "hyperparameter matrix" statistical method for performing the joint analyses of multiple correlated astronomical data sets, in which the weights of data sets are determined by their own statistical properties. This method is a generalization of the hyperparameter method constructed by Lahav et al. (2000) and Hobson et al. (2002) which was designed to combine independent data sets. The advantage of our method is to treat correlations between multiple data sets and gives appropriate relevant weights of multiple data sets with mutual correlations. We define a new "element-wise" product, which greatly simplifies the likelihood function with hyperparameter matrix. We rigorously prove the simplified formula of the joint likelihood and show that it recovers the original hyperparameter method in the limit of no covariance between data sets. We then illustrate the method by applying it to a demonstrative toy model of fitting a straight line to two sets of data. We show that the hyperparameter matrix method can detect unaccounted systematic errors or underestimated errors in the data sets. Additionally, the ratio of Bayes' factors provides a distinct indicator of the necessity of including hyperparameters. Our example shows that the likelihood we construct for joint analyses of correlated data sets can be widely applied to many astrophysical systems.

1. Introduction

Due to the fast development of astronomical observations such as the measurements of the cosmic microwave background temperature anisotropy (e.g. *WMAP* Hinshaw et al., 2013 and *Planck* Ade et al., 2013b satellites) and observations of galaxy clustering (e.g. 6dF Magoulas et al., 2012 and SDSS Nuza et al., 2013 galaxy surveys), more and more large-scale data sets are available for studying a variety of astrophysical systems. It is, therefore, a common practice in astronomy to combine different data sets to obtain the joint likelihood for astrophysical parameters of interest. The standard approach for this joint analysis assumes that the data sets are independent, therefore the joint likelihood is simply the product of the likelihood of each data set. The joint likelihood function can then be used to determine optimal parameter values and their associated uncertainties. In the frequentist approach to parameter estimation, this is equivalent to the weighted sum of the parameter constraints from the individual data sets, where the weight of each data set is the inverse variance. Data sets with small errors provide stronger constraints on the parameters.

There is a long history discussing the appropriate way to combine observations from different experiments. In the context of cosmology, the discussion can be traced back to Godwin and Lynden-Bell (1987) and Press (1996), where weight parameters were assigned to different data sets to obtain joint constraints on the velocity field and Hubble parameter H_0. In these approaches, however, the assignment of weights to data sets with differing systematic errors was, in some ways, ad-hoc. For instance, if a data set has large systematic error and is not reliable, it is always assigned a weight of zero and is effectively excluded from the joint analysis. On the other hand, a more trustworthy data set can be assigned a higher relative weighting.

Due to the subjectivity and limitations of this traditional way of assigning weights to different data sets, Lahav et al. (2000) and Hobson et al. (2002, hereafter HBL02) developed the original hyperparameter method. This allows the statistical properties of the data themselves to determine the relative weights of each data set. In the framework developed by Lahav et al. (2000) and HBL02, a set of hyperparameters is introduced to weight each

* Corresponding author at: Department of Physics and Astronomy, University of British Columbia, Vancouver, V6T 1Z1, BC, Canada. Tel.: +1 6048222945.
E-mail addresses: mayinzhe@phas.ubc.ca (Y.-Z. Ma), berndsen@phas.ubc.ca (A. Berndsen).

http://dx.doi.org/10.1016/j.ascom.2014.04.005
2213-1337/

independent data set, and the posterior distribution of the model parameters is recovered by marginalization over the hyperparameters. The marginalization can be carried out with a brute-force grid evaluation of the hyperparameters, or it can be explored by using Monte Carlo methods which directly sample the posterior distribution. Such possibilities include Markov chain Monte Carlo (MCMC) algorithms such as Metropolis–Hastings and Simulated Annealing, or non-MCMC methods such as Nested Sampling (Skilling, 2004). The application of hyperparameters was considered for a variety of cases by HBL02. For instance, if the error of a data set is underestimated, the direct combination of data sets (no hyperparameter) results in an underestimated error-budget, providing unwarranted confidence in the observation and producing a fake detection of the signal. The hyperparameter method, however, was shown to detect such a phenomenon and act to broaden the error-budget, thus recovering the true variance of the data sets. By using the hyperparameter method, the results of joint constraints become more robust and reliable. This approach has also been applied to the joint analysis of the primordial tensor mode in the cosmic microwave background radiation (CMB) (Ma et al., 2010), the distance indicator calibration (Erdogdu et al., 2003), the study of mass profile in galaxy clusters (Host and Hansen, 2011), and the cosmic peculiar velocity field study (Ma et al., 2012).

Notably, the hyperparameter method established by Lahav et al. (2000) and HBL02 is limited to independent data sets, where "no correlation between data sets" is assumed in the joint analysis. In the analysis of cosmology and many other astrophysical systems, the data sets sometimes are correlated. For instance, in the study of the angular power spectrum of the CMB temperature fluctuations, the data from the Atacama Cosmology Telescope (ACT), South Pole Telescope (SPT) and *Planck* satellite share a large range of multipole moments ℓ (see Fig. 1 of Cheng et al., 2013 and Fig. 11 of Ade et al., 2013a). When combining these observations, one needs to consider the correlated cosmic variance term since these data are drawn from a close region of the sky. In addition, in the study of the cosmic velocity field (Ma and Scott, 2013), the bulk flows from different peculiar velocity surveys are drawn from the same underlying matter distribution so, in principle, a non-zero correlation term exists between different peculiar velocity samples. Therefore, a method both using hyperparameter method and taking into account the correlation between different data sets is needed in the study of astrophysics. Providing such a method is the main aim of this paper.

For a clear presentation, we build up our method step-by-step from the most basic level, explaining the concepts and derivation process in a pedagogical way. The structure of the paper is as follows. In Section 2, we review Bayes' theorem (Section 2.1) and the standard multivariate Gaussian distribution (Section 2.2) in the absence of any hyperparameters. Section 2.3 provides a review of the hyperparameter method as proposed in HBL02. In Section 2.4 we present the hyperparameter matrix method, which is the core of the new method proposed in this paper. We quote the appropriate likelihood function for the hyperparameter matrix method for correlated data in Section 2.4, leaving its derivation and proofs of its salient features in Appendix A. The proof of the functional form for the joint likelihood of correlated data sets makes use of several recondite matrix operations and lemmas. These are laid out in Appendices B and C, while the main text simply quotes their results. In Section 3, we apply our method to a straight-line model while fitting two independent data sets. We vary the error-budget and systematic errors in each data set to test the behaviour of the hyperparameter matrix method. In Section 3.4, we also discuss the improvement of our hyperparameter matrix method over the original method proposed by HBL02. The conclusion and discussion are presented in the last section.

Table 1
Jeffreys' empirical criterion for strength of evidence (Jeffreys, 1961).

K value	Strength of evidence
<1	Negative (supporting H_0)
1–3	Weak
3–10	Substantial
10–30	Strong
30–100	Very Strong
>100	Decisive

2. Statistical method

2.1. Bayes theorem

Let us suppose that our data set is represented by D and the parameters of interest are represented by vector $\vec{\theta}$. Then by Bayes' theorem, the posterior distribution $\Pr(\vec{\theta}|D)$ is given by

$$\Pr(\vec{\theta}|D) = \frac{\Pr(D|\vec{\theta})\Pr(\vec{\theta})}{\Pr(D)}, \tag{1}$$

where $\Pr(D|\vec{\theta})$ is called the likelihood function,[1] $\Pr(\vec{\theta})$ is the prior distribution of parameters and $\Pr(D)$ is the Bayesian evidence, an important quantity for model selection.

Given a data set D, let us suppose we have two alternative models (or hypotheses) for D, namely H_0 and H_1. One can calculate the Bayesian evidence for each hypothesis $H \in \{H_0,\ H_1\}$ as

$$\Pr(D|H) = \int \Pr(D|\vec{\theta})\Pr(\vec{\theta})\,\mathrm{d}\vec{\theta}, \tag{2}$$

where the integral is performed over the entire parameter space $\vec{\theta}$ of each model H. Note that the models may have different sets of parameters. The evidence is an important quantity in the Bayesian approach to parameter fitting, and it plays a central role in model selection (Jeffreys, 1961; Kass and Raftery, 1995). Specifically, if we have no prior preference between models H_1 and H_0, the ratio between two Bayesian evidences gives a model selection criterion, or Bayes' factor

$$K = \frac{\Pr(H_1|D)}{\Pr(H_0|D)} = \frac{\Pr(D|H_1)}{\Pr(D|H_0)}. \tag{3}$$

The value of K indicates whether the model H_1 is favoured over model H_0 by data D. Jeffreys (1961) gave an empirical scale for interpreting the value of K, as listed in Table 1. We will use this table as a criterion to assess the improvement of statistical significance when using the hyperparameter matrix method.

2.2. Multivariate Gaussian distribution

Let us now consider the combination of multiple data sets, coming from a collection of different surveys S. Each survey provides n_i number of measurements (D_i) of the quantity we are trying to fit, whose expectation value by our hypothesis is μ_i. For each survey S_i we form the data vector $\vec{x}^{S_i}$ with the following elements:

$$x_j^{S_i} \equiv D_j - \mu_j, \quad j \in \{1, \ldots, n_i\}. \tag{4}$$

The data vector is the difference between the observed value and the expected value, characterizing the error in the measurement. As such, it is also referred to as the error vector. We combine the

[1] Sometimes it is written as $L(\vec{\theta})$, but here we stick to the notation $\Pr(D|\vec{\theta})$.