

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

Yin-Zhe Ma^{a,b,*}, Aaron 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.*

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, Bridle, & Lahav (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.

Keywords: Bayesian analysis, data analysis, statistical method, observational cosmology

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* (Planck results XVI., 2013) 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 & 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, Bridle, & Lahav (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 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

*Email Address: mayinzhe@phas.ubc.ca

**Email Address: berndsen@phas.ubc.ca

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, Zhao, & Brown, 2010), the distance indicator calibration (Erdogdu, Ettori, & Lahav, 2003), the study of mass profile in galaxy clusters (Host & Hansen, 2011), and the cosmic peculiar velocity field study (Ma, Branchini, & Scott, 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, Huang, & Ma 2013 and Fig. 11 of Planck results XV. 2013). 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 & Scott, 2012), 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 Appendix B and Appendix 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.

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

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

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

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

where $\Pr(D|\vec{\theta})$ is called the likelihood function¹, $\Pr(\vec{\theta})$ 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

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

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, 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

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

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

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

μ_i . For each survey S_i we form the data vector \vec{x}^{S_i} with the following elements

$$x_j^{S_i} \equiv D_j - \mu_j, j \in \{1, \dots, n_i\}. \quad (4)$$

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 different data sets by forming a total data vector \vec{x} from the individual survey vectors \vec{x}^{S_i}

$$\vec{x} = \begin{pmatrix} \vec{x}^{S_1} \\ \vec{x}^{S_2} \\ \dots \\ \vec{x}^{S_N} \end{pmatrix}, \quad (5)$$

resulting in a vector with dimension

$$\dim(\vec{x}) = \sum_{i=1}^N n_i = N_t. \quad (6)$$

In the particular case where all of the data sets have the same number of samples, the individual data vectors \vec{x}^{S_i} have the same dimension $\dim(\vec{x}^{S_i}) = n_i \equiv n$ ($i = 1, \dots, N$), and $N_t = n \cdot N$.

The covariance matrix² is, generically,

$$\begin{aligned} \tilde{C} &= \langle \vec{x} \vec{x}^T \rangle \\ &= \begin{pmatrix} \langle \vec{x}^{S_1} \vec{x}^{S_1 T} \rangle & \langle \vec{x}^{S_1} \vec{x}^{S_2 T} \rangle & \dots & \langle \vec{x}^{S_1} \vec{x}^{S_N T} \rangle \\ \langle \vec{x}^{S_2} \vec{x}^{S_1 T} \rangle & \langle \vec{x}^{S_2} \vec{x}^{S_2 T} \rangle & \dots & \langle \vec{x}^{S_2} \vec{x}^{S_N T} \rangle \\ \dots & \dots & \dots & \dots \\ \langle \vec{x}^{S_N} \vec{x}^{S_1 T} \rangle & \langle \vec{x}^{S_N} \vec{x}^{S_2 T} \rangle & \dots & \langle \vec{x}^{S_N} \vec{x}^{S_N T} \rangle \end{pmatrix} \\ &= \begin{pmatrix} (C^{S_1}) & (C^{S_1 S_2}) & \dots & (C^{S_1 S_N}) \\ (C^{S_1 S_2}) & (C^{S_2}) & \dots & (C^{S_2 S_N}) \\ \dots & \dots & \dots & \dots \\ (C^{S_1 S_N}) & (C^{S_2 S_N}) & \dots & (C^{S_N}) \end{pmatrix}, \quad (7) \end{aligned}$$

where each $C^{S_i S_j}$ is an $N_i \times N_j$ matrix, characterizing the auto- or cross-correlation between the vectors \vec{x}^{S_i} and \vec{x}^{S_j} .

Finally, the χ^2 statistic for the combined data vector \vec{x} is

$$\chi^2 = \vec{x}^T \tilde{C}^{-1} \vec{x}, \quad (8)$$

and the Gaussian likelihood is

$$\Pr(D|\vec{\theta}) = \frac{1}{(2\pi)^{\frac{N_t}{2}} \sqrt{\det \tilde{C}}} \exp\left(-\frac{1}{2} \vec{x}^T \tilde{C}^{-1} \vec{x}\right). \quad (9)$$

Equation (9) is the Gaussian likelihood function of μ_i ($i = 1, \dots, N$) with respect to the data. However, the likelihood is a multivariate Gaussian in parameter space only if the μ_i is a linear function of the parameters of interest. In a more general case, both \tilde{C} and μ_i ($i = 1, \dots, N$) in Eq. (9) may have a dependence on the model parameters $\vec{\theta}$, so the likelihood function (9) is not Gaussian in parameter space. But this is not a problem if we evaluate the likelihood function numerically.

²Note, in Section 2.4 of this paper we use C to represent the covariance matrix with hyperparameters. \tilde{C} is the special case of C evaluated with all hyperparameters set to unity.

Note that when we combine multiple surveys with correlated data as in Eq. (9), we give each data set equal weight, and combine them all together without distinguishing whether some data set has poorer estimated error or unaccounted systematic errors. If a data sets' error and systematic effects are properly accounted for, this method can give an unbiased estimate of the parameters of interest. However, if errors or systematic errors exist, the method can give biased results or exaggerated significance. We provide several such examples in Section 3, and compare with our hyperparameter matrix method.

2.3. Combining independent data sets: Original hyperparameter method

The original hyperparameter method, as proposed by Lahav et al. (2000) and HBL02, assumes that different data sets are independent of one another. That is $C^{S_i S_j} = \delta_{ij} C^{S_i S_i}$, δ_{ij} being the Kronecker-delta, in which case the covariance matrix becomes block diagonal. "Hyperparameters" α_i are introduced as a rescaling of the error vector

$$\vec{x}_i \rightarrow \vec{x}_i / \sqrt{\alpha_i}. \quad (10)$$

This is equivalent to rescaling the individual blocks, or data sets, of the covariance matrix

$$C^{S_i} \rightarrow \alpha_i^{-1} C^{S_i}, \quad (11)$$

for the i th survey.

With the hyperparameter rescaling of Equation (10) and the assumption of independent data sets, the total covariance matrix becomes

$$C = \begin{pmatrix} \alpha_1^{-1} C^{S_1} & 0 & \dots & 0 \\ 0 & \alpha_2^{-1} C^{S_2} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \alpha_N^{-1} C^{S_N} \end{pmatrix}. \quad (12)$$

Since the autocorrelation C^{S_i} is the covariance of the i th data set, the hyperparameters clearly act to re-weight the internal errors of the survey S_i . Exploring different values of the hyperparameters is equivalent to exploring potential systematic errors and error estimates of the data set (see our examples in Sec. 3). In this case the total χ^2 for N combined data sets becomes

$$\chi^2 = \sum_{i=1}^N \alpha_i \chi_i^2, \quad (13)$$

and the joint likelihood, including hyperparameter and parameters of interest, becomes

$$\Pr(D|\vec{\theta}, \vec{\alpha}) = \prod_{i=1}^N \frac{\alpha_i^{n_i/2}}{\sqrt{(2\pi)^{n_i} \det(C^{S_i})}} \exp\left(-\frac{1}{2} \alpha_i \chi_i^2\right). \quad (14)$$

Equation (14) is obtained in HBL02 (eq. (32)) as the general result of a likelihood function with hyperparameters. By re-deriving it here, we emphasize the assumption of independent data sets and show the effect of introducing hyperparameters. Specifically, a large hyperparameter α_i increases the error-budget of the i th data set and reduces its' constraint in the likelihood function. Conversely, a small hyperparameter α_i increases the significance of the i th data set.

2.4. Combining correlated data sets: Hyperparameter matrix method

The original hyperparameter method shown in Section 2.3 is only for the case where different data sets do not have cross-correlation terms, i.e. all off-diagonal matrix entries $C^{S_i S_j} = 0$ for $i \neq j$. In this section we generalize the hyperparameter formalism to the case where correlations between individual data sets is non-negligible, i.e. generalize to the case when C includes $C^{S_i S_j} \neq 0$ if $(i \neq j)$.

As before, for each experiment i we introduce a hyperparam-

eter α_i as a rescaling of the error vector

$$\vec{x}_i \rightarrow \vec{x}_i / \sqrt{\alpha_i}, \quad (15)$$

but we drop the assumption of independent data sets. The (sub) covariance matrices become

$$C^{S_i S_j} \rightarrow C^{S_i S_j} / \sqrt{\alpha_i \alpha_j}. \quad (16)$$

Therefore, for N correlated data sets, the full covariance matrix with hyperparameters becomes

$$C = \begin{pmatrix} \alpha_1^{-1} C^{S_1} & (\alpha_1 \alpha_2)^{-1/2} C^{S_1 S_2} & \dots & (\alpha_1 \alpha_N)^{-1/2} C^{S_1 S_N} \\ (\alpha_1 \alpha_2)^{-1/2} (C^{S_1 S_2})^T & \alpha_2^{-1} C^{S_2} & \dots & (\alpha_2 \alpha_N)^{-1/2} C^{S_2 S_N} \\ \dots & \dots & \dots & \dots \\ (\alpha_1 \alpha_N)^{-1/2} (C^{S_1 S_N})^T & (\alpha_2 \alpha_N)^{-1/2} (C^{S_2 S_N})^T & \dots & \alpha_N^{-1} C^{S_N} \end{pmatrix}, \quad (17)$$

where each C^{S_i} ($i = 1, \dots, N$) is an $n_i \times n_i$ symmetric, positive-definite matrix, while $C^{S_i S_j}$ is an $n_i \times n_j$ asymmetric matrix if $n_i \neq n_j$.

To simplify the matrix calculations in this case, we define an $N \times N$ hyperparameter matrix P with elements $P_{ij} = (\alpha_i \alpha_j)^{-1/2}$ ($i, j = 1, \dots, N$). Thus

$$P = \begin{pmatrix} \alpha_1^{-1} & (\alpha_1 \alpha_2)^{-1/2} & \dots & (\alpha_1 \alpha_N)^{-1/2} \\ (\alpha_1 \alpha_2)^{-1/2} & \alpha_2^{-1} & \dots & (\alpha_2 \alpha_N)^{-1/2} \\ \dots & \dots & \dots & \dots \\ (\alpha_1 \alpha_N)^{-1/2} & (\alpha_2 \alpha_N)^{-1/2} & \dots & \alpha_N^{-1} \end{pmatrix}. \quad (18)$$

Note that the covariance matrices C (Eq. (17)) and \tilde{C} (Eq. (7)) are $N_i \times N_i$ matrices, while P is an $N \times N$ matrix for the N data sets, $N \leq N_i$. The relation between \tilde{C} , P , \vec{C} cannot be linked by any ordinary matrix product. Here we define a new ‘‘element-wise’’ product ‘‘ \odot ’’ which multiplies any $N \times N$ hyperparameter matrix with any $N_i \times N_i$ covariance matrix, i.e.

$$C = P \odot \tilde{C}. \quad (19)$$

The \odot operation proceeds as follows. We first expand each hyperparameter P_{ij} to an $n_i \times n_j$ matrix by multiplying the $(\alpha_i \alpha_j)^{-1/2}$ value to an $n_i \times n_j$ unit matrix $J_{n_i n_j}$, where all elements are equal to one³, while keeping the partition of P_{ij} values the same as the hyperparameter matrix (18)—this is equivalent to the Kronecker product. Then we do a Hadamard product (see Appendix B) for the extended hyperparameter matrix with the covariance matrix \tilde{C} (Eq. (7)) to obtain the total covariance matrix C (Eq. (17)).

We can now write the likelihood function which includes both parameters of interest $\vec{\theta}$ and hyperparameter vector $\vec{\alpha}$ as

$$\Pr(D|\vec{\theta}, \vec{\alpha}) = \frac{1}{(2\pi)^{\frac{N}{2}} \sqrt{\det(C(\vec{\alpha}))}} \exp\left(-\frac{1}{2} \vec{x}^T C(\vec{\alpha})^{-1} \vec{x}\right), \quad (20)$$

³In order to distinguish the unit matrix from the identity matrix, we denote this matrix with J and identity matrix with I ; see Appendix B for illustration.

where we indicate, explicitly, the dependence of C on the hyperparameter vector $\vec{\alpha}$.

Since the values of hyperparameters $\vec{\alpha}$ can take any values between 0 and infinity, we might be worried for the positive definiteness of the covariance matrix $C(\vec{\alpha})$. Fortunately there are several important properties of the hyperparameter covariance matrix C that make it positive definite, and therefore invertible with a positive determinant. The rigorous proofs of these properties can be found in Appendix A. Here, we exploit these useful properties to greatly simplify the generalized χ^2 calculation, so that Eq. (20) can be re-expressed as

$$\Pr(D|\vec{\theta}, \vec{\alpha}) = \left[\prod_{i=1}^N \left(\frac{\alpha_i}{2\pi} \right)^{n_i/2} \right] \frac{1}{\sqrt{\det \tilde{C}}} \exp\left(-\frac{1}{2} \vec{x}^T (\hat{P} \odot \tilde{C}^{-1}) \vec{x}\right). \quad (21)$$

In the above expression, \hat{P} is the Hadamard inverse of the hyperparameter matrix P , \tilde{C}^{-1} is the inverse matrix of the correlation matrix (Eq. (7)) without hyperparameters, and \odot is the ‘‘element-wise’’ product.

This form of the likelihood function is a key result of this work, which is a generalized expression for the joint distribution of combined, correlated data sets. As a consistency check, we note that in the case of independent data sets, \tilde{C}^{-1} is block diagonal, and Eq. (21) reduces to the original hyperparameter likelihood function (Eq. (14)). As well, in the case of equal weights to all data sets the hyperparameter matrix P is the unit matrix, and one recovers the standard multivariate Gaussian distribution, Eq. (9).

3. Example of fitting a straight line

Having derived the joint likelihood function for correlated data sets together with hyperparameters, we now investigate a simple demonstrative example of fitting data with a straight line. The goal is to combine two different data sets for improved constraints on the posterior distribution $\Pr(\vec{\theta}|D)$. As was

shown in HBL02, the original hyperparameter method is particularly useful for overcoming the common problems of inaccurately quoted error bars and the presence of systematic errors in the measurements. We reproduce the results of HBL02 here as a validation of our method, and show that the hyperparameter matrix extension to correlated data sets also overcomes these problems, and provides a preferred method for describing the model parameters.

Starting with the assumption that the underlying model for some process is a straight line with slope $m = 1$ and intercept $c = 1$

$$y(x) = mx + c, \quad (22)$$

we generated two independent sets of measurements D_1 and D_2 for the quantity y . For each data set five x -values were randomly drawn from a uniform distribution over $(0, 1)$, and the corresponding y -values were drawn from a Gaussian distribution of known variance σ_k and mean $\mu_k = mx_k + c$. In this way we know the “true” values of the model parameters m and c , which we attempt to recover. Following the notation of Section 2, the parameters of interest are $\vec{\theta} = (m, c)$, and the hyperparameters for the two data sets are $\vec{\alpha} = (\alpha_1, \alpha_2)$.

In this simple case, where the number of data sets, measurements and parameters is small, one could determine the posterior distribution (Eq. (1)) by evaluating the likelihood function and prior distributions on a grid. However, this method scales geometrically with the number of free parameters and exponentially with the number of grid points, and can quickly become impractical to evaluate. Instead, in this work the posterior distributions of the parameters and hyperparameters (if present) are obtained using Monte-Carlo Markov-Chains (MCMC). Specifically, we use the default settings in the PyMC (Patil et al., 2010) framework, which uses a Metropolis-Hastings sampling of the prior distributions. In this way, the marginalized posterior distributions for each parameter are recovered from the traces of the MCMC runs, and the evidence integrals are determined from the trace of the likelihood function (Kass, 1995; Raftery, 2007). Weinberg (2010) points out that using the mean or harmonic mean of the likelihood function can produce spurious results if there is a lot of variance in the likelihood, but we have checked the evidence ratios are consistent with his quadrature formulation.

In the following subsections we investigate the behaviour of the original hyperparameter and hyperparameter matrix method in comparison to the standard non-hyperparameter method. We consider three different cases, as listed in Section 3.1-3.2. The arrangement of these case studies is similar to that of HBL02, though we add the case of correlated data sets. This facilitates readers to directly compare the original hyperparameter method and our hyperparameter matrix method. For ease of comparison, the Bayes’ factor from each of the different cases is presented in Table 2.

In all cases the prior distributions on the slope m and intercept c are uniform over the interval $(0, 2)$, and the prior for hyperparameters is $\Pr(\alpha) = \exp(-\alpha)$ in the range $(0, 10)^4$. Recall that a

hyperparameter of unity is equivalent to no additional weighting, removing the effects of the weights. As such, it is natural to use prior distributions for the hyperparameters which give a mean of one, preferring an analysis with no re-weighting (as did in HBL02). For $\Pr(\alpha) = \exp(-\alpha)$, it is a properly normalized prior function ($\int_0^\infty \Pr(\alpha)d\alpha = 1$) with mean value equal to unity ($\int_0^\infty \Pr(\alpha)\alpha d\alpha = 1$). Therefore in the following we will adopt such prior function, and thus confirm that our results of Bayes’ factor are consistent with the values given by HBL02.

The posterior distributions for the parameters of interest are obtained by $O(10^5)$ MCMC steps, with a burn-in of 5000. We denote the non-hyperparameter method as hypothesis H_0 , with the likelihood given by Eq. (9). H_1 is reserved for the original hyperparameter method, appropriate for data sets with no correlation, and whose likelihood is given by Eq. (14). Finally, we denote the hyperparameter matrix method as H_2 , whose likelihood is given by Eq. (21), which allows for correlated data sets.

3.1. Accurate error-bars and no systematic error

3.1.1. Independent data sets

In this first case, both data sets D_1 and D_2 are drawn from the correct model $m = 1, c = 1$, with a noise rms of $\sigma_1 = \sigma_2 = 0.1$. From the experimental side, both data sets are (correctly) assumed to have an rms of 0.1 in the likelihood, or $C^{S_1} = C^{S_2} = 0.01$. The two data sets and the underlying model are shown in the left panel of Fig. 1, with the middle panel depicting the posterior distributions $\Pr(m, c|D, H_i)$ from the standard non-hyperparameter analysis (H_0) and the original hyperparameter analysis (H_1), and the right panel showing the posterior distributions of the hyperparameters, $\Pr(\vec{\alpha}|D, H_1)$. In this case both hyperparameters are consistent with unity, indicating that the hyperparameter method is not playing an important role in parameter estimation.

Both hypotheses contain the true parameter values $(m, c) = (1, 1)$ within the 1σ confidence level, however the Bayesian evidence ratio

$$\frac{\Pr(D|H_1)}{\Pr(D|H_0)} = 0.61 \quad (23)$$

indicates that the introduction of hyperparameters is marginally disfavoured. Here we see one of the powerful results of a Bayesian approach to combining data sets: the Bayes’ factor offers a simple but distinct method for model selection (Jeffreys, 1961; Kass, 1995). The preference of H_0 is not surprising in this case, since both experiments “estimated” the correct variance in the underlying distributions, $\sigma_1 = \sigma_2 = 0.1$.

3.1.2. Correlated data sets

In this section we apply the hyperparameter matrix method, but where the data sets are (anti) correlated at the 10% level, $\rho = C^{S_1 S_2} / \sqrt{C^{S_1} C^{S_2}} = -0.1$. As before, both data sets D_1 and D_2 are drawn from the correct model $m = 1, c = 1$, with internal errors of $\sigma_1 = \sigma_2 = 0.1$. In addition, however, we

⁴The range of values are chosen in order to give enough sampling space for

hyperparameters.

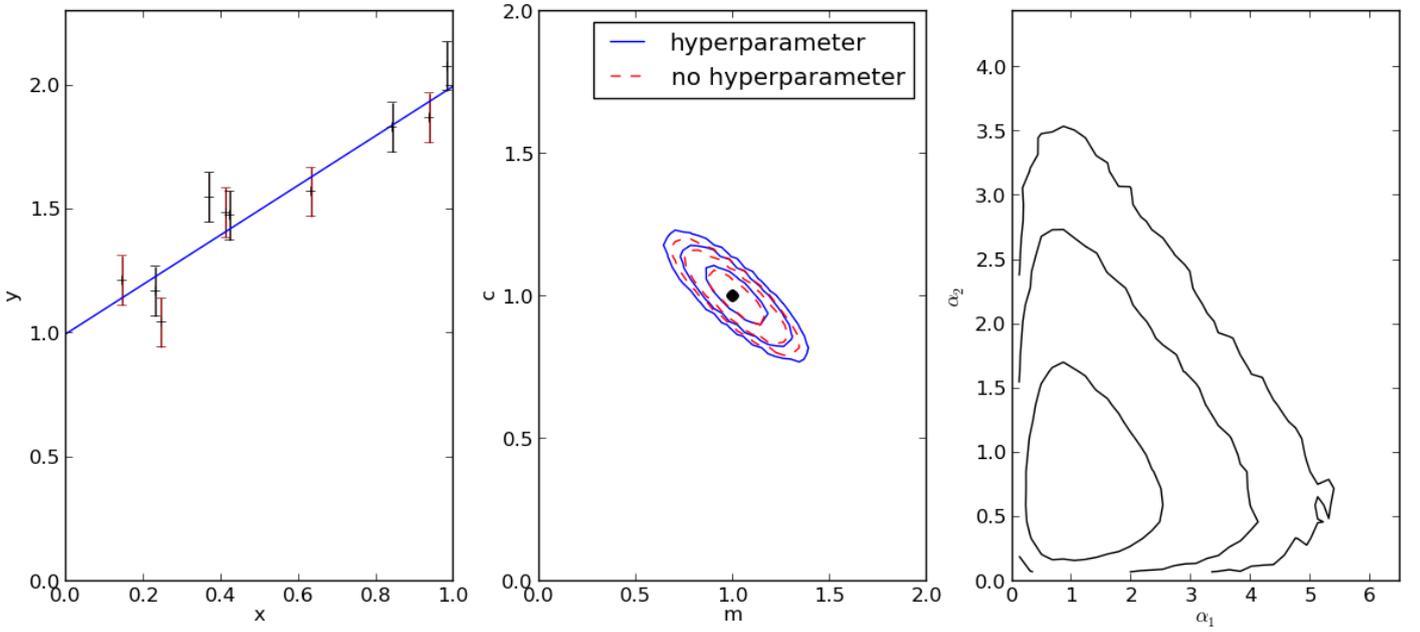


Figure 1: *Left*: the two data sets D_1 and D_2 , both drawn from a Gaussian distribution of mean $\mu = x + 1$ and rms $\sigma = 0.1$. *Middle*: the posterior distributions $\Pr(m, c|D, H_i)$ for the hyperparameter approach of HBL02 (H_1 , blue solid lines) and traditional, error-weighted approach (H_0 , red dashed lines) approach of parameter estimation. Significance contours of 68.3%, 95.4% and 99.7% are shown. A black dot indicates the true values of the model parameters $(m, c) = (1, 1)$. *Right*: the posterior distributions of the hyperparameters $\Pr(\vec{\alpha}|D, H_1)$. Values of unity correspond to no re-weighting of the data sets, as expected in this case.

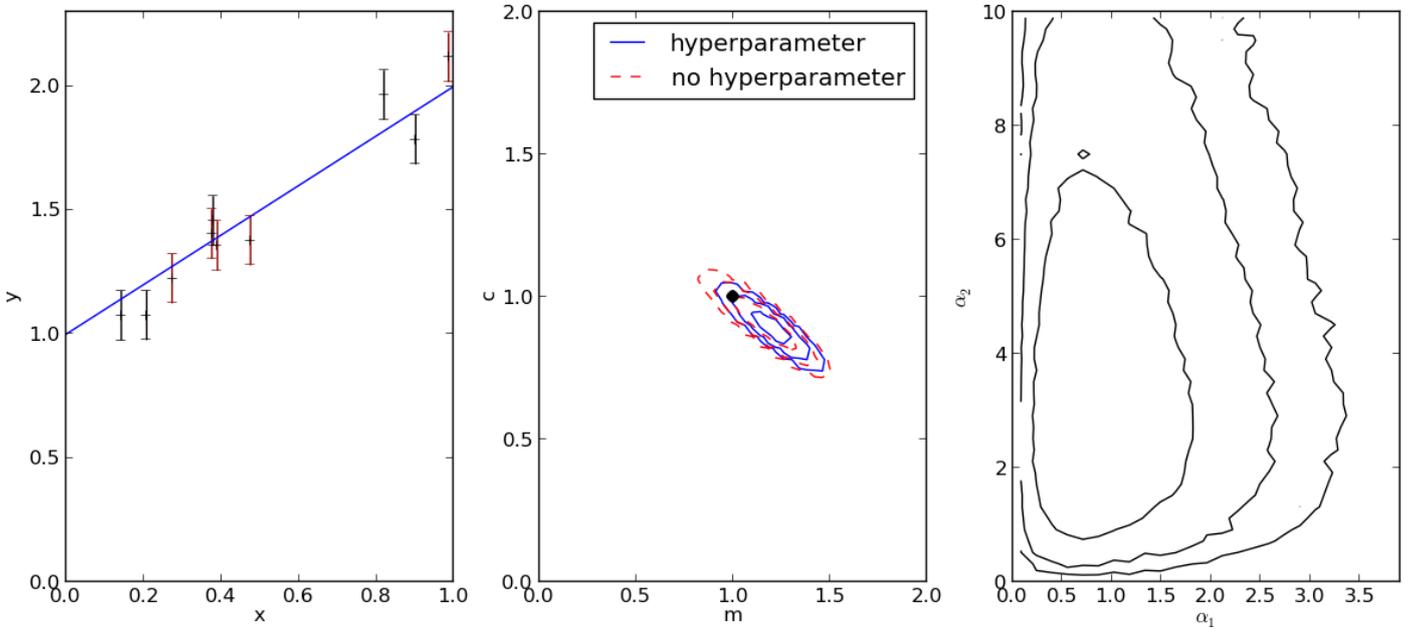


Figure 2: *Left*: same as Fig. 1 but two data sets have a correlation coefficient $\rho = C^{S_1 S_2} / \sqrt{C^{S_1} C^{S_2}} = -0.1$. *Middle*: the posterior distributions $\Pr(m, c|D, H_i)$ for the standard approach to parameter estimation (H_0 , red dashed lines) and the hyperparameter matrix method (H_2 , blue solid lines). *Right*: the posterior distribution of hyperparameters $\Pr(\vec{\alpha}|D, H_2)$.

(correctly) assume a covariance between the two data sets of $C^{S_1 S_2} = -1 \times 10^{-3}$, a tenth of the variance of the two data sets.

With reference to Fig. 2 we see, again, that the correct model parameters are consistent with both the original and hyperparameter matrix methods, but with an increased Bayesian evidence factor of 2.56. This is weak support for the hyperparameter matrix hypothesis (Table 1); however, even in this simplest of examples, we begin to see that the extended hyperparameter method provides a better fit to the correlated data sets than the non-hyperparameter method.

3.2. Inaccurate error-bars and no systematic error

3.2.1. Independent data sets

In this case the data sets D_1 and D_2 are drawn from the same distributions as in Section 3.1.1, but in the parameter estimation procedure, we assume the values of $\sigma_1 = 0.02$ (underestimated by a factor of 5) and $\sigma_2 = 0.1$ in the likelihood function. In the frequentist approach to parameter estimation, this underestimation of the noise in D_1 would over-weight its contribution to the parameter fits. With reference to Fig. 3, we see that the standard non-hyperparameter approach also underestimates the noise in the parameter fits, so the true value is well outside the 3σ confidence level. The original hyperparameter approach, H_1 , is consistent with the true parameter values at the 3.5σ level and the Bayesian evidence ratio between the two approaches is 2.5×10^4 , heavily favouring the hyperparameter approach. It should be noted that this value is very consistent with the Bayes' factor obtained by HBL02 in the same case (sec. 6.2 in HBL02).

3.2.2. Correlated data

In this case we draw the two data sets with a correlation coefficient of $\rho = C^{S_1 S_2} / \sqrt{C^{S_1} C^{S_2}} = 0.01$, so the off-diagonal component of the covariance matrix in the joint likelihood becomes $C^{S_1 S_2} = 2 \times 10^{-5}$. A comparison of the two posteriors in Fig. 4 reveals very different distributions, with the standard non-hyperparameter method being tightly constrained about the maximum but inconsistent with the true value. This is a consequence of the artificially low noise reported for the data set D_1 . As before, the Bayesian evidence strongly favours the hyperparameter matrix approach (8.3×10^{11}), and a comparison to the evidence ratio for the case with no correlation between data sets (Section 3.2.1) reveals that the hyperparameter matrix approach deals better with correlated data sets.

3.2.3. Interpretation

Let us now understand the values of the hyperparameters. In both cases of correlated and uncorrelated data sets, the joint constraint of hyperparameters reveal $\alpha_1 \simeq 0.05$, and $\alpha_2 \simeq 1$ (right panels of Figs. 3 and 4). Recalling that the hyperparameters act to rescale the error vector $\vec{x} \rightarrow \vec{x} / \sqrt{\alpha}$, and that the error reported for data set D_1 was underestimated by a factor of 5, we observe that the error recovered by the hyperparameters is $\sigma_1 / \sqrt{\alpha_1} \simeq 0.1$, close to the true value. Broadly, since α_1 is most likely less than α_2 , the average effect of the hyperparameters is to reduce the reported weight of the first data set relative to the second.

To show the importance of the generalized hyperparameter (matrix) method, we redo the analysis ignoring the data set covariance, $C^{S_1 S_2} = 0$, as would be done in the original hyperparameter method, despite the fact that the data were drawn from a correlated distribution. A comparison of the evidence for the two cases gives a Bayes' factor of 1.46, so recognizing the data sets having a covariance is a weakly-favoured hypothesis. That is, the hyperparameter method (H_1) is strongly favoured over the standard joint analysis (H_0), and the hyperparameter matrix method (H_2) is weakly favoured over the original hyperparameter method (H_1) when errors are mis-reported and correlation between data set is present. In Sec. 3.4, we will sample the correlation strength ρ and show that our hyperparameter matrix approach provides more reliable fits than the original method of HBL02.

3.3. Accurate error-bars with a systematic error

We have seen that the hyperparameter matrix approach to combining data sets provides better model fitting than the standard approach when the reported error bars differ from the true underlying error. This is true for both the case of uncorrelated and correlated data sets. In this section we explore another issue that can corrupt a joint analysis of data sets: systematic errors. We introduce a systematic error into the data set D_1 by drawing its "observed" data from a straight line with $m = 0.5$ and $c = 0.5$, while D_2 is still drawn from $m = 1$, $c = 1$.

3.3.1. Independent data sets

Figure 5 shows the two data sets D_1 and D_2 together with the underlying straight line models from which they were drawn (left panel). The systematic differences of the two models are quite apparent, which is reflected in the posterior distribution of the hyperparameter analysis $\Pr(m, c | D, H_1)$ (middle panel, blue solid lines). $\Pr(m, c | D, H_1)$ clearly indicates a bimodal distribution, recovering the underlying models $(m, c) = (0.5, 0.5)$ and $(m, c) = (1, 1)$ at the 2σ level. In contrast, the standard non-hyperparameter approach does not indicate the presence of a systematic difference between the two data sets, and fails to recover either of the models with any significance. The result of the joint constraints clearly reports a wrong parameter space, outside the input values by more than 3σ confidence level. The evidence ratio of 6.1×10^{12} heavily favours the original hyperparameter approach.

3.3.2. Correlated data sets

In this case, we compare the non-hyperparameter likelihood analysis (H_0) with the hyperparameter matrix approach when a systematic is present in one of the data sets, and there is a correlation between the two with coefficient $\rho = 0.01$ (i.e. $C^{S_1 S_2} = 1 \times 10^{-4}$). The posterior distributions recovered in this situation are shown in Fig. 6, with the similar result that the hyperparameter approach reveals a bimodal distribution, indicating the presence of a systematic difference in the data sets. The Bayes' factor comparing H_2 to H_0 is 1.5×10^{15} .

Right panels in Figures 5 and 6 show the marginalized distribution of hyperparameters α_1 and α_2 . One can see that since the

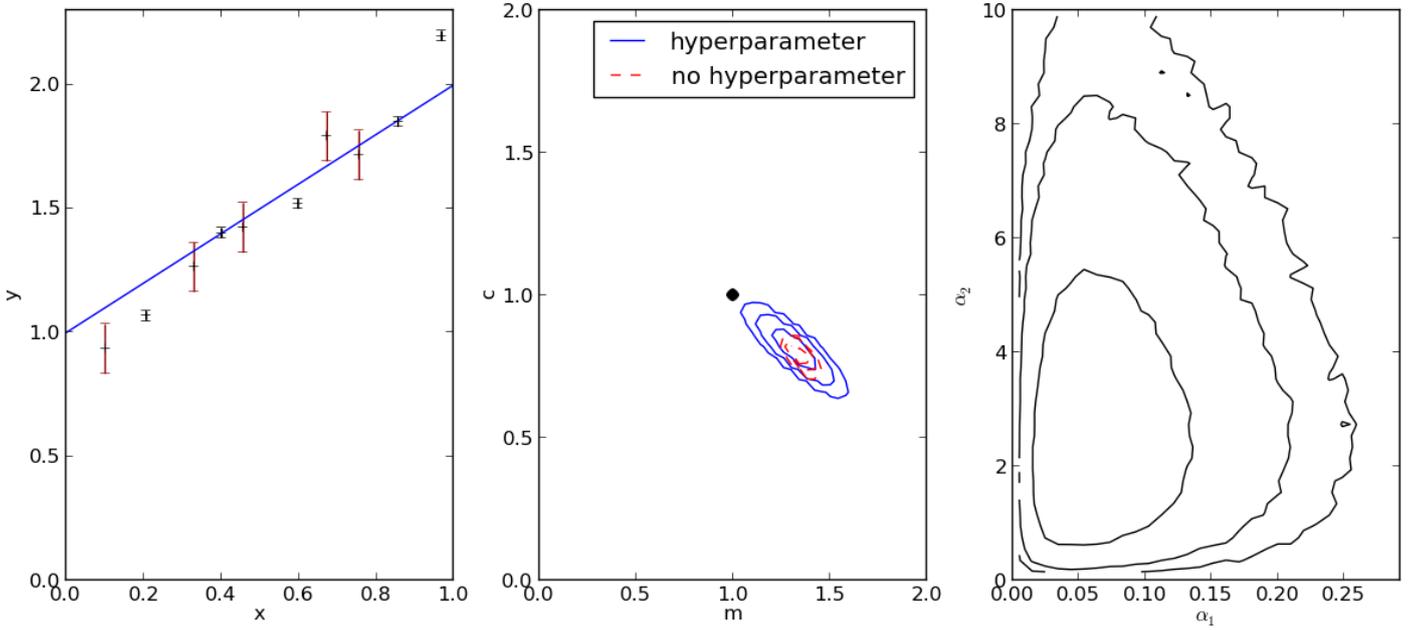


Figure 3: *Left*: same as Fig. 1, but with the reported error bars on data set D_1 underestimated by a factor of 5. *Middle*: the posteriors $\Pr(m, c|D, H_i)$ corresponding to the standard approach to parameter estimation (H_0 , red dashed lines) and the hyperparameter method (H_1 , blue solid lines). *Right*: the posterior distribution of hyperparameters $\Pr(\vec{\alpha}|D, H_1)$.

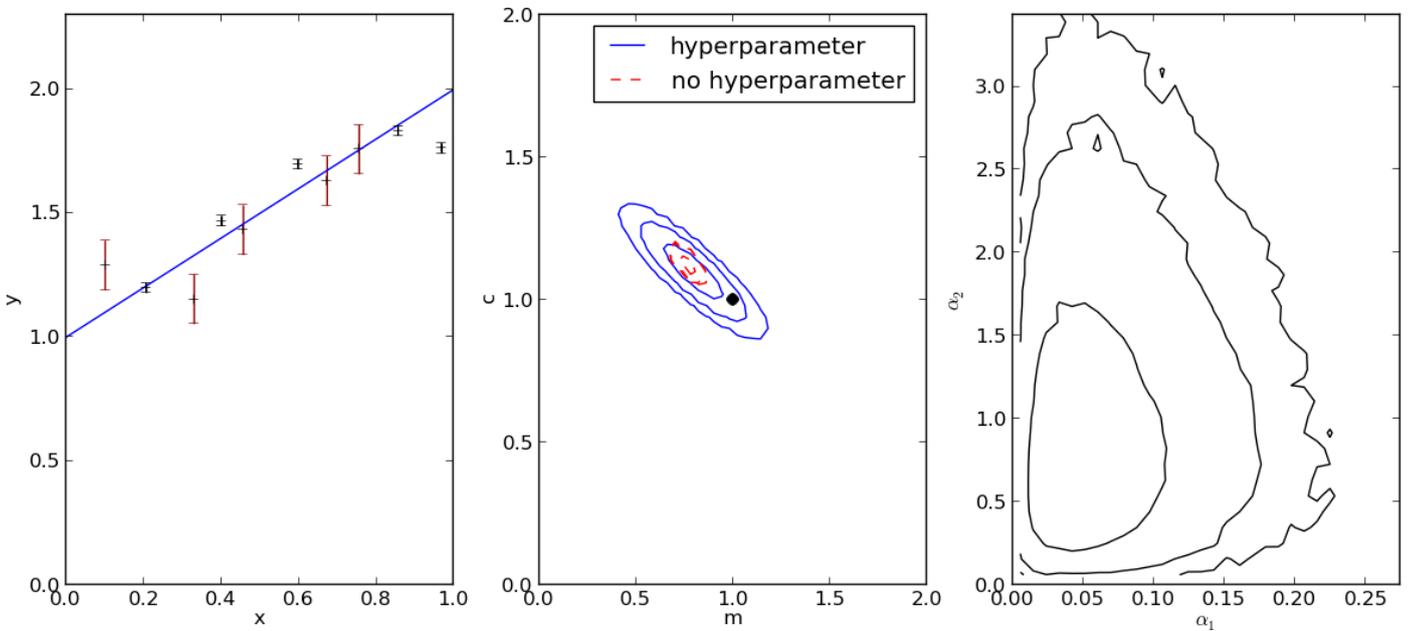


Figure 4: *Left*: same as Fig. 3 (D_1 underestimated error by a factor of 5) but with a correlation coefficient $\rho = C^{S_1 S_2} / \sqrt{C^{S_1} C^{S_2}} = 0.01$ between two data sets. *Middle*: the posteriors $\Pr(m, c|D, H_i)$ for the standard (H_0 , red dashed lines) and hyperparameter matrix (H_2 , blue solid lines) approach of parameter estimation. *Right*: the posterior $\Pr(\vec{\alpha}|D, H_2)$ distributions of the hyperparameters.

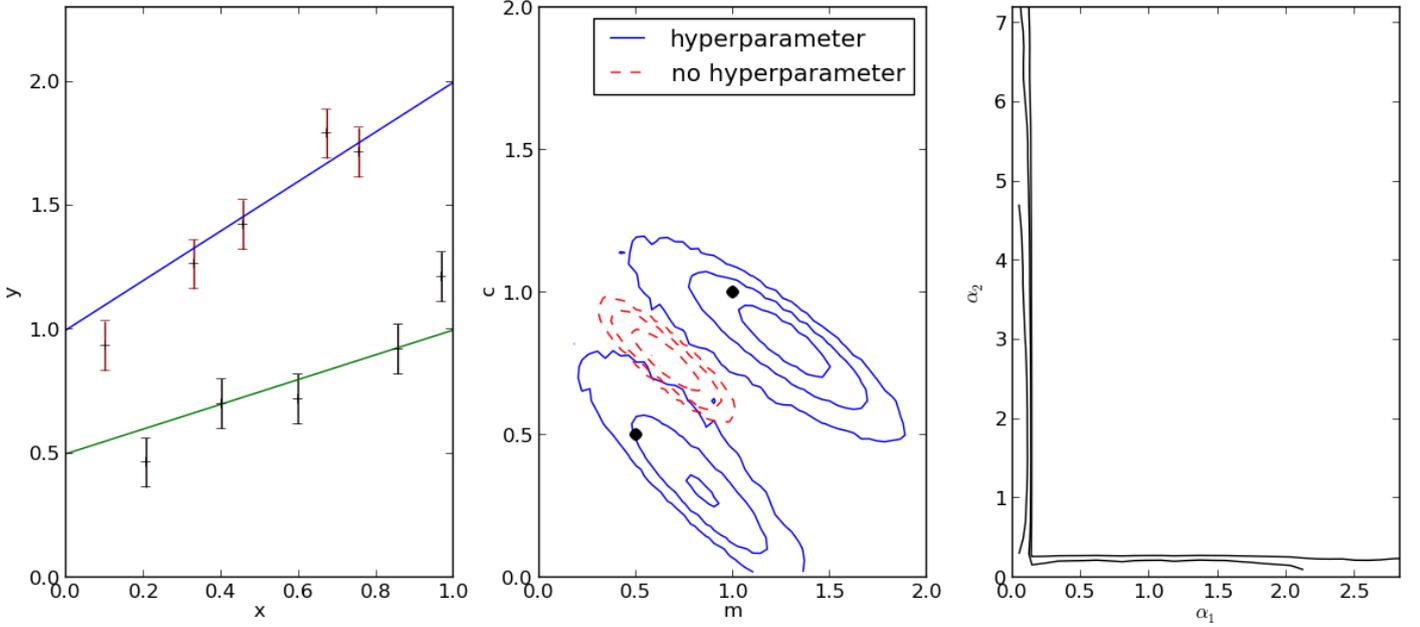


Figure 5: *Left*: the two data sets D_1 and D_2 with a systematic difference. One set is drawn from a Gaussian distribution of mean $\mu = \frac{1}{2}(x + 1)$ and rms $\sigma = 0.1$, the other from $\mu = x + 1$, $\sigma = 0.1$. *Middle*: the posteriors $\Pr(m, c | D, H_i)$ corresponding to the standard approach of parameter estimation (H_0 , red dashed contours) and the hyperparameter approach (H_1 , blue solid contours). *Right*: the posterior distribution of hyperparameters $\Pr(\vec{\alpha} | D, H_1)$.

accurate error bars	systematic error	correlated data sets	Bayes' Factor
Y	N	N	0.6
Y	N	Y	2.6
N	N	N	2.5×10^4
N	N	Y	8.3×10^{11}
Y	Y	N	6.1×10^{12}
Y	Y	Y	1.5×10^{15}

Table 2: Ratio of Bayes' evidence factors of the hyperparameter analysis to the standard non-hyperparameter analysis under varying cases of systematic errors, inaccurate error bars, and correlated data sets. The last column is Bayes' factor K (Eq. (3)). We calculate this K factor with the original hyperparameter method (H_1) over standard non-hyperparameter analysis (H_0) for uncorrelated data sets, and the hyperparameter matrix method (H_2) over standard Gaussian likelihood analysis (H_0) for correlated data sets. Note that throughout the calculation we adopt the exponential prior on hyperparameters ($\Pr(\alpha) = \exp(-\alpha)$).

two data sets have systematic errors, the constraints on hyperparameters have two branches. In each branch, one parameter takes an ordinary value while the other is close to zero. This is a consequence of the presence of a systematic, since the error is reduced by ignoring one of the data sets entirely instead of combining them jointly.

3.4. The improvement on the original hyper-parameter method

The hyperparameter matrix method we propose here is the most general method which can be used to combine arbitrary number of multi-correlated experimental data. This greatly breaks up the limitation of the original hyperparameter method (Lahav et al. (2000) and HBL02) which can only deal with multiple independent data sets. It is always important, to include all

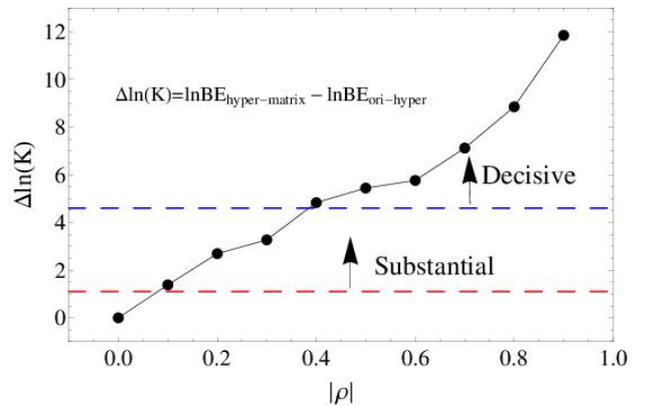


Figure 7: The difference between logarithmic Bayesian evidence (the factor BE is defined as Eq. (2)) as a function of the correlation strength ρ . ρ is sampled from 0 to -0.9 with each step -0.1 . $\Delta \ln BE$ is equal to the value of the Bayesian evidence with our hyperparameter matrix method to consider full covariance between data sets, minus the value of Bayesian evidence from the original hyperparameter method (ignore the correlation between data sets). For the specific experiment, please refer to Sec. 3.4.

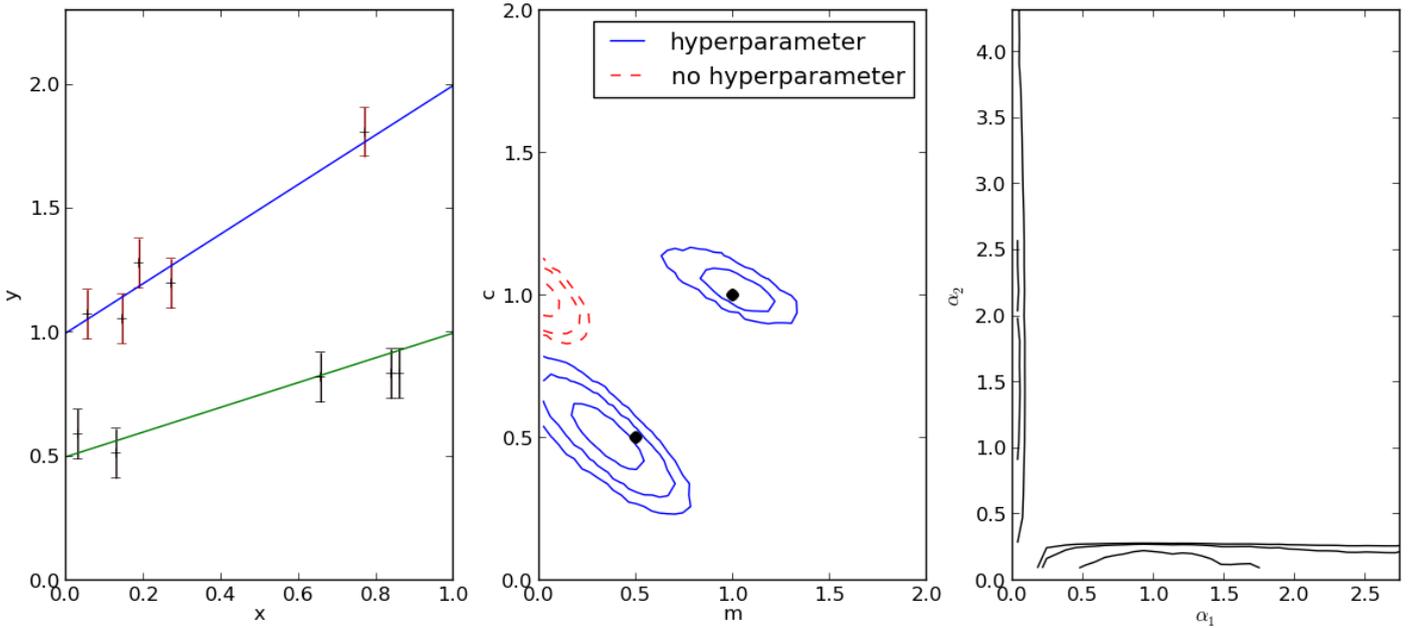


Figure 6: *Left*: same as Fig. 5 but two data sets have a correlation coefficient $\rho = C^{S_1 S_2} / \sqrt{C^{S_1} C^{S_2}} = 0.01$. *Middle*: the posteriors $\Pr(m, c | D, H_0)$ corresponding to the standard approach of parameter estimation (H_0 , red dashed lines) and the hyperparameter matrix method (H_2 , blue solid lines). *Right*: the posterior distribution $\Pr(\alpha | D, H_2)$ of the hyperparameter matrix method.

of the correlation information between data sets to obtain correct parameter values and justify the goodness of fit.

To see the importance of our method, we design an illustrative experiment to demonstrate this. We generate two data sets with $N = 5$. For each data set, we generate the samples with mean 1.0 and 0.0 with Gaussian error $\sigma = 0.1$ but correlated between the two data sets. We take the correlation strength ρ as 0.0, -0.1, -0.2, ..., -0.9.

Then we use these correlated data sets to do a parameter estimation. We first use our hyperparameter matrix method, which considers the full covariance matrix between two data sets. Then in order to check the behaviour of the original hyperparameter method, we *ignore* the correlation part of the two experiments and treat them as individual data sets. We calculate the Bayesian evidence value (Eq. (2)) for both cases, and obtain the difference between the two Bayesian evidence (BE) values.

In Fig. 7, we plot the difference between BE value for our hyperparameter matrix method and for the original hyperparameter method. First, one can see that when $\rho = 0.0$, the two methods are the same one so $\Delta \ln(K) = 0$. But as the correlation strength increases, the $\Delta \ln(K)$ increases as well, indicating that the hyperparameter matrix method provides better and better fits than the original hyperparameter method. This can be understood as the danger of ignoring correlation between data sets, since the model becomes inadequate to fit the data if the correlation is not included. In Fig. 7, one can see that if $|\rho| > 0.1$, the Bayes' factor becomes "Substantial", and if $|\rho| > 0.4$, the Bayes' factor becomes "Decisive". This strongly indicates that when combining multiple correlated data sets, it is very necessary to use our hyperparameter matrix method rather than the original hyperparameter method.

4. Conclusion

In this paper we have reviewed the standard approach to parameter estimation when there are multiple data sets. This is an important aspect to most scientific enquiries, where multiple experiments are attempting to observe the same quantity. In the context of a Bayesian analysis, the data can also be used for model selection and tests of the null hypothesis. We reviewed the original hyperparameter method of HBL02 for combining independent data sets, showing how it can overcome inaccurate error bars and systematic differences between multiple data sets.

Here we developed the hyperparameter matrix method for the case of correlated data sets, and we have shown that it is a preferred model to the standard non-hyperparameter approach of parameter estimation. We rigorously prove that the hyperparameter matrix likelihood can be greatly simplified and be easily implemented. From this form of the likelihood, we can recover the simple case of no hyperparameters where all of the data sets have equal weights. As well, the original hyperparameter approach is recovered in the limit of no inter-data set covariance ($C^{S_i S_j} = 0$ if $i \neq j$), so our likelihood function provides a generalized form which covers hyperparameter and non-hyperparameter analysis, as well as correlated and uncorrelated data sets.

We test this statistical model by fitting two data sets to a straight line, and looked at the consequences of mis-reported error bars, as well as systematic differences between correlated data sets. In all cases, with the assistance of Bayesian evidence, we find that the hyperparameter matrix method is heavily favoured over the traditional joint analysis. By using an illustrative example to calculate the difference of Bayesian

evidence value between the hyperparameter matrix method, and the original hyperparameter method, we demonstrate that the Bayes' factor becomes very substantial (decisive) if $|\rho|$ is greater than 0.1 (0.4). This suggests that for the case where two experiments are strongly correlated, our hyperparameter matrix method is heavily favoured over the original hyperparameter method.

The method proposed here can be used in a variety of astrophysical systems. In the context of cosmology, when cosmic variance is a common component to all large-scale observations, the data sets drawn from the same underlying density or temperature field will be correlated to some degree. For instance, in the study of CMB where multiple data sets drawn from the same region of the sky are combined (such as *Planck* (Planck results XVI., 2013), *WMAP* (Hinshaw et al., 2013), SPT (Hou et al., 2012) and ACT (Sievers et al., 2013)), it is necessary to consider the correlation between data sets since they follow the same underlying temperature distribution. Therefore our method can be an objective metric to quantify the posterior distribution of cosmological parameters estimated from the CMB. In addition, in the analysis of the galaxy redshift surveys for cosmic density and velocity fields, when combining two surveys data drawn from the similar cosmic volume, the cosmic variance between different data sets should also be considered as a part of the total covariance matrix since they all follow the same underlying matter distribution. In the future survey of 21 cm, if two or more surveys sample the neutral hydrogen in the same (or close) cosmic volume, the correlation between surveys should also be considered when combining data sets. In this sense, our hyperparameter matrix method provides an objective metric to quantify the probability distribution of the parameters of interest when multiple data sets are combined.

In summary, when combining correlated data sets, the hyperparameter matrix method can provide an unbiased and objective approach that can wisely detect and down-weight any unaccounted experimental errors or systematic errors, in this way it provides the most robust and reliable constraints on astrophysical parameters.

5. Acknowledgements

We would like to thank Chris Blake, Andrew Johnson, Douglas Scott and Jasper Wall for helpful discussions. Y.Z.M. is supported by a CITA National Fellowship. This research is supported by the Natural Science and Engineering Research Council of Canada.

Appendix A. Theorem: positive-definiteness of the hyperparameter covariance matrix

The generalized form of the likelihood function for the hyperparameter analysis in the presence of correlated data sets (Eq. (20)) must satisfy several properties in order to serve as a probability density function. In particular, the generalized hyperparameter covariance matrix $C = P \odot \tilde{C}$ (Eq. (17)) must have positive determinant, and must be invertible. However, since

the matrix P is a function of the hyperparameters $\vec{\alpha}$ which, in principle, vary from zero to infinity, the positive definiteness and invertibility of C are not immediately clear.

The following theorem guarantees the feasibility of inverting the total covariance matrix C , and the positive definiteness of the determinant.

Theorem: The likelihood function of combining N correlated data sets with hyperparameter matrix, i.e. Eq. (20) is equivalent to

$$\Pr(D|\vec{\theta}, \vec{\alpha}) = \left[\prod_{i=1}^N \left(\frac{\alpha_i}{2\pi} \right)^{n_i/2} \right] \frac{1}{\sqrt{\det \tilde{C}}} \exp \left(-\frac{1}{2} \vec{x}^T (\hat{P} \odot \tilde{C}^{-1}) \vec{x} \right), \quad (\text{A.1})$$

where n_i is the dimension of the i th data set, \tilde{C} is the covariance matrix between N data sets without the inclusion of hyperparameter (Eq. (7)), \odot is the element-wise product (same as Eq. (19)), and \hat{P} is the ‘‘Hadamard inverse’’ of the P matrix (see Appendix B).

We first prove the inverse relation,

$$\begin{aligned} C^{-1} &\equiv (P \odot \tilde{C})^{-1} \\ &= \hat{P} \odot \tilde{C}^{-1}. \end{aligned} \quad (\text{A.2})$$

Proof. (1) Let us multiply matrices $(P \odot \tilde{C})$ and $(\hat{P} \odot \tilde{C}^{-1})$, then take the block element (i, j) of the matrix, i.e. ‘‘ i, j, k ’’ are the block element which can take any value between $(1, \dots, N)$

$$\begin{aligned} &[(P \odot \tilde{C})(\hat{P} \odot \tilde{C}^{-1})]_{ij} \\ &= \sum_k (P \odot \tilde{C})_{ik} (\hat{P} \odot \tilde{C}^{-1})_{kj} \\ &= \sum_k (\tilde{C}_{ik} * (\alpha_i \alpha_k)^{-1/2}) (\tilde{C}_{kj}^{-1} * (\alpha_k \alpha_j)^{1/2}) \\ &= \sum_k (\tilde{C}_{ik} \tilde{C}_{kj}^{-1}) (\alpha_j / \alpha_i)^{1/2} \\ &= (\delta_{ij}) I_{n_i \times n_j} (\alpha_j / \alpha_i)^{1/2} \\ &= (\delta_{ij}) I_{n_i \times n_i}, \end{aligned} \quad (\text{A.3})$$

where in the second step, we use the property of block matrix product. The final line of Eq. (A.3) indicates that, only if $i = j$, the product is an $n_i \times n_i$ identity matrix, otherwise it is all zeros. Thus we prove the inverse relation (Eq. (A.2)). \square

Next, let us prove the determinant relation

$$\det(C) = \det(P \odot \tilde{C}) = \det \tilde{C} * \left(\prod_{i=1}^N \alpha_i^{-n_i} \right), \quad (\text{A.4})$$

where C is given by Eq. (17), \tilde{C} is given by (Eq. (7)) and n_i is the dimension of the i th block matrix.

Proof. (2) In Appendix C, we have proved that a matrix of type \tilde{C} (7) follows the determinant Eqs. (C.3)-(C.7). We now use Eqs. (C.3)-(C.7) to prove Eq. (A.4). From Eq. (C.3), we have

$$\begin{aligned} \det(C) &= \prod_{k=1}^N \det(\alpha_{kk}^{(N-k)}) \\ &= \det(\alpha_{11}^{(N-1)}) * \det(\alpha_{22}^{(N-2)}) * \dots \\ &* \det(\alpha_{N-1, N-1}^{(1)}) * \det(\alpha_{NN}^{(0)}), \end{aligned} \quad (\text{A.5})$$

where the α matrix stands for Eqs. (C.4)-(C.7) but replacing A matrix for C matrix.

We then apply the same equation for the covariance matrix \tilde{C}

$$\begin{aligned}\det \tilde{C} &= \prod_{k=1}^N \det(\tilde{\alpha}_{kk}^{(N-k)}) \\ &= \det(\tilde{\alpha}_{11}^{(N-1)}) * \det(\tilde{\alpha}_{22}^{(N-2)}) * \dots \\ &* \det(\tilde{\alpha}_{N-1,N-1}^{(1)}) * \det(\tilde{\alpha}_{NN}^{(0)}),\end{aligned}\quad (\text{A.6})$$

where the $\tilde{\alpha}$ matrix stands for Eqs. (C.4)-(C.7) but replacing A matrix with \tilde{C} matrix.

Now we compare the last terms in Eqs. (A.5) and (A.6). Since α_{NN}^0 is indeed C_{NN} as given by Eq. (C.4), we have

$$\begin{aligned}\det(\alpha_{NN}^{(0)}) &= \det(\alpha_N^{-1} \tilde{C}_{NN}) \\ &= \alpha_N^{-n_N} \det(\tilde{C}_{NN}) \\ &= \alpha_N^{-n_N} \det(\tilde{\alpha}_{NN}^{(0)}).\end{aligned}\quad (\text{A.7})$$

We then calculate the i th term; following Eq. (C.4), we have

$$\begin{aligned}\det(\alpha_{ii}^{(N-i)}) &= C_{ii} - \sigma_{i,i+1} (C_{N-i})^{-1} \eta_{i+1,i} \\ &= C_{ii} - \sum_{m=i+1}^N \sum_{n=i+1}^N C_{im} (C)_{mn}^{-1} C_{ni}.\end{aligned}\quad (\text{A.8})$$

By using Eq. (A.2), we obtain

$$(C^{-1})_{mn} = (\alpha_m \alpha_n)^{1/2} (\tilde{C}^{-1})_{mn}.\quad (\text{A.9})$$

Therefore we have

$$\begin{aligned}\det(\alpha_{ii}^{(N-i)}) &= \alpha_i^{-1} \tilde{C}_{ii} - \sum_{m=i+1}^N \sum_{n=i+1}^N (\alpha_m \alpha_n)^{-1/2} \\ &\times \tilde{C}_{im} (\alpha_m \alpha_n)^{1/2} (C)_{mn}^{-1} (\alpha_i \alpha_n)^{-1/2} C_{ni} \\ &= \alpha_i^{-1} \left(\tilde{C}_{ii} - \sum_{m=i+1}^N \sum_{n=i+1}^N \tilde{C}_{im} (\tilde{C})_{mn}^{-1} \tilde{C}_{ni} \right) \\ &= \alpha_i^{-1} \det(\tilde{\alpha}_{ii}^{(N-i)}).\end{aligned}\quad (\text{A.10})$$

Thus, by mathematical induction, we have proved that all of the terms in Eqs. (A.5) and (A.6) follow Eq. (A.10). Therefore the relationship between Eqs. (A.5) and (A.6) is

$$\det(C) = \det(\tilde{C}) * \left(\prod_{i=1}^N \alpha_i^{-n_i} \right),\quad (\text{A.11})$$

i.e. we have proved Eq. (A.4). \square

Combining Proofs (1) and (2), we have shown that, in general, when combining multiple correlated data sets with hyperparameters, the inverse and determinant of the covariance matrix follow Eqs. (A.2) and (A.4). Therefore the likelihood function for combined correlated data sets is Eq. (A.1).

Equation (A.1) greatly simplifies the computation of hyperparameter likelihood, since one can always calculate the covariance matrix for correlated data sets \tilde{C} and then use ‘‘element-wise’’ product \odot to calculate the covariance matrix with hyperparameters, and then numerically solve for the maximum likelihood solution.

Appendix B. Hadamard product and inverse

The Hadamard product is the element-wise product of any two matrices with the same dimension. If A and B are the two matrices with the same dimension $m \times n$, the Hadamard product $A \circ B$ is a matrix with the same dimension with element (i, j) equal to

$$(A \circ B)_{ij} = A_{ij} \cdot B_{ij}.\quad (\text{B.1})$$

The Hadamard inverse is an inverse operation which requires that each element of the matrix is nonzero, so that each element of the Hadamard inverse matrix is

$$\hat{A}_{ij} = A_{ij}^{-1}.\quad (\text{B.2})$$

Here we use a hat to denote the Hadamard inverse. Therefore the Hadamard product of an $m \times n$ matrix and its Hadamard inverse becomes a unit matrix where all elements are equal to one, i.e.

$$A \circ \hat{A} = (J)_{m \times n}.\quad (\text{B.3})$$

Appendix C. A lemma for determinant

We will use the following Lemma to prove the determinant relation of the covariance matrix of hyperparameter likelihood, Eq. (A.4).

Let A be an $(N_t \times N_t)$ real or complex matrix, which is partitioned into $N \times N$ blocks, each of size is $n_i \times n_j$, which satisfies

$$\sum_{i=1}^N n_i = N_t.\quad (\text{C.1})$$

$$A = \begin{pmatrix} (A_{11})_{n_1 \times n_1} & (A_{12})_{n_1 \times n_2} & \dots & (A_{1N})_{n_1 \times n_N} \\ (A_{12})_{n_2 \times n_1}^T & (A_{22})_{n_2 \times n_2} & \dots & (A_{2N})_{n_2 \times n_N} \\ \dots & \dots & \dots & \dots \\ (A_{1N})_{n_N \times n_1}^T & (A_{2N})_{n_N \times n_2}^T & \dots & (A_{NN})_{n_N \times n_N} \end{pmatrix}.\quad (\text{C.2})$$

The determinant of A is given by

$$\det A = \prod_{k=1}^N \det(\alpha_{kk}^{(N-k)}),\quad (\text{C.3})$$

where $\alpha^{(k)}$ is defined as

$$\begin{aligned}\alpha_{ij}^{(0)} &= A_{ij} \\ \alpha_{ij}^{(k)} &= A_{ij} - \sigma_{i,N-k+1} (\bar{A}_k)^{-1} \eta_{N-k+1,j}, (k \geq 1),\end{aligned}\quad (\text{C.4})$$

where vectors σ_{ij}^T and η_{ij} are defined as

$$\sigma_{ij} = (A_{ij}, A_{i,j+1}, \dots, A_{i,N}),\quad (\text{C.5})$$

$$\eta_{ij} = (A_{ij}, A_{i+1,j}, \dots, A_{N,j})^T,\quad (\text{C.6})$$

and \bar{A}_k is defined as

$$\bar{A}_k = \begin{pmatrix} A_{N-k+1,N-k+1} & A_{N-k+1,N-k+2} & \dots & A_{N-k+1,N} \\ A_{N-k+2,N-k+1} & A_{N-k+2,N-k+2} & \dots & A_{N-k+2,N} \\ \dots & \dots & \dots & \dots \\ A_{N,N-k+1} & A_{N,N-k+2} & \dots & A_{N,N} \end{pmatrix}.\quad (\text{C.7})$$

A particular case of this lemma, where each block matrix has the same dimension $n \times n$, is shown as a theorem in Powell (2011). Here we extend the theorem shown in Powell (2011) to a more general case where each diagonal block matrix may have a different size, so the off-diagonal matrix can be a rectangular matrix.

Proof. We start from the simplest case, where $N = 2$, i.e. A is a 2×2 symmetric block matrix

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{12}^T & A_{22} \end{pmatrix}, \quad (\text{C.8})$$

where A_{11} and A_{22} are $p \times p$ and $q \times q$ semi-positive definite symmetric matrix respectively, and A_{12} is a $p \times q$ matrix. The determinant of A is

$$\begin{aligned} \det A &= \det(A_{11} - A_{12}A_{22}^{-1}A_{12}^T) \det(A_{22}) \\ &= \det(A_{22} - A_{12}A_{11}^{-1}A_{12}^T) \det(A_{11}). \end{aligned} \quad (\text{C.9})$$

We can immediately check that this is indeed the simplest case for Eqs. (C.3)-(C.7) where $N = 2$. Since if $N = 2$, Eq. (C.4) gives $\det A = \det(\alpha_{11}^{(1)}) \det(\alpha_{22}^{(0)})$, where $\alpha_{22}^{(0)} = A_{22}$, and $\alpha_{11}^{(1)} = A_{11} - A_{12}A_{22}^{-1}A_{12}^T$, which is exactly Eq. (C.9).

Now we can use Eq. (C.9) for the $N = 2$ case to inductively derive general equations (C.3)-(C.7). Let us treat matrix (C.2) as a 2-by-2 matrix, where all of the matrices $A_{22}, A_{33}, \dots, A_{NN}$ are grouped into a big matrix \tilde{A}_{22} :

$$A = \begin{pmatrix} A_{11} & \tilde{A}_{12} \\ \tilde{A}_{12}^T & \tilde{A}_{22} \end{pmatrix}, \quad (\text{C.10})$$

where

$$\tilde{A}_{22} = \begin{pmatrix} A_{22} & A_{23} & \dots & A_{2N} \\ A_{23}^T & A_{33} & \dots & A_{3N} \\ \dots & \dots & \dots & \dots \\ A_{2N}^T & A_{3N}^T & \dots & A_{NN} \end{pmatrix}, \quad (\text{C.11})$$

is exactly \bar{A}_{N-1} (Eq. (C.7)), and

$$\tilde{A}_{12} = \begin{pmatrix} A_{12} & A_{13} & \dots & A_{1N} \end{pmatrix}, \quad (\text{C.12})$$

is exactly the definition of σ_{12} (Eq. (C.5)). In addition,

$$\tilde{A}_{12}^T = \begin{pmatrix} A_{21} & A_{31} & \dots & A_{N1} \end{pmatrix}^T, \quad (\text{C.13})$$

is exactly η_{21} (Eq. (C.6)). Now applying the second line of Eq. (C.9) to this matrix, one has

$$\begin{aligned} \det(A) &= \det(\tilde{A}_{22}) * \det(A_{11} - \tilde{A}_{12}\tilde{A}_{22}^{-1}\tilde{A}_{12}^T) \\ &= \det(\tilde{A}_{22}) * \det(A_{11} - \sigma_{12}\bar{A}_{N-1}^{-1}\eta_{21}). \end{aligned} \quad (\text{C.14})$$

Now proceeding to $\det(\tilde{A}_{22})$, again, \tilde{A}_{22} can be separated into two big matrices as

$$\tilde{A}_{22} = \begin{pmatrix} A_{22} & \tilde{A}_{23} \\ \tilde{A}_{23}^T & \tilde{A}_{33} \end{pmatrix}, \quad (\text{C.15})$$

where

$$\tilde{A}_{33} = \begin{pmatrix} A_{33} & A_{34} & \dots & A_{3N} \\ A_{34}^T & A_{44} & \dots & A_{4N} \\ \dots & \dots & \dots & \dots \\ A_{3N}^T & A_{4N}^T & \dots & A_{NN} \end{pmatrix} = \bar{A}_{N-2}, \quad (\text{C.16})$$

and

$$\begin{aligned} \tilde{A}_{23} &= \begin{pmatrix} A_{23} & A_{24} & \dots & A_{2N} \end{pmatrix} = \sigma_{23}, \\ \tilde{A}_{23}^T &= \begin{pmatrix} A_{32} & A_{42} & \dots & A_{N2} \end{pmatrix}^T = \eta_{32}, \end{aligned} \quad (\text{C.17})$$

therefore

$$\begin{aligned} \det(\tilde{A}_{22}) &= \det(\tilde{A}_{33}) * \det(A_{22} - \tilde{A}_{23}\tilde{A}_{33}^{-1}\tilde{A}_{23}^T) \\ &= \det(\tilde{A}_{33}) * \det(A_{22} - \sigma_{23}\bar{A}_{N-2}^{-1}\eta_{32}), \end{aligned} \quad (\text{C.18})$$

so combining Eqs. (C.18) and (C.14), we have

$$\begin{aligned} \det(A) &= \det(\tilde{A}_{33}) * \det(A_{22} - \sigma_{23}\bar{A}_{N-2}^{-1}\eta_{32}) \\ &\quad * \det(A_{11} - \sigma_{12}\bar{A}_{N-1}^{-1}\eta_{21}). \end{aligned} \quad (\text{C.19})$$

Repeating this operation until breaking down the first term, one can eventually reach A_{NN} , therefore the determinant of A is

$$\begin{aligned} \det(A) &= \det(A_{NN}) \\ &\quad * \det(A_{N-1,N-1} - A_{N-1,N}\bar{A}_1^{-1}A_{N,N-1}) \\ &\quad * \dots * \det(A_{22} - \sigma_{23}\bar{A}_{N-2}^{-1}\eta_{32}) \\ &\quad * \det(A_{11} - \sigma_{12}\bar{A}_{N-1}^{-1}\eta_{21}). \end{aligned} \quad (\text{C.20})$$

By comparing the brackets in Eq. (C.20) with Eq. (C.4), one can find that each term is exactly the same, therefore the determinant is given by Eq. (C.3). \square

References

- Ade P. A. R. et al., 2013a. Planck results XV., arXiv: 1303.5075 [astro-ph.CO].
Ade P. A. R. et al., 2013b. Planck results XVI., arXiv: 1303.5076 [astro-ph.CO].
Cheng C., Huang Q. G., & Ma Y. Z., 2013, JCAP, 07, 018
Erdogdu P., Ettori S., & Lahav O., 2003, MNRAS, 340, 573
Godwin P., & Lynden-Bell D., 1987, MNRAS, 229, 7
Hinshaw G. et al., 2013, ApJS, 208, 19
Hobson M. P., Bridle S. L., & Lahav O., 2002, MNRAS, 335, 377 (HBL02)
Host O., & Hansen S. H., 2011, ApJ, 736, 52
Hou Z. et al., arXiv: 1212.6267 [astro-ph.CO].
Jeffreys H., *The Theory of Probability*, Oxford University Press, 1961.
Lahav O., Bridle S. L., Hobson M. P., Lasenby A. N., & Sodre L., 2000, MNRAS, 315, 45
Ma Y. Z., Zhao W., & Brown M. L., 2010, JCAP, 1010, 007
Ma Y. Z., Branchini E., & Scott D., 2012, MNRAS, 425, 2880
Ma Y. Z., & Scott D., 2013, MNRAS, 428, 2017
Magoulas C. et al., 2012, MNRAS, 427, 245
Nuza S. E. et al., 2013, MNRAS, 432, 743
Powell P. D., arXiv: 1112.4379, [math.RA]
Patil A., Huard D. & Fonnesbeck C.J., 2010, Journal of Statistical Software, 35, 4
Press W.H., 1996, in *Unsolved Problems in Astrophysics*, Proc. Conference in Honour of John Bahcall, ed. J.P. Ostriker. Princeton University Press
Sievers J. L. et al., 2013, JCAP, 10, 060
Skilling J., in *Nested Sampling*, AIP Conference Proc., 2004, **735**, 395.
Kass, R. E., & Raftery, A. E., 1995, Journal of the American Statistical Association, 90, 773
Raftery A. E., Newton M. A., Satagopan J. M., & Krivitsky P. N., 2007, *Bayesian Statistics*, 8, 1-45.
Weinberg M. D., arXiv: 0911.1777, [astro-ph.IM]