Bayesian model selection applied to linear regressions with weighted data

Bayesian model selection applied to linear regressions with weighted data

G Mana, E Massa and C P Sasso

INRIM - Istituto Nazionale di Ricerca Metrologica, Str. delle Cacce 91, 10135 Torino, Italy

E-mail: g.mana@inrim.it

Abstract. This paper uses the Bayesian model-selection to find the linear regression that is most supported by the data. It uses a hierarchical model to improve and to extend to weighted data a previous investigation. As an application example, it shows how the results can be used to investigate the consistency of measurements carried out to determine the central second-moment of the angular power-spectrum of a laser beam.

Submitted to: *Metrologia*

PACS numbers: 02.50.Cw, 02.50.Tt,07.05.Kf

1. Introduction

A problem in linear regression is to test the goodness of fit and to select the model most supported by the data. If the data explanation is uncertain, for instance, when building calibration curves, the problem is to chose how many basis functions to include in the regression. If the data uncertainties are underestimated, different hypotheses are possible, for instance, in the presence of random or fixed effects [1, 2, 3].

To make meaningful assessments of the odds of competing explanations one must assign a prior distribution to the regression parameters in such a way that the posterior probabilities of regressions that differ only by their parametrisation are the same. This requirement was not considered in previous works [3, 4, 5, 6]. These prior distributions (named after H. Jeffreys) are proportional to the invariant volume-element of the model manifold [7, 8], where the regression parameters are the manifold coordinates and whose metric is the Fisher information [9, 10, 11, 12].

A difficulty impedes to pursue this line of thought: Jeffrey's distribution of the regression parameters is improper. To go around this problem, we build on previous proposals [4, 6] and use a hierarchical model to improve and extend to weighted and correlated data the results given in [6].

Section 2 frames the study and introduces the representations of the data and design-matrix that make the algebra the simplest. The Bayesian procedure to compare competing models is shortly summarised in section 3 and applied to the problem at hand in section 4. Eventually, a numerical example and an application to experimental data demonstrate the practical use of model selection in data analysis.

2. Linear regressions

Let us explain N observations, $\boldsymbol{y} = \{y_1, y_2, ... y_N\}^T$, by a set of competing (mutually exclusive) linear models

$$\boldsymbol{y} = W_l \boldsymbol{b} + \boldsymbol{u}, \tag{1}$$

where l is the regression order, $\boldsymbol{u}$ are zero-mean errors having variance-covariance matrix Σ_u, $W_l = \{\boldsymbol{w}_1, \boldsymbol{w}_2, ... \boldsymbol{w}_l\}$ are full rank $N \times l$ design matrices, $\boldsymbol{w}_i$ are the model basis-vectors, and $\boldsymbol{b} = \{b_1, b_2, ... b_l\}^T$ are model parameters. The problem is to find the model most supported by the data. The basis functions, whose discrete samples $\boldsymbol{w}_i$ form the basis vectors, may be polynomials but, in general, they are any set of linearly independent functions. For the sake of simplicity, we indexed the design matrices W_l by the regression order, but competing regressions having the same order are not excluded.

In order to avoid awkward algebra, it is convenient to normalize the data in such a way that they have unit variance and are uncorrelated. This corresponds to use an orthonormal basis in the $\boldsymbol{y}$ space and it is done by the transformations $\boldsymbol{h} = U^{-1}\boldsymbol{y}$, $V_l = U^{-1}W_l$, and $\boldsymbol{\epsilon} = U^{-1}\boldsymbol{u}$, where the lower triangular matrix U is the Cholesky factor of Σ_u, that is, $\Sigma_u = UU^T$ [13]. Hence, we rewrite the data model (1) as $\boldsymbol{h} = V_l\boldsymbol{b} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ are zero-mean uncorrelated errors having unit variance-covariance matrix, $V_l = \{\boldsymbol{v}_1, \boldsymbol{v}_2, ... \boldsymbol{v}_l\}$, and $\boldsymbol{v}_i$ are the new design matrices and basis vectors.

Since we are not interested in any specific parametrisation, a further simplification is obtained by using an orthonormal basis in the subspace spanned by $V_l\boldsymbol{b}$, which is done by orthonormalizing the $\{\boldsymbol{v}_1, \boldsymbol{v}_2, ... \boldsymbol{v}_l\}$ set and by scaling the parameters according to the decomposition $V_l = Q_l R$, where Q_l is column-orthonormal, $Q_l^T Q_l = \mathbb{1}$, $\mathbb{1}$ is the unit matrix, and R is upper triangular [13].

Eventually, the maximum-entropy distribution of the normalized data, constrained by $\boldsymbol{h} = Q_l\boldsymbol{a} + \boldsymbol{\epsilon}$ and $\Sigma_h = \mathbb{1}$ is

$$L(\boldsymbol{h}|\boldsymbol{a}, l) = \sqrt{\frac{1}{(2\pi)^N}} \exp\left[-\frac{|\boldsymbol{h} - Q_l\boldsymbol{a}|^2)}{2}\right], \tag{2}$$

where $\boldsymbol{a} = R\boldsymbol{b} = \{a_1, a_2, ... a_l\}^T$. Since we are not interested in a specific model-parametrisation, but only in the rank l of the design matrix, no generality is lost.

3. Bayesian model comparison

Let $\{Q_l\}$ be a complete set of mutually exclusive models – indexed by the regression order l – that compete to explain the dataset $\boldsymbol{h}$. Prior the measurement, the joint distribution of the data, model, and model parameters $\boldsymbol{a}$ can be written in terms of conditional distributions as

$$P(\boldsymbol{h}, \boldsymbol{a}, l) = L(\boldsymbol{h}|\boldsymbol{a}, l)\pi(\boldsymbol{a}|l)\Pi(l), \tag{3}$$

where the likelihood $L(\boldsymbol{h}|\boldsymbol{a}, l)$ is the sampling distribution of $\boldsymbol{h}$, $\pi(\boldsymbol{a}|l)$ is the pre-data distribution of the model parameters, and $\Pi(l)$ is the prior probability of Q_l being true. Accordingly, l – hence, the model – is sampled from $\Pi(l)$; then, the model parameters are sampled from $\pi(\boldsymbol{a}|l)$; eventually, the data are sampled from $L(\boldsymbol{h}|\boldsymbol{a}, l)$. The prior distribution $\Pi(l)$ must synthesize information available about the data model. In the absence of information, the probabilities that maximizes the $\Pi(l)$ entropy are equal.

By marginalisation of $P(\boldsymbol{h}, \boldsymbol{a}, l)$ over the model parameters and conditioning on the data, the posterior probability of the l-th model provided by the data is

$$P(l|\boldsymbol{h}) = \frac{Z(\boldsymbol{h}|l)\Pi(l)}{\sum_l Z(\boldsymbol{h}|l)\Pi(l)}. \tag{4}$$

Here, the marginal likelihood (or evidence),

$$Z(\boldsymbol{h}|l) = \int_A L(\boldsymbol{h}|\boldsymbol{a}, l)\pi(\boldsymbol{a}|l)\mathrm{d}\boldsymbol{a}, \tag{5}$$

where the integration is carried out on the parameter space, is the sampling distribution of the data given the model, independently of what the parameter values may be. It summarises the uncertainty about what is the model explaining the data.

Not unlike the case where the Bayes theorem is applied to determine the posterior probabilities of the model-parameter values, the distribution (4) embeds all the statements about the uncertainty after the data have been observed. In parameter estimation, uncertainty summaries are the posterior standard deviation and credible intervals. Also, the optimal estimate under quadratic loss of a parameter function (i.e., a measurand) is its expectation with respect the posterior distribution of the parameter values. The same is true for the posterior model-distribution. A simple way to select a model is to chose the mode. When $P(l|\boldsymbol{h})$ does not allow for a reliable comparison of posterior probabilities and no single model stands out, we can consider all the models and average over them. Whether and how to synthesize $P(l|\boldsymbol{h})$ – e.g., by selection or averaging – depends on decision theory considerations.

To investigate how much $P(l|\boldsymbol{h})$ depends on the assumptions made – that is, on the priors $\Pi(l)$ and $\pi(\boldsymbol{a}|l)$ and likelihood $L(\boldsymbol{h}|\boldsymbol{a}, l)$ – we can apply the Bayesian analysis to all plausible combinations of priors and likelihood. Then, the $P(l|\boldsymbol{h})$ variability with these changes delivers information about how much confidence we can place on the model W_l. This sensitivity analysis is out of the scope of this manuscript. Besides, the $\Pi(l)$ and $L(\boldsymbol{h}|\boldsymbol{a}, l)$ distributions have been uniquely set by the available information via entropy maximization. Furthermore, sections 4.1 and 4.2, will prove that also $\pi(\boldsymbol{a}|l)$ is uniquely imposed to result in equal posterior probabilities of models that differ only in parametrization.

The marginal likelihood and posterior model-probability are proportional to the ability of the model to explain the data – the higher the fitness, the higher $Z(\boldsymbol{h}|l)$ – but inversely proportional to the volume of the parameter space – the higher the model freedom, the lesser $Z(\boldsymbol{h}|l)$. Therefore, models with fewer parameters, which have a smaller parameter space, are preferred. This bias is known as the Ockham's razor and penalises the models having greater freedom in explaining the data.

4. Application to linear regression

4.1. Prior distribution of the model parameters

In order to find the model probabilities via (4) and (5), we need the pre-data distribution $\pi(\boldsymbol{a}|l)$. The probability calculus updates prior distributions into posterior ones, but it does not tell how to assign the prior probabilities.

In the absence of pre-data information, the posterior model-probabilities $P(l|\boldsymbol{h})$ and, consequently, the marginal likelihood $Z(\boldsymbol{h}|l)$ must be independent of the model parameters, that is, they must be invariant for model reparametrisations and transform as probability distributions. For instance, if the explaining model is a polynomial, (4)

and (5) must depend only on the polynomial degree l and not on what basis, orthogonal or not, is used.

This is ensured by the Jeffreys' prior $\pi_J(\boldsymbol{a}|l) \propto \sqrt{|F(\boldsymbol{a}|l)|}$, where $F(\boldsymbol{a}|l)$ is the Fisher information that the data carry about the parameters and the vertical bars denote the determinant. To calculate $\pi_J(\boldsymbol{a}|l)$, let us observe that the log-likelihood Jacobian of (2) is [14]

$$J = \partial_{\boldsymbol{a}}\big[\ln(L)\big] = -\partial_{\boldsymbol{a}}(\boldsymbol{x}^T\boldsymbol{x})/2 = -\boldsymbol{x}^T(\partial_{\boldsymbol{a}}\boldsymbol{x}) = \boldsymbol{x}^T Q_l \tag{6}$$

where $\boldsymbol{x} = \boldsymbol{h} - Q_l\boldsymbol{a}$. Hence,

$$F(\boldsymbol{a}|l) = -\langle\partial_{\boldsymbol{a}} J\rangle_{h|a,l} = Q_l^T Q_l = \mathbb{1} \tag{7}$$

where $\langle\rangle_{h|a,l}$ indicates the average with respect to the data conditioned to the model and model parameters. Since (7) does not depend on the model parameters, $\pi_J(\boldsymbol{a}|l) \propto$ const. is improper.

The use of an improper prior to calculate the posterior distribution of the model parameters,

$$P(\boldsymbol{a}|\boldsymbol{h}, l) = \frac{L(\boldsymbol{h}|\boldsymbol{a}, l)\pi_J(\boldsymbol{a}|l)}{Z(\boldsymbol{h}|l)}, \tag{8}$$

is justified by showing that $P(\boldsymbol{a}|\boldsymbol{h}, l)$ is the limit of the posteriors obtained from proper priors defined on increasingly large bounded-supports [15]. However, this argument encodes that $|\boldsymbol{a}|$ is greater than any positive number. This is irrelevant to the $\boldsymbol{a}$ posterior, but, it is not so for $Z(\boldsymbol{h}|l)$ and $P(l|\boldsymbol{h})$. In fact, since $Z(\boldsymbol{h}|l)$ and $P(l|\boldsymbol{h})$ are defined only up to model-dependent, but unknown, scale factors, an improper prior makes their calculation meaningless.

4.2. Hierarchical model

A way to escape from this difficulty is to complement the problem with measurable prior information, which information makes the $Z(\boldsymbol{h}|l)$ and $P(l|\boldsymbol{h})$ invariance irrelevant. Since we know in advance that both the expected value and variance of the data are bounded [4], we can encode in the parameter prior a prejudice towards repeated measurements of the same quantity, whose value, α, is unknown. This prejudice is modelled by setting $\boldsymbol{w}_1 = \{1, 1, \ldots 1\}^T$ in (1), where $\boldsymbol{w}_1$ is the first column of W_l, and $\langle\boldsymbol{y}\rangle_{b,u} = \alpha\boldsymbol{w}_1$. The subscript indicates that we take the mean over the joint (prior) distribution of $\boldsymbol{b}$ and $\boldsymbol{u}$.

After switching to the normalized data and orthonormal basis, the relationships $|\boldsymbol{q}_1| = 1$ and $Q_l Q_l^T \boldsymbol{q}_1 = \boldsymbol{q}_1$ hold. It is convenient to ensure that the estimate $\hat{a}_1 = \boldsymbol{q}_1^T\boldsymbol{h}$ is proportional to the weighted mean of the data,

$$y_{LS} = (\boldsymbol{w}_1^T\Sigma_u^{-1}\boldsymbol{w}_1)^{-1}\boldsymbol{w}_1^T\Sigma_u^{-1}\boldsymbol{y}. \tag{9a}$$

This is achieved by starting the orthonormalisation from $\boldsymbol{v}_1 = U^{-1}\boldsymbol{w}_1$; hence, $\boldsymbol{q}_1 = \boldsymbol{v}_1/|\boldsymbol{v}_1|$. In fact,

$$y_{LS} = (\boldsymbol{v}_1^T\boldsymbol{v}_1)^{-1}\boldsymbol{v}_1^T\boldsymbol{h} = \hat{a}_1/|\boldsymbol{v}_1|, \tag{9b}$$

where we used $\Sigma_u^{-1} = U^{-T}U^{-1}$, $U^{-1}\boldsymbol{y} = \boldsymbol{h}$, and $\boldsymbol{q}_1^T\boldsymbol{q} = 1$.

That done, the repeated measurements of a constant prejudice is still encoded by $\langle\boldsymbol{h}\rangle_{a,\epsilon} = \alpha\boldsymbol{q}_1$, where we redefined the hyperparameter α. Hence, $Q\langle\boldsymbol{a}\rangle_a = \langle\boldsymbol{h}\rangle_{a,\epsilon} = \alpha\boldsymbol{q}_1$ and

$$\langle\boldsymbol{a}\rangle_a = \alpha Q_l^T\boldsymbol{q}_1 = \alpha\boldsymbol{p}_1, \tag{10}$$

where $\boldsymbol{p}_1 = Q_l^T \boldsymbol{q}_1 = \{1, 0, ...0\}^T$ and we used $Q_l^T Q_l = \mathbb{1}$. Next, we encoded the bounded data-variance prejudice by

$$\Sigma_a = \langle \boldsymbol{a}\boldsymbol{a}^T \rangle_a = (\beta^2 - 1)\mathbb{1}, \tag{11}$$

where we used $\langle \boldsymbol{h}\boldsymbol{h}^T \rangle_{a,\epsilon} = \beta^2 \mathbb{1} = Q_l \Sigma_a Q_l^T + \mathbb{1}$ and $\beta > 1$.

The prior distribution of the model parameters constrained by (10) and (11) and maximizing the relative entropy with respect to the uniform distribution is

$$\pi(\boldsymbol{a}|\alpha, \beta, l) = \sqrt{\frac{1}{(2\pi)^l (\beta^2 - 1)^l}} \exp\left[-\frac{|\boldsymbol{a} - \alpha \boldsymbol{p}_1|^2}{2(\beta^2 - 1)} \right], \tag{12}$$

where the hyper-parameters α and β are unknown. It is worth noting that, when $\beta \to \infty$, $\pi(\boldsymbol{a}|\alpha, \beta, l)$ reproduces the uniform prior.

4.3. Sampling distribution of the data

The distribution of the normalized data, given the hyper-parameters α and β and the orthonormal model Q_l,

$$\begin{aligned} \mathcal{L}(\boldsymbol{h}|\alpha, \beta, l) &= \int_{\mathbf{R}^l} L(\boldsymbol{h}|\boldsymbol{a}, l) \pi(\boldsymbol{a}|\alpha, \beta, l)\, \mathrm{d}\boldsymbol{a} \\ &= \sqrt{\frac{1}{(2\pi)^n \beta^{2l}}} \exp\left(-\frac{(\boldsymbol{h} - \alpha \boldsymbol{q}_1)^T \left[\mathbb{1} + (\beta^2 - 1) Q_l Q_l^T \right]^{-1} (\boldsymbol{h} - \alpha \boldsymbol{q}_1)}{2} \right), \end{aligned} \tag{13}$$

is obtained by marginalization over the model parameters. The relevant integration is carried out in the Appendix A.

4.4. Prior distribution of the hyper-parameters

To ensure that $P(l|\boldsymbol{h})$ is invariant for model reparametrisations, we must compute the Jeffreys' prior $\pi_J(\alpha, \beta|l)$. As it is shown in the Appendix B, the Fisher information that $\boldsymbol{h}$ carries about α and β is

$$F(\alpha, \beta) = \frac{1}{\beta^2} \begin{pmatrix} 1 & 0 \\ 0 & 2l \end{pmatrix}. \tag{14}$$

Therefore, the Jeffrey's distribution of α and β given the Q_l model is

$$\pi_J(\alpha, \beta|l) \propto \sqrt{|F(\alpha, \beta|l)|} \propto 1/\beta^2, \tag{15}$$

whose support is $\beta > 1$, $-\infty < \alpha < +\infty$. Although it is improper, since now the same normalizing-factor is included in all the marginal likelihoods, (15) does not impede the Q_l comparison. In fact, after the normalization, the model probabilities (4) do not depend on the $\pi_J(\alpha, \beta|l)$ support in the α space.

4.5. Marginal likelihood

The computation of the marginal likelihood (the data sampling-distribution given W_l) is given in Appendix C. Although it was derived using uncorrelated data having the same unit variance and an orthonormal basis in the model subspace, the result,

$$Z(\boldsymbol{y}|l) \propto \frac{\mathrm{e}^{-|\hat{\boldsymbol{u}}|^2/2}\, \gamma(l/2, |\hat{\boldsymbol{y}}'|^2/2)}{\left(|\hat{\boldsymbol{y}}'|/\sqrt{2}\right)^l}, \tag{16}$$

is independent of the transformations made and it is written terms of the weighted least-squares estimate of the original data

$$\hat{\boldsymbol{y}}' = W_l(W_l^T\Sigma_u^{-1}W_l)^{-1}W_l^T\Sigma_u^{-1}\boldsymbol{y}', \tag{17}$$

where $\boldsymbol{y}' = \boldsymbol{y} - \boldsymbol{w}_1 y_{LS}$ and y_{LS} is the weighted mean ($9a$), and the residuals $\hat{\boldsymbol{u}}$. Unessential factors independent of the regression order l and data have been left out and $\gamma(s, z_2)$ is the lower incomplete gamma function [16]. The weighted norms $|\hat{\boldsymbol{u}}|^2$ and $|\hat{\boldsymbol{y}}'|^2$ are calculated according to the Σ_u^{-1} metric, that is, $|\hat{\boldsymbol{u}}|^2 = \hat{\boldsymbol{u}}^T\Sigma_u^{-1}\hat{\boldsymbol{u}}$ and $|\hat{\boldsymbol{y}}'|^2 = \hat{\boldsymbol{y}}'^T\Sigma_u^{-1}\hat{\boldsymbol{y}}'$.

Eventually, with the assumed uniform $\Pi(l)$ prior distribution, the posterior probability of each W_l model is obtained by the normalization to $\sum_l Z(\boldsymbol{y}|l) = 1$.

When we consider independent and identically distributed data $\Sigma_u = \sigma_0^2\mathbb{1}$ and

$$Z(\boldsymbol{y}|l) \propto \frac{\mathrm{e}^{-|\hat{\boldsymbol{u}}|^2/(2\sigma_0^2)}\gamma\big[l/2, |\hat{\boldsymbol{y}}'|^2/(2\sigma_0^2)\big]}{\big[|\hat{\boldsymbol{y}}'|/(\sqrt{2}\sigma_0)\big]^l}, \tag{18}$$

where, now, y_{LS} is the arithmetic mean of the data, $\hat{\boldsymbol{y}}' = W_l(W_l^TW_l)^{-1}W_l^T\boldsymbol{y}'$ is the ordinary least-squares estimate, and $|\hat{\boldsymbol{u}}|^2 = \hat{\boldsymbol{u}}^T\hat{\boldsymbol{u}}$ and $|\hat{\boldsymbol{y}}'|^2 = \hat{\boldsymbol{y}}'^T\hat{\boldsymbol{y}}'$ are the norms of the residuals and data-estimates calculated according the Euclidean metric. We observe that (18) is the same as the equation (B.4) in [6]. In fact, since $\hat{\boldsymbol{y}}^T(\boldsymbol{y} - \hat{\boldsymbol{y}}) = 0$ (because $\hat{\boldsymbol{y}}$ is the $\boldsymbol{y}$ projection in the $W_l\boldsymbol{b}$ sub-space), $\boldsymbol{y}^T\hat{\boldsymbol{u}} = (\boldsymbol{y} - \hat{\boldsymbol{y}})^T(\boldsymbol{y} - \hat{\boldsymbol{y}}) = |\hat{\boldsymbol{u}}|^2$.

Extending the comparison to non-linear models would require determining the Jeffreys' prior of the model parameters, which, in general, is a difficult problem to solve and devoid of a general solution. However, since it does not depend on the W_l representation, but only on the regression order, residuals, and central moments of the data, (16) can be heuristically applied also to non-linear models, provided that their curvature in the parameter subspace of interest can be neglected – which means that a linear approximation holds.

4.6. Asymptotic behaviour

For a large data sample, provided $|\hat{\boldsymbol{y}}'|^2 \gg l$,

$$\ln\big[Z(\boldsymbol{y}|l)\big] \propto -|\hat{\boldsymbol{u}}|^2/2 - l\ln(|\hat{\boldsymbol{y}}'|), \tag{19a}$$

where we omitted the terms independent of l and used $\gamma(l/2, |\hat{\boldsymbol{y}}'|^2/2) \approx \Gamma(l/2)$ and $\Gamma(l/2) \ll |\hat{\boldsymbol{y}}'|^l$ [17]. Furthermore, for large $l \ll |\hat{\boldsymbol{y}}'|^2$, $|\hat{\boldsymbol{y}}'|^2$ is independent of the regression order and proportional to the sample size N. Hence,

$$\ln\big[Z(\boldsymbol{y}|l)\big] \propto -|\hat{\boldsymbol{u}}|^2/2 - l\ln(N)/2, \tag{19b}$$

always omitting the terms independent of l.

The model having the maximum posterior-probability maximises the log-likelihood. The asymptotic behaviour ($19a$) shows that the optimal order minimises the weighted residuals by keeping the number of free parameters as small as possible, thus handicapping the overfitting. Among the regressions having the same order, the most supported is that whose associated sum of the (weighted) squared residuals is minimum.

To compare (16) with the classical indicators, we compare ($19a$-b) with the marginal likelihoods of the asymptotic information criteria due to Akaike [18, 19],

$$\ln\big[Z_{AIC}(\boldsymbol{y}|l)\big] \propto -|\hat{\boldsymbol{u}}|^2/2 - l, \tag{20a}$$

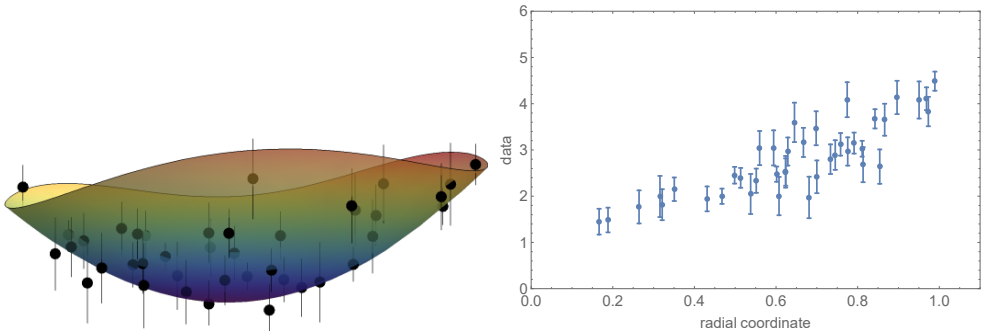

Figure 1. Left: three dimensional scatter plot of the data sampled from (21). The bars indicate 95% confidence intervals. The best-fit regression to the data is also shown. Right: radial plot of the same data.

and Schwarz [20, 21, 22],

$$\ln\left[Z_{BIC}(\boldsymbol{y}|l)\right] \propto -|\hat{\boldsymbol{u}}|^2/2 - l\ln(N)/2, \tag{20b}$$

where we omitted the terms that are independent of the regression order (which does not matter, because all the models are Gaussian). The identity of (19b) and (20b) shows that the model probabilities derived from (16) shed light and encompass these (approximate) criteria.

5. Numerical example

Figure 1 shows a set of 40 data randomly sampled in the unit disk from the aberrated wavefront

$$y = b_1 Z_0^0(x_1,x_2) + b_2 Z_1^1(x_1,x_2) + b_3 Z_2^0(x_1,x_2) + b_4 Z_3^{-1}(x_1,x_2), \tag{21}$$

where $Z_n^m(x_1,x_2)$ are Zernike polynomials, $\{Z_0^0, Z_1^1, Z_2^0, Z_3^{-3}\}$ is the basis set, and the parameter values are $\boldsymbol{b} = \{3, 0.5, 1.5, 0.5\}$. Zero mean uncorrelated Gaussian errors – having random standard-deviations $\sigma_0(i)$ ranging from 5% to 23% – were added to the data.

To explain the data, we carried out regressions (having order ranging from $l = 1$ to $l = 11$) using the 1024 basis sets corresponding to subsets – containing Z_0^0 – of

$$\{Z_0^0, Z_1^{-1}, Z_1^1, Z_2^0, Z_3^{-3}, Z_3^3, Z_3^{-1}, Z_3^1, Z_4^0\}. \tag{22}$$

We calculated the marginal likelihood and posterior probability of each basis by application of (16) and (4), where the bases were assumed to have the same prior probability.

To assess the efficiency of (16) in detecting the right model, we repeated the regressions 100 times – by using 100 independent datasets – and averaged the posterior model-probabilities. After sorting the results into decreasing probabilities, Fig. 2 (left) shows the top eight posterior-probabilities. The most sustained basis is that used to generate the data.

In a second test, we used exponentially correlated errors. Therefore, the variance-covariance matrix of the data is

$$\Sigma_u(i,j) = \sigma_0(i)\sigma_0(j)\mathrm{e}^{-5r_{ij}}, \tag{23}$$

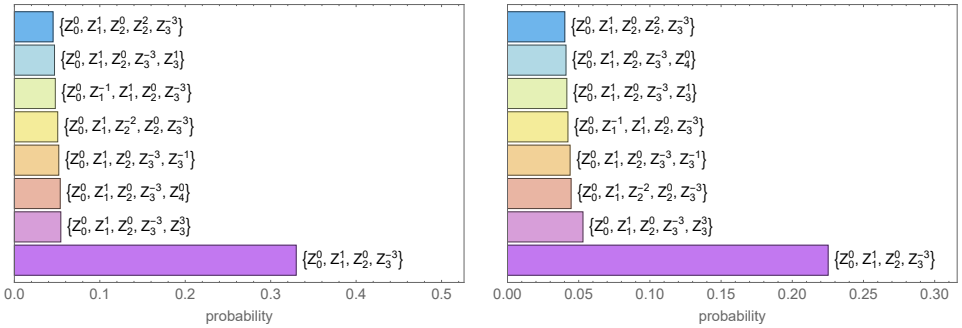

Figure 2. Top eight probabilities of the subsets of (22) averaged over 100 Monte Carlo simulations. Left: uncorrelated data. Right: exponentially correlated data.

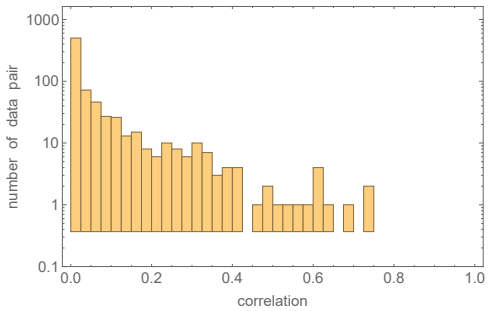

Figure 3. Histogram of the correlations between the data sampled from (21).

where r_{ij} is the distance of the datum i from the j one. Figure 3 shows an example of the data correlations. Also in this case, we repeated the regressions 100 times – by using 100 independent datasets – and averaged the posterior model-probabilities. Figure 2 (right) shows the result. Though the mean probability of the $\{Z_0^0, Z_1^1, Z_2^0, Z_3^{-3}\}$ basis decreases from 33% to 23%, it is still the most fostered by the data.

6. Application example

The marginal likelihood given in (16) was used to investigate doubtful measurements carried out to determine the central second-moment of the angular power-spectrum of a laser beam, which is a pivotal quantity when correcting for the diffraction effects the length measurements carried out by optical interferometry. We do not want here describe these measurements nor draw conclusions about them, but only show how the results obtained can be used in practice.

The second moment is measured by imaging the power spectrum in the focal plane of a telescope and by numerically integrating the normalised image by a sum over the camera pixels [23]. We repeated the measurement by placing the telescope at 137 cm, 394 cm, and 1047 cm distance from the beam source and grabbed defocused images by putting the camera in the focal plane and at $(-15, -10, -5, +5, +10, +15)$ mm from it. The results are shown in Fig. 4. The paraxial propagation of a scalar beam in free space predicts that the focal-plane images do not depend on the telescope

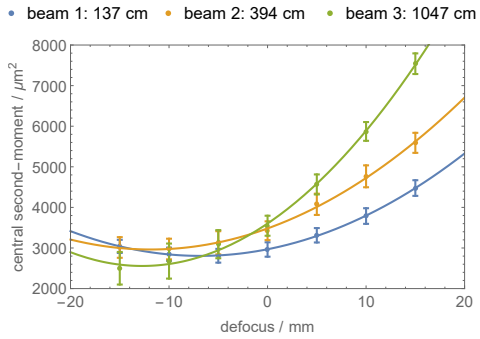

Figure 4. Central second-moment of the defocused images. Each set of seven data corresponds to a different telescope distance. The solid lines are the best parabolas fitting the data.

distance and that the seven central-moments depend quadratically on the defocus.

Figure 4 shows that the best-fit parabolas to the data do not intersect where the defocus is null, as we would have expected. To quantify how much the data support this observation and the quadratic dependence on the defocus, we compared three hypotheses: the data belong to i) non-intersecting parabolas, that is,

$$\boldsymbol{y} = W_9 \boldsymbol{b} + \boldsymbol{u} \tag{24a}$$

or

$$y_{ij} = b_{i0} + b_{i1} z_j + b_{i2} z_j^2 + u_{ij}, \tag{24b}$$

where the data and parameters are assembled in $\boldsymbol{y}$ and $\boldsymbol{b}$, z is the defocus, $i = 1, 2, 3$ labels the data series, and $j = 1, 2, \ldots 7$ labels the defocus; ii) non-intersecting quartics functions, that is,

$$\boldsymbol{y} = W_{12} \boldsymbol{b} + \boldsymbol{u} \tag{25a}$$

or

$$y_{ij} = b_{i0} + b_{i1} z_j + b_{i2} z_j^2 + b_{i4} z_j^4 + u_{ij}; \tag{25b}$$

and iii) parabolas intersecting at null defocus, that is

$$\boldsymbol{y} = W_7 \boldsymbol{b} + \boldsymbol{u} \tag{26a}$$

or

$$y_{ij} = b_0 + b_{i1} z_j + b_{i2} z_j^2 + u_{ij}. \tag{26b}$$

The design matrices are given in the Appendix D.

With assumed equal prior probabilities of the models – after removing the weighted mean from the data and calculating their weighted least-square estimates and residuals – the marginal likelihood (16) and its normalization to $\sum_l Z(\boldsymbol{y}|l) = 1$ delivers the posterior probabilities shown in Fig. 5 (left). It confirms that the second moment depends quadratically on the defocus and the non-intersecting parabolas hypothesis overcomes by a 99.99 % probability the other two. We don't yet have an explanation for this unexpected behaviour.

For a Gaussian beam, the second moment and its curvature at the vertex are constrained by the unity value of the quality factor. Since a lens does not change the unit value of the quality factor, the beams downstream of the telescope lens must be

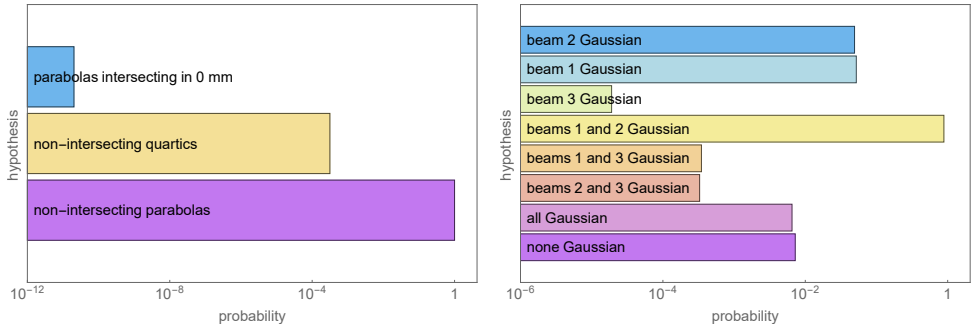

Figure 5. Posterior probabilities of the listed explanations of the data shown in Fig. 4

all Gaussian or all paraxial. Therefore, as an additional application of (16) (this time heuristically comparing non-linear models), we calculated the probabilities that none, one, two, or all the downstream beams are Gaussian.

A unit quality factor implies that the second moment propagates as $y^2 = a+c(z-z_0)^2$, where $k\sqrt{ac} = 1$ constrains the a and c parameters, k is the wavenumber, and z_0 the vertex position. Hence, the data models are

$$y_{ij} = a_i + \frac{(z_j - z_{i0})^2}{a_i k^2} + u_{ij}, \tag{27}$$

if the i-th beam is Gaussian, and

$$y_{ij} = a_i + c_i(z_j - z_{i0})^2 + u_{ij}, \tag{28}$$

if the i-th beam is paraxial. Eventually, eight (non-linear) data models were built by arranging in a eight systems three equations (corresponding to $i = 1, 2, 3$) drawn from (27) and/or (28).

Figure 5 (right) shows the posterior probabilities, where all the explanations were assumed to have the same prior probability. Also in this case (16) identifies a single explanation, assigns it an 88.4 % probability, and points out a severe inconsistency. This result was confirmed by calculating the M^2 value of the three beams.

7. Conclusions

This paper showed how to select the linear regression most supported by the data and extended previous works that did not account for weights and correlations [4, 6]. The selection requires that the marginal likelihood and posterior regression probabilities be independent of the parametrisation used, which previous investigations did not consider [3, 5]. To comply with this requirement, we derived Jeffreys priors from the volume element of each model-manifold equipped with the information metric. Also, to avoid inconsistencies due to improper priors, we used hierarchical modelling.

Although the computation of the marginal likelihood required complex integrations, the final formula (16), which can also be heuristically used for non-linear models, involves only standard least-squares estimates and simple algebra. Our result makes Bayesian model selection easily accessible and extends this toolbox to the analysis of arbitrarily correlated and weighted data.

Appendix A. Data likelihood

To calculate the sampling-distribution $\mathcal{L}(\boldsymbol{h}|\alpha,\beta,l)$ we observe that

$$|\boldsymbol{h}-Q_l\boldsymbol{a}|^2=|\boldsymbol{a}-\hat{\boldsymbol{a}}|^2+|\boldsymbol{h}|^2-|\hat{\boldsymbol{a}}|^2, \tag{A.1}$$

where $\hat{\boldsymbol{a}}=Q_l^T\boldsymbol{h}=\{\hat{a}_1,\hat{a}_2,...,\hat{a}_l\}^T$ are the least squares estimates of the model parameters. Furthermore,

$$|\boldsymbol{a}-\alpha\boldsymbol{p}_1|^2=|\boldsymbol{a}'|^2+(a_1-\alpha)^2. \tag{A.2}$$

and

$$|\boldsymbol{a}-\hat{\boldsymbol{a}}|^2=|\boldsymbol{a}'-\hat{\boldsymbol{a}}'|^2+(a_1-\hat{a}_1)^2, \tag{A.3}$$

where $\hat{a}_1=\boldsymbol{q}_1^T\boldsymbol{h}$, $\boldsymbol{a}'=\{a_2,a_3,...a_l\}^T$ and $\hat{\boldsymbol{a}}'=\{\hat{a}_2,\hat{a}_3,...\hat{a}_l\}^T$. Hence, by carrying out the needed integrations with the aid of Mathematica® [17],

$$\begin{aligned}
\mathcal{L}(\boldsymbol{h}|\alpha,\beta,l)&=\sqrt{\frac{1}{(2\pi)^{n+l}(\beta^2-1)^l}}\int_{-\infty}^{+\infty}\mathrm{d}a_1\int_{\mathbf{R}^{l-1}}\exp\left(-\frac{|\boldsymbol{h}-Q_l\boldsymbol{a}|^2}{2}\right)\exp\left[-\frac{|\boldsymbol{a}-\alpha\boldsymbol{p}_1|^2}{2(\beta^2-1)}\right]\mathrm{d}\boldsymbol{a}'\\
&=\sqrt{\frac{1}{(2\pi)^{n+l}(\beta^2-1)^l}}\exp\left(-\frac{|\boldsymbol{h}|^2-|\hat{\boldsymbol{a}}|^2}{2}\right)\int_{-\infty}^{+\infty}\exp\left(-\frac{(a_1-\hat{a}_1)^2}{2}\right)\exp\left(-\frac{(a_1-\alpha)^2}{2(\beta^2-1)}\right)\mathrm{d}a_1\\
&\times\int_{\mathbf{R}^{l-1}}\exp\left(-\frac{|\boldsymbol{a}'-\hat{\boldsymbol{a}}'|^2}{2}\right)\exp\left(-\frac{|\boldsymbol{a}'|^2}{2(\beta^2-1)}\right)\mathrm{d}\boldsymbol{a}'\\
&=\sqrt{\frac{1}{(2\pi)^{(n+1)}\beta^{2(l-1)}(\beta^2-1)}}\exp\left(-\frac{|\boldsymbol{h}|^2-|\hat{\boldsymbol{a}}|^2}{2}-\frac{|\hat{\boldsymbol{a}}'|^2}{2\beta^2}\right)\\
&\times\int_{-\infty}^{+\infty}\exp\left(-\frac{(a_1-\hat{a}_1)^2}{2}\right)\exp\left(-\frac{(a_1-\alpha)^2}{2(\beta^2-1)}\right)\mathrm{d}a_1\\
&=\sqrt{\frac{1}{(2\pi)^n\beta^{2l}}}\exp\left(-\frac{|\boldsymbol{h}|^2-|\hat{\boldsymbol{a}}|^2}{2}-\frac{|\hat{\boldsymbol{a}}'|^2}{2\beta^2}+\frac{(\hat{a}_1-\alpha)^2}{2\beta^2}\right)\\
&=\sqrt{\frac{1}{(2\pi)^n\beta^{2l}}}\exp\left(-\frac{|\boldsymbol{h}|^2-|\hat{\boldsymbol{a}}|^2}{2}-\frac{|\hat{\boldsymbol{a}}-\alpha\boldsymbol{p}_1|^2}{2\beta^2}\right)\\
&=\sqrt{\frac{1}{(2\pi)^n\beta^{2l}}}\exp\left(-\frac{\boldsymbol{h}^T(\mathbb{1}-Q_lQ_l^T)\boldsymbol{h}}{2}-\frac{(\boldsymbol{h}-\alpha\boldsymbol{q}_1)^TQ_lQ_l^T(\boldsymbol{h}-\alpha\boldsymbol{q}_1)}{2\beta^2}\right).
\end{aligned} \tag{A.4}$$

It is worth noting that $\mathcal{L}(\boldsymbol{h}|\alpha,\beta,l)$ is a multivariate normal distribution having

$$\langle\boldsymbol{h}\rangle_{h|\alpha,\beta,l}=\alpha\boldsymbol{q}_1 \tag{A.5}$$

mean and

$$\Sigma_h=\mathbb{1}+(\beta^2-1)Q_lQ_l^T \tag{A.6}$$

variance-covariance matrix. To prove (A.5), it is enough to find the $\mathcal{L}(\boldsymbol{h}|\alpha,\beta,l)$ maximum by observing that the gradient of the quadratic form $\boldsymbol{h}^T\Sigma\boldsymbol{h}$, where Σ is symmetric, is $\partial_{\boldsymbol{h}}(\boldsymbol{h}^T\Sigma\boldsymbol{h})=2\boldsymbol{h}^T\Sigma$ and that $\boldsymbol{q}_1^TQ_lQ_l^T=\boldsymbol{q}_1^T$ [14].

Next, we observe that $\partial^2_{\boldsymbol{h}}(\boldsymbol{h}^T\Sigma\boldsymbol{h})=2\Sigma$. Hence, the opposite of the Hessian of the $\mathcal{L}(\boldsymbol{h}|\alpha,\beta,l)$ exponent,

$$\Sigma_h^{-1}=\mathbb{1}-\frac{\beta^2-1}{\beta^2}Q_lQ_l^T, \tag{A.7}$$

is the inverse of (A.6) [24, 25].

Appendix B. Fisher information matrix

The entries of the Fisher information matrix (14) are

$$F_{\alpha\alpha} = -\langle\partial_{\alpha\alpha}\big[\ln(\mathcal{L})\big]\rangle_{h|\alpha,\beta,l} = |\boldsymbol{p}_1|^2/\beta^2 = 1/\beta^2, \tag{B.1}$$

$$F_{\alpha\beta} = -\langle\partial_{\alpha\beta}\big[\ln(\mathcal{L})\big]\rangle_{h|\alpha,\beta,l} = \langle\hat{\boldsymbol{a}} - \alpha\boldsymbol{p}_1\rangle_{h|\alpha,\beta,l}\boldsymbol{p}_1/\beta^3 = 0, \tag{B.2}$$

$$F_{\beta\beta} = -\langle\partial_{\beta\beta}\big[\ln(\mathcal{L})\big]\rangle_{h|\alpha,\beta,l} = \frac{1}{\beta^2}\left(\frac{3}{\beta^2}\langle|\hat{\boldsymbol{a}} - \alpha\boldsymbol{p}_1|^2\rangle_{h|\alpha,\beta,l} - l\right) = \frac{2l}{\beta^2}, \tag{B.3}$$

where, apart from unessential terms independent of α and β,

$$\ln(\mathcal{L}) = -|\hat{\boldsymbol{a}} - \alpha\boldsymbol{p}_1|^2/(2\beta^2) - l\ln(\beta). \tag{B.4}$$

To evaluate $\langle\hat{\boldsymbol{a}} - \alpha\boldsymbol{p}_1\rangle_{h|\alpha,\beta,l}$, we observe that $\hat{\boldsymbol{a}} = Q_l^T\boldsymbol{h}$ and use (A.5) and $Q_l^T\boldsymbol{q}_1 = \boldsymbol{p}_1$. The evaluation of $\langle|\hat{\boldsymbol{a}} - \alpha\boldsymbol{p}_1|^2\rangle_{h|\alpha,\beta,l}$ is a bit more tricky. First, we note that $\hat{\boldsymbol{a}} = Q_l^T\boldsymbol{h}$ and $\langle\hat{\boldsymbol{a}}\rangle_{h|\alpha,\beta,l} = Q_l^T\langle\boldsymbol{h}\rangle_{h|\alpha,\beta,l} = \alpha Q_l^T\boldsymbol{q}_1 = \alpha\boldsymbol{p}_1$. Hence,

$$\langle|\hat{\boldsymbol{a}} - \alpha\boldsymbol{p}_1|^2\rangle_{h|\alpha,\beta,l} = \mathrm{Tr}(\Sigma_{\hat{\boldsymbol{a}}}), \tag{B.5}$$

where $\Sigma_{\hat{\boldsymbol{a}}}$ is the variance-covariance matrix of $\hat{\boldsymbol{a}}$. Next, since $\hat{\boldsymbol{a}} = Q_l^T\boldsymbol{h}$, by using $Q_l^TQ_l = \mathbb{1}$ and (A.6),

$$\Sigma_{\hat{\boldsymbol{a}}} = Q_l^T\Sigma_h Q_l = \beta^2\mathbb{1}. \tag{B.6}$$

Eventually, $\mathrm{Tr}(\Sigma_{\hat{\boldsymbol{a}}}) = l\beta^2$.

Appendix C. Marginal likelihhod

By carrying out the needed integrations with the aid of Mathematica® [17], The marginal likelihood is

$$\begin{aligned}
Z(\boldsymbol{h}|l) &= \int_1^{+\infty}\int_{-\infty}^{+\infty}\mathcal{L}(\boldsymbol{h}|\alpha,\beta,l)\pi_J(\alpha,\beta|)\,\mathrm{d}\alpha\mathrm{d}\beta \\
&\propto \exp\left(-\frac{|\boldsymbol{h}|^2 - |\hat{\boldsymbol{a}}|^2}{2}\right)\int_1^{+\infty}\left[\frac{1}{\beta^{l+2}}\exp\left(\frac{|\hat{\boldsymbol{a}}'|^2}{2\beta^2}\right)\int_{-\infty}^{+\infty}\exp\left(\frac{(\hat{a}_1-\alpha)^2}{2\beta^2}\right)\mathrm{d}\alpha\right]\mathrm{d}\beta \\
&\propto \exp\left(-\frac{|\hat{\boldsymbol{\epsilon}}|^2}{2}\right)\int_1^{+\infty}\frac{1}{\beta^{l+1}}\exp\left(\frac{|\hat{\boldsymbol{a}}'|^2}{2\beta^2}\right)\mathrm{d}\beta \\
&= \frac{\sqrt{2^{l-1}}\exp\big(-|\hat{\boldsymbol{\epsilon}}|^2/2\big)\gamma(l/2,|\hat{\boldsymbol{a}}'|^2/2)}{|\hat{\boldsymbol{a}}'|^l},
\end{aligned} \tag{C.1}$$

where $\gamma(s,z_2)$ is the lower incomplete gamma function [16], $|\hat{\boldsymbol{\epsilon}}|^2 = \hat{\boldsymbol{\epsilon}}^T\hat{\boldsymbol{\epsilon}}$ is the sum of the squared residuals, $\hat{\boldsymbol{\epsilon}} = \boldsymbol{h} - \hat{\boldsymbol{h}}$ are the residuals, and $\hat{\boldsymbol{h}} = Q_l\hat{\boldsymbol{a}}$ is the least-squares estimate of $\boldsymbol{h}$. To prove that $|\boldsymbol{h}|^2 - |\hat{\boldsymbol{a}}|^2 = |\hat{\boldsymbol{\epsilon}}|^2$, which has been used to obtain (C.1), we observe that $\hat{\boldsymbol{h}}^T(\boldsymbol{h} - \hat{\boldsymbol{h}}) = 0$, because $\hat{\boldsymbol{h}}$ is the projection of $\boldsymbol{h}$ in the $Q_l\boldsymbol{a}$ sub-space. Furthermore, $|\hat{\boldsymbol{a}}|^2 = \boldsymbol{h}^TQ_lQ_l^T\boldsymbol{h} = \boldsymbol{h}^T\hat{\boldsymbol{h}}$. Hence,

$$|\boldsymbol{h}|^2 - |\hat{\boldsymbol{a}}|^2 = \boldsymbol{h}^T(\boldsymbol{h} - \hat{\boldsymbol{h}}) = (\boldsymbol{h} - \hat{\boldsymbol{h}})^T(\boldsymbol{h} - \hat{\boldsymbol{h}}) = \hat{\boldsymbol{\epsilon}}^T\hat{\boldsymbol{\epsilon}}. \tag{C.2}$$

Since $Q^T\boldsymbol{h}' = Q^T\boldsymbol{h}$ and $QQ^T\boldsymbol{q}_1 = \boldsymbol{q}_1$, the marginal likelihood (C.1) is invariant against the $\boldsymbol{h}' = \boldsymbol{h} + g\boldsymbol{q}_1$ transformation, where g is any numerical constant. This can be used to rewrite it in a form that is independent of the data normalization and the orthonormalization of the design matrix.

To this end, firstly, we shift the $\boldsymbol{y}$ data in such a way that their generalised mean (9*a*-*b*) is null. In such a way $\hat{a}_1 = \boldsymbol{q}_1^T\boldsymbol{h}$ shifts to zero. Since $|\boldsymbol{q}_1| = \boldsymbol{q}_1^T\boldsymbol{q}_1 = 1$,

this corresponds to shift $\boldsymbol{h}$ to $\boldsymbol{h} - (\boldsymbol{q}_1^T\boldsymbol{h})\boldsymbol{q}_1$ and keeps the $Z(\boldsymbol{h}|l)$ value unchanged. Now, when $\hat{a}_1 = 0$, we have $|\hat{\boldsymbol{a}}'|^2 = |\hat{\boldsymbol{a}}|^2$. In addition, by using $Q_l^T Q_l = \mathbb{1}$, we get $|\hat{\boldsymbol{a}}|^2 = \hat{\boldsymbol{a}}^T(Q_l^T Q_l)\hat{\boldsymbol{a}} = |\hat{\boldsymbol{h}}|^2$. Eventually, in (C.1), we have $|\hat{\boldsymbol{a}}'|^2 = |\hat{\boldsymbol{h}}|^2$.

Secondly, we observe that $|\hat{\boldsymbol{h}}|^2 = \hat{\boldsymbol{y}}^T\Sigma_u^{-1}\hat{\boldsymbol{y}} = |\hat{\boldsymbol{y}}|^2$ and $|\hat{\boldsymbol{\epsilon}}|^2 = \hat{\boldsymbol{u}}^T\Sigma_u^{-1}\hat{\boldsymbol{u}} = |\hat{\boldsymbol{u}}|^2$, where $\hat{\boldsymbol{y}}$ are the generalised least-squares estimate of the data (17) after subtracting their generalised mean and $\hat{\boldsymbol{u}} = \boldsymbol{y} - \hat{\boldsymbol{y}}$.

To conclude, we can rewrite (C.1) by substituting $|\boldsymbol{u}|$ and $|\hat{\boldsymbol{y}}|$ for $|\hat{\boldsymbol{\epsilon}}|$ and $|\hat{\boldsymbol{a}}'|$, so that (16) gives the marginal likelihood in terms of the original data and variance-covariance matrix.

Appendix D. Design matrices

The central second moment of a paraxial beam depends quadratically on the propagation distance, z. Therefore, after assembling the data – three series of seven measured values – in the vector $\boldsymbol{y} = (y_{11}, \ldots y_{17}, y_{21}, \ldots y_{27}, y_{31}, \ldots y_{37})^T$ and the nine parameters of three non-intersecting parabolas in $\boldsymbol{b} = (a_1, b_1, c_1, a_2, b_2, c_2, a_3, b_3, c_3)^T$, the design matrix is

$$W_9 = \begin{pmatrix} 1 & z_1 & z_1^2 & 0 & 0 & 0 & 0 & 0 & 0 \\ \cdots & & & & & & & & \\ 1 & z_7 & z_7^2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & z_1 & z_1^2 & 0 & 0 & 0 \\ \cdots & & & & & & & & \\ 0 & 0 & 0 & 1 & z_7 & z_7^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & z_1 & z_1^2 \\ \cdots & & & & & & & & \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & z_7 & z_7^2 \end{pmatrix}. \tag{D.1}$$

A quartic polynomial was used as an alternative explanation to test the measurement quality. The design matrix,

$$W_{12} = \begin{pmatrix} 1 & z_1 & z_1^2 & z_1^4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \cdots & & & & & & & & & & & \\ 1 & z_7 & z_7^2 & z_7^4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & z_1 & z_1^2 & z_1^4 & 0 & 0 & 0 & 0 \\ \cdots & & & & & & & & & & & \\ 0 & 0 & 0 & 0 & 1 & z_7 & z_7^2 & z_7^4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & z_1 & z_1^2 & z_1^4 \\ \cdots & & & & & & & & & & & \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & z_7 & z_7^2 & z_7^4 \end{pmatrix}, \tag{D.2}$$

is obtained from (D.1) by adding three columns related to the z^4 basis and by supplementing the parameter vector, $\boldsymbol{b} = (a_1, b_1, c_1, d_1, a_2, b_2, c_2, d_2, a_3, b_3, c_3, d_3)^T$, accordingly.

Eventually, the expectation that the three parabolas intersect in the null-defocus configuration was modelled by imposing the constraint $a_1 = a_2 = a_3 = a_0$. This was

done by using the design matrix

$$W_7 = \begin{pmatrix} 1 & z_1 & z_1^2 & 0 & 0 & 0 & 0 \\ \cdots & & & & & & \\ 1 & z_7 & z_7^2 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & z_1 & z_1^2 & 0 & 0 \\ \cdots & & & & & & \\ 1 & 0 & 0 & z_7 & z_7^2 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & z_1 & z_1^2 \\ \cdots & & & & & & \\ 1 & 0 & 0 & 0 & 0 & z_7 & z_7^2 \end{pmatrix}, \tag{D.3}$$

where the seven parameters are $\boldsymbol{b} = (a_0, b_1, c_1, b_2, c_2, b_3, c_3)^T$.

References

[1] Birge R T 1932 *Phys. Rev.* **40**(2) 207–227
[2] Mana G, Massa E and Predescu M 2012 *Metrologia* **49** 492–500
[3] Wuebbeler G, Bodnar O and Elster C 2016 *Metrologia* **53** 1131–1138
[4] Gull G F 1988 Bayesian inductive inferences and maximum entropy *Maximum entropy and Bayesian meyhods in science and engineering* ed Erickson G and Smith C (Dordrecht: Kluwer Academic Publishers) pp 53–74
[5] Nesseris S and García-Bellido J 2013 *Journal of Cosmology and Astroparticle Physics* **2013** 036
[6] Mana G, Albo P A G and Lago S 2014 *Measurement* **55** 564–570
[7] Amari S, Nagaoka H and Harada D 2007 *Methods of Information Geometry* Translations of Mathematical Monographs vol. 191 (Oxford: Oxford University Press)
[8] Arwini K and Dodson C 2008 *Information Geometry: Near Randomness and Near Independence* Lecture Notes in Mathematics 1953 (Berlin Heidelberg: Springer)
[9] Jeffreys H 1946 *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* **186** 453–461
[10] Jeffreys H 1998 *The Theory of Probability* (Oxford: Oxford University Press)
[11] Kass R E and Wasserman L 1996 *Journal of the American Statistical Association* **91** 1343–1370
[12] Costa S I, Santos S A and Strapasson J E 2014 *Discrete Applied Mathematics* **197** 59–69
[13] Golub G and Van Loan C 1996 *Matrix Computations* Johns Hopkins Studies in the Mathematical Sciences (Johns Hopkins University Press)
[14] Dhrymes P 1978 *Mathematics for Econometrics* (Berlin Heidelberg: Springer-Verlag)
[15] Berger J O, Bernardo J M and Sun D 2009 *Ann. Statist.* **37** 905–938
[16] Weisstein E W Incomplete gamma function `http://mathworld.wolfram.com/IncompleteGammaFunction.html` from MathWorld – A Wolfram Web Resource Accessed: 2017-01-12
[17] Wolfram Research Inc 2017 Mathematica 11.2
[18] Akaike H 1974 *IEEE Transactions on Automatic Control* **19** 716–723
[19] Burnham K and Anderson D 2002 *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach* (New York: Springer)
[20] Schwarz G 1978 *Ann. Statist.* **6** 461–464
[21] Konishi S and Kitagawa G 2008 *Information Criteria and Statistical Modeling* Springer series in statistics (New York: Springer)
[22] Bhat H and Kumar N 2010 On the derivation of the bayesian information criterion Tech. rep. School of Natural Sciences, University of California, Merced
[23] Mana G, Massa E, Sasso C P, Andreas B and U K 2017 *Metrologia* **54** 559–565
[24] Henderson H V and Searle S R 1981 *SIAM Review* **23** 53–60
[25] Tylavsky D J and Sohie G R L 1988 *Proceedings of the IEEE* **74** 1050–52