A&A 464, 399–404 (2007)
DOI: 10.1051/0004-6361:20066170

Astronomy & Astrophysics

Why your model parameter confidences might be too optimistic. Unbiased estimation of the inverse covariance matrix

J. Hartlap, P. Simon, and P. Schneider

Argelander-Institut* für Astronomie, Universität Bonn, Auf dem Hügel 71, 53121 Bonn, Germany
e-mail: hartlap@astro.uni-bonn.de

Received 3 August 2006 / Accepted 24 November 2006

ABSTRACT

Aims. The maximum-likelihood method is the standard approach to obtain model fits to observational data and the corresponding confidence regions. We investigate possible sources of bias in the log-likelihood function and its subsequent analysis, focusing on estimators of the inverse covariance matrix. Furthermore, we study under which circumstances the estimated covariance matrix is invertible.

Methods. We perform Monte-Carlo simulations to investigate the behaviour of estimators for the inverse covariance matrix, depending on the number of independent data sets and the number of variables of the data vectors.

Results. We find that the inverse of the maximum-likelihood estimator of the covariance is biased, the amount of bias depending on the ratio of the number of bins (data vector variables), p, to the number of data sets, n. This bias inevitably leads to an – in extreme cases catastrophic – underestimation of the size of confidence regions. We report on a method to remove this bias for the idealised case of Gaussian noise and statistically independent data vectors. Moreover, we demonstrate that marginalisation over parameters introduces a bias into the marginalised log-likelihood function. Measures of the sizes of confidence regions suffer from the same problem. Furthermore, we give an analytic proof for the fact that the estimated covariance matrix is singular if $p > n$.

Key words. methods: analytical – methods: data analysis – gravitational lensing

1. Introduction

The maximum-likelihood method (e.g. Barlow 1991) is common practice to obtain the best-fit parameters $\boldsymbol{\pi}_0$ and confidence regions from a measured data vector $\boldsymbol{d} \in \mathbb{R}^p$ for a model $\boldsymbol{m}(\boldsymbol{\pi})$. It usually consists of finding the maximum of the log-likelihood function

$$\mathcal{L}(\boldsymbol{d}|\boldsymbol{\pi}) \propto -\frac{1}{2}\left[\boldsymbol{d}-\boldsymbol{m}(\boldsymbol{\pi})\right]^{\mathrm{t}} \boldsymbol{\Sigma}^{-1} \left[\boldsymbol{d}-\boldsymbol{m}(\boldsymbol{\pi})\right], \tag{1}$$

where $\boldsymbol{\pi}$ is the parameter vector and a Gaussian distribution of the measurement errors is assumed. The confidence regions for the maximum-likelihood fit are then defined by the surfaces of constant $\Delta\mathcal{L} \equiv \mathcal{L}_{\max} - \mathcal{L}$, where $\mathcal{L}_{\max}$ is the maximum value of the log-likelihood function.

For the evaluation of the log-likelihood the population covariance matrix $\boldsymbol{\Sigma}$ and its inverse $\boldsymbol{\Sigma}^{-1}$ or estimates thereof are needed. In most cases, no exact analytical expression for $\boldsymbol{\Sigma}$ can be given, although numerous authors make use of analytical approximations. An example from the field of weak gravitational lensing is Semboloni et al. (2006), who use the Gaussian approximation to the covariance matrix of the shear correlation functions given by Schneider et al. (2002). Other possibilities are to estimate $\boldsymbol{\Sigma}$ from the data themselves (e.g. Hetterscheidt et al. 2006; Budavári et al. 2003) or to obtain it from a simulated data set whose properties are comparable to the original data (e.g. Pan & Szapudi 2005). In the latter paper, the authors observed that the estimated covariance matrix becomes singular if p, the number of entries of the data vectors, exceeds the number of observations / simulated data vectors. As a remedy, they propose to use the Singular Value Decomposition (SVD, Press et al. 1992) to obtain a pseudo-inverse of the covariance matrix, but do not investigate the properties of the resulting estimate of $\boldsymbol{\Sigma}^{-1}$ in detail. In this paper, we prove analytically that the rank of the standard estimator of the covariance matrix cannot exceed the number of observations. We then point out that, even if this estimator is not singular, simple matrix inversion yields a biased estimator of $\boldsymbol{\Sigma}^{-1}$. This may, if not corrected for, cause a serious underestimate of the size of the confidence regions. This problem has also been noticed by Hirata et al. (2004) and Mandelbaum et al. (2006), who use Monte-Carlo simulations to determine the correct confidence contours in cases where the covariance matrix is noisy. We report on the existence of a simpler method to remove this bias, which can be derived for Gaussian noise and statistically independent data vectors, and test the validity of this method when these assumptions are violated.

* Founded by merging of the Sternwarte, Radioastronomisches Institut and Institut für Astrophysik und Extraterrestrische Forschung der Universität Bonn.

2. The covariance matrix

2.1. Estimators

Let $\boldsymbol{d}$ be a vector of p random variables with components d_i, drawn from a multi-variate Gaussian distribution with population covariance matrix $\boldsymbol{\Sigma}$ and mean $\boldsymbol{\mu}$:

$$P(\boldsymbol{d}) = \frac{1}{(2\pi)^{p/2}\sqrt{\det \boldsymbol{\Sigma}}} \exp\left(-\frac{1}{2}(\boldsymbol{d}-\boldsymbol{\mu})^{\mathrm{t}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{d}-\boldsymbol{\mu})\right). \tag{2}$$

Furthermore, let $\boldsymbol{d}^{(k)}$ denote the kth realisation of this random vector, where $k \in [1, n]$ and n is the total number of realisations.

The well-known maximum-likelihood estimator for the components of the covariance matrix is given by (Barlow 1991)

$$\hat{\mathsf{C}}_{ij}^{\mathrm{ML}} = \frac{1}{n}\sum_{k=1}^{n}\left(d_i^{(k)} - \mu_i\right)\left(d_j^{(k)} - \mu_j\right), \tag{3}$$

which in the case of a known mean vector $\boldsymbol{\mu}$ is unbiased. If, however, $\boldsymbol{\mu}$ has to be estimated from the data, a correction factor of $n/(n-1)$ has to be applied to (3).

2.2. The rank of $\hat{C}^{ML}$

In the following, we prove that $\hat{\mathsf{C}}^{\mathrm{ML}}$ is singular for $p > n$ in case of known mean vector, and for $p > n-1$ if the mean vector is obtained from the data as well. For the first case, this can be seen by rewriting (3) as

$$\hat{\mathsf{C}}^{\mathrm{ML}} = \frac{1}{n}\sum_{k=1}^{n}\boldsymbol{d}^{(k)}\,\boldsymbol{d}^{(k)\mathrm{t}}, \tag{4}$$

where we presume, without loss of generality, that the mean vector is zero. Since the data vectors $\boldsymbol{d}^{(k)}$ are statistically independent, we can safely assume that they are linearly independent for $n \le p$ (for a continuous distribution, the probability to draw linearly dependent data vectors is zero). Therefore, $\left\{\boldsymbol{d}^{(k)}\right\}$ span an n-dimensional subspace U of $\mathbb{R}^p$. To check whether $\hat{\mathsf{C}}^{\mathrm{ML}}$ is singular we now try to find a vector $\boldsymbol{y} \neq \boldsymbol{0}$ for which $\hat{\mathsf{C}}^{\mathrm{ML}}\boldsymbol{y} = \boldsymbol{0}$. Looking at (4), we see that this is only possible for $p > n$, since in this case we can always choose a vector $\boldsymbol{y}$ from the subspace orthogonal to U, for which $\boldsymbol{d}^{(k)} \cdot \boldsymbol{y} = 0\ \forall\ k$. If $p \le n$, $\left\{\boldsymbol{d}^{(k)}\right\}$ already spans the whole of $\mathbb{R}^p$, and no vector can be found that is orthogonal to all $\boldsymbol{d}^{(k)}$. This proves that $\hat{\mathsf{C}}^{\mathrm{ML}}$ is singular for known mean vector if $p > n$.

We now prove our statement for an unknown mean vector $\boldsymbol{\mu}$, which is estimated from the data using

$$\boldsymbol{\mu} = \frac{1}{n}\sum_{k=1}^{n}\boldsymbol{d}^{(k)}. \tag{5}$$

For this, we define a new set of independent data vectors $\left\{\boldsymbol{w}^{(k)}\right\}$ by forming linear combinations of $\left\{\boldsymbol{d}^{(k)}\right\}$, specified by the orthogonal transformation B, of which we demand that the last (nth) row be given by $(1/\sqrt{n}, \ldots, 1/\sqrt{n})$ (Anderson 2003):

$$\boldsymbol{w}^{(k)} = \sum_{l=1}^{n}\mathsf{B}_{kl}\,\boldsymbol{d}^{(l)}. \tag{6}$$

Thanks to our choice of B_{nl}, we have $\boldsymbol{w}^{(n)} = \sqrt{n}\,\boldsymbol{\mu}$. Next, we rewrite $\hat{\mathsf{C}}^{\mathrm{ML}}$ by means of the new data vectors:

$$\hat{\mathsf{C}}^{\mathrm{ML}} = \frac{1}{n}\sum_{k=1}^{n}\boldsymbol{d}^{(k)}\,\boldsymbol{d}^{(k)\mathrm{t}} - \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{t}} \tag{7}$$

$$= \frac{1}{n}\sum_{k=1}^{n}\boldsymbol{w}^{(k)}\,\boldsymbol{w}^{(k)\mathrm{t}} - \frac{1}{n}\boldsymbol{w}^{(n)}\,\boldsymbol{w}^{(n)\mathrm{t}} \tag{8}$$

$$= \frac{1}{n}\sum_{k=1}^{n-1}\boldsymbol{w}^{(k)}\,\boldsymbol{w}^{(k)\mathrm{t}}. \tag{9}$$

The last expression is of the same form as (4) (except for the sum, which has one addend less), and so the same line of reasoning as above can be applied to show that $\hat{\mathsf{C}}^{\mathrm{ML}}$ is singular for $p > n-1$.

Another interesting implication of Eq. (9) is that the mean vector and the estimated covariance matrix are distributed independently (again see Anderson 2003), although they are computed from the same data vectors. First, note that $\boldsymbol{w}^{(i)}$ and $\boldsymbol{w}^{(j)}$ are statistically independent for $i \neq j$. This can be seen by computing the covariance between the two vectors:

$$\mathrm{Cov}\left(\boldsymbol{w}^{(i)}, \boldsymbol{w}^{(j)}\right) = \left\langle\left(\boldsymbol{w}^{(i)} - \boldsymbol{\nu}^{(i)}\right)\left(\boldsymbol{w}^{(j)} - \boldsymbol{\nu}^{(j)}\right)^{\mathrm{t}}\right\rangle \tag{10}$$

$$= \sum_{k,l=1}^{n}\mathsf{B}_{ik}\mathsf{B}_{jl}\left\langle\left(\boldsymbol{d}^{(k)} - \boldsymbol{\mu}\right)\left(\boldsymbol{d}^{(l)} - \boldsymbol{\mu}\right)^{\mathrm{t}}\right\rangle \tag{11}$$

$$= \sum_{k,l=1}^{n}\mathsf{B}_{ik}\mathsf{B}_{jl}\,\delta_{kl}\,\boldsymbol{\Sigma} \tag{12}$$

$$= \delta_{ij}\,\boldsymbol{\Sigma}. \tag{13}$$

Here, $\langle\cdot\rangle$ denotes the expectation value and (see Eq. (6))

$$\boldsymbol{\nu}^{(i)} = \left\langle\boldsymbol{w}^{(i)}\right\rangle = \sum_{j=1}^{n}\mathsf{B}_{ij}\left\langle\boldsymbol{d}^{(j)}\right\rangle = \boldsymbol{\mu}\sum_{j=1}^{n}\mathsf{B}_{ij} \tag{14}$$

is the mean value of $\boldsymbol{w}^{(i)}$.

Since $\hat{\mathsf{C}}^{\mathrm{ML}}$ does not depend on $\boldsymbol{w}^{(n)}$, which in turn is statistically independent of the remaining $\boldsymbol{w}^{(i)}$, this shows the independence of estimated mean and covariance.

3. The inverse covariance matrix

3.1. An unbiased estimator for $\boldsymbol{\Sigma}^{-1}$

From (3), an estimator for $\boldsymbol{\Sigma}^{-1}$ can be obtained by matrix inversion:

$$\hat{\mathsf{C}}_{*}^{-1} = \left(\hat{\mathsf{C}}^{\mathrm{ML}}\right)^{-1}. \tag{15}$$

This estimator is consistent, but *not unbiased* due to noise in $\hat{\mathsf{C}}^{\mathrm{ML}}$: the inverse of an unbiased estimator for some statistical variable X is in general not an unbiased estimator for X^{-1}. Indeed, in our case of Gaussian errors and statistically independent data vectors one can show (Anderson 2003) that the expectation value of $\hat{\mathsf{C}}_{*}^{-1}$ is *not* the inverse of the population covariance, but

$$\left\langle\hat{\mathsf{C}}_{*}^{-1}\right\rangle = \frac{N}{N-p-1}\,\boldsymbol{\Sigma}^{-1} \text{ for } p < N-1, \tag{16}$$

where $N = n$ if $\boldsymbol{\mu}$ is known and $N = n-1$ if the mean is estimated from the data. In the following, we will only pursue the latter case.

The amount of bias in $\hat{\mathsf{C}}_{*}^{-1}$ thus depends essentially on the ratio of the number of entries p in the data vectors (henceforth referred to as the number of bins) to the number of independent observations n. It leads to an underestimation of the size of confidence regions by making the log-likelihood function steeper, but it does not change the maximum-likelihood point and thus does not affect the parameter estimates themselves.

From (16) it follows that an unbiased estimator of $\boldsymbol{\Sigma}^{-1}$ is given by[1]

$$\hat{\mathsf{C}}^{-1} = \frac{n-p-2}{n-1}\,\hat{\mathsf{C}}_{*}^{-1} \text{ for } p < n-2. \tag{17}$$

3.2. Monte-Carlo experiments

To illustrate Eq. (17), and also to probe how the pseudo-inverse of the estimated covariance obtained by the Singular Value Decomposition behaves (see below), we perform the following experiment: first, we choose an analytical form for the population covariance $\mathsf{\Sigma}$. We use three different models:

$$\mathsf{\Sigma}^{\rm d,c}_{ij} = \sigma^2\,\delta_{ij}, \tag{18}$$

$$\mathsf{\Sigma}^{\rm d,l}_{ij} = \sigma^2\,\left[1 - i/(1+p_1)\right]\,\delta_{ij} \quad \text{and} \tag{19}$$

$$\mathsf{\Sigma}^{\rm nd}_{ij} = \sigma^2/(1+\epsilon|i-j|), \tag{20}$$

which initially are $p_1 \times p_1$ matrices. ϵ can be used to tune the degree of correlation in model (20); we choose $\epsilon = 0.05$.

We then create n data vectors of length p_1 according to $\boldsymbol{d}^{(k)} = \boldsymbol{m} + \boldsymbol{\gamma}^{(k)}(\mathsf{\Sigma}_1)$, where $\boldsymbol{\gamma}^{(k)}(\mathsf{\Sigma}_1)$ is a noise vector drawn from a multivariate Gaussian distribution with mean zero and covariance $\mathsf{\Sigma}_1$. The choice of the model vector $\boldsymbol{m}$ is arbitrary, and in fact for the present purpose it would be sufficient to set $\boldsymbol{m} = \boldsymbol{0}$. For later use, however, we choose the linear model $m_i = ax_i + b$, where $x_i = (x_{\rm max} - x_{\rm min})(i + 1/2)/p_1$ is the value of the free variable corresponding to the centre of the ith bin.

From this synthetic set of observations we estimate the mean data vector and the covariance matrix, which yields the estimator $\hat{\mathsf{C}}^{\rm ML}$. Next, both $\mathsf{\Sigma}$ and $\hat{\mathsf{C}}^{\rm ML}$ are inverted using the Singular Value Decomposition (see below). Finally, we compute the unbiased estimate $\hat{\mathsf{C}}^{-1}$ of the inverse covariance as given in (17).

To probe the dependence of the bias of the estimators for $\mathsf{\Sigma}^{-1}$, n new data vectors are created subsequently with $p_j = p_1/j$ bins, for all integer $j \in [2, \ldots, p_1/2]$, where the population covariance $\mathsf{\Sigma}_j$ for p_j bins can be obtained from the original $\mathsf{\Sigma}_1$ by averaging over $(j \times j)$-sub-blocks of $\mathsf{\Sigma}_1$. This strategy of re-binning has the advantage that the true covariance is known exactly for all p_j.

Since the bias in Eq. (16) is just a scalar factor, we record the traces of the estimators $\hat{\mathsf{C}}_*^{-1}$ and $\hat{\mathsf{C}}^{-1}$ for each number of bins p. To improve our statistics, we repeat the procedure outlined above 10^4 times and average over the traces computed in each step.

In Fig. 1, we plot the ratios of the trace of $\mathsf{\Sigma}^{-1}$ to the traces of $\hat{\mathsf{C}}_*^{-1}$ and $\hat{\mathsf{C}}^{-1}$, respectively. Not using the bias-corrected $\hat{\mathsf{C}}^{-1}$ can have considerable impact on the size of confidence regions of parameter estimates: for $p < n - 2$, the components of $\hat{\mathsf{C}}_*^{-1}$ will be too large compared to those of the true inverse covariance, and the log-likelihood will decrease too steeply, resulting in confidence contours too small.

We also plot the traces of $\hat{\mathsf{C}}_*^{-1}$ for the different covariance models beyond $p \geq n - 1$, where the estimator $\hat{\mathsf{C}}^{\rm ML}$ is singular. These data points have been obtained using the Singular Value Decomposition to invert the covariance matrix, yielding a decomposition of the form

$$\mathsf{C} = \mathsf{U}\mathsf{W}\mathsf{V}^{\rm t}, \tag{21}$$

where U and V are orthogonal matrices and W is a diagonal matrix containing the singular values. Since C is symmetric, one has in addition $\mathsf{U} = \mathsf{V}$, while W contains the moduli of the eigenvalues of C. The inverse of C is then given by $\mathsf{C}^{-1} = \mathsf{V}\mathsf{W}^{-1}\mathsf{U}^{\rm t}$. If C is singular, some of the entries of W will be zero or comparable to machine precision. We therefore can only compute a pseudo-inverse of C by replacing the inverses of these singular values in W^{-1} by zero, as has been suggested in Press et al. (1992) and Pan & Szapudi (2005). Figure 1 shows that the bias of $\hat{\mathsf{C}}_*^{-1}$ in this regime depends significantly on the covariance model chosen and does not depend on binning in a simple way. Therefore we strongly discourage from the use of the SVD for $p > n - 1$.

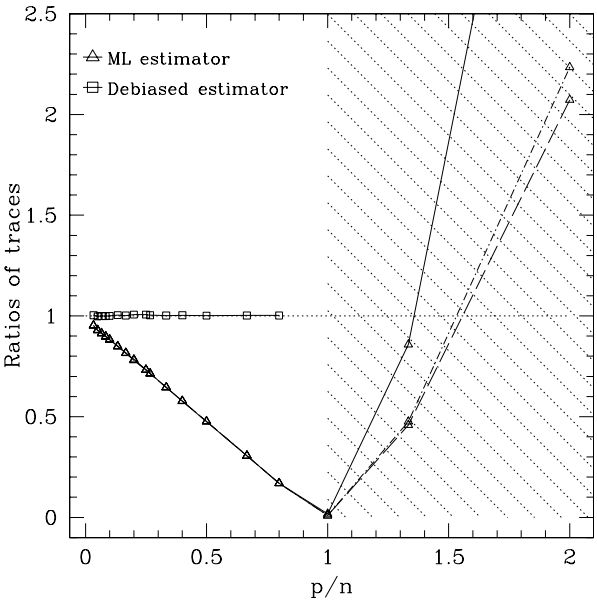

Fig. 1. Ratios of the trace of $\mathsf{\Sigma}^{-1}$ to the traces of C_*^{-1} (triangles) and $\hat{\mathsf{C}}^{-1}$ (squares), respectively. The dashed line is for the covariance model (18), the solid line for (19) and the dot-dashed-line for (20). The original data vectors had $p_1 = 240$ bins, and were rebinned by subsequently joining 2, 3, ... of the original bins. The number of independent observations is $n = 60$. Error bars are comparable to the symbol size and therefore omitted.

4. Implications for likelihood analysis

Having obtained an unbiased estimator of the inverse covariance matrix, one may still be concerned about a possible bias in the log-likelihood function, since it consists of the product of $(\boldsymbol{d} - \boldsymbol{\mu})$ and $\hat{\mathsf{C}}^{-1}$ (Eq. (1)), since $\boldsymbol{\mu}$ and $\hat{\mathsf{C}}^{-1}$ are estimated from the same set of observations. In other words, the question is if it is possible to write

$$\langle \mathcal{L}(\boldsymbol{d}|\boldsymbol{\pi})\rangle = -\frac{1}{2}\,\left\langle (\boldsymbol{\mu} - \boldsymbol{m})^{\rm t}\,\hat{\mathsf{C}}^{-1}\,(\boldsymbol{\mu} - \boldsymbol{m})\right\rangle \tag{22}$$

$$= -\frac{1}{2}\,(\langle\boldsymbol{\mu}\rangle - \boldsymbol{m})^{\rm t}\left\langle\hat{\mathsf{C}}^{-1}\right\rangle(\langle\boldsymbol{\mu}\rangle - \boldsymbol{m})\,. \tag{23}$$

Luckily, this is indeed the case, since we have shown at the end of Sect. 2.2 that mean vector and covariance matrix are distributed independently.

4.1. Marginalised likelihood

Usually, one is not only interested in the full parameter space of a problem, but also in values of single parameters and the corresponding errors. In the Bayesian framework, this is usually

[1] Note that there is a typing error in Anderson's book, where he gives an expression corresponding to

$$\hat{\mathsf{C}}^{-1} = \frac{n-p-2}{n-\mathbf{2}}\,\hat{\mathsf{C}}_*^{-1} \text{ for } p < n-2.$$

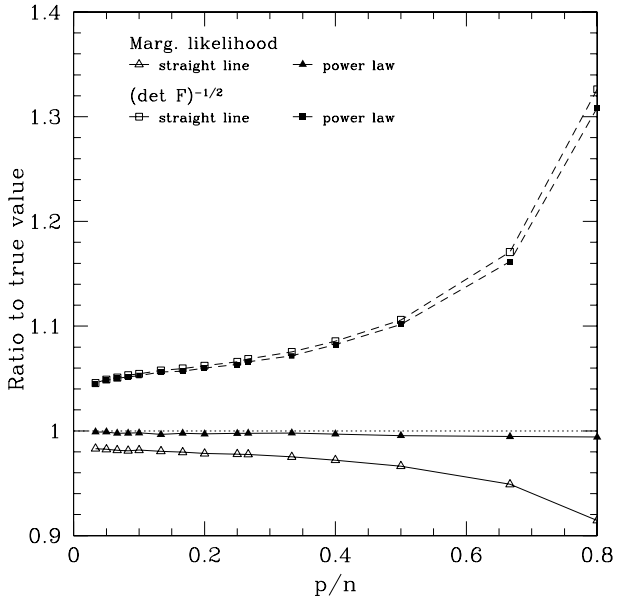

Fig. 2. *Triangles, solid lines*: ratio of the sum over all pixels of the marginalised likelihood computed using $\hat{\mathsf{C}}^{-1}$ and the true marginalised likelihood. Filled triangles are for the power-law fit (marginalised over the power-law index), open triangles are for the straight line fit (marginalised over the intercept). *Squares, dashed lines*: ratio of $\sqrt{\det \mathsf{F}^{-1}}$ using $\hat{\mathsf{C}}^{-1}$ to the true one, computed with $\mathsf{\Sigma}$. For both cases $\mathsf{\Sigma} = \mathsf{\Sigma}^{\mathrm{d,c}}$.

achieved by marginalising over the "uninteresting" parameters: the log-likelihood function $\mathcal{L}_i$ for the single parameter π_i is computed using

$$\mathcal{L}_i(\boldsymbol{d}|\pi_i) = \log\left\{\left[\prod_{j\neq i}\int \mathrm{d}\pi_j\right] \exp\left[\mathcal{L}(\boldsymbol{d}|\boldsymbol{\pi})\right]\right\}. \tag{24}$$

There is no reason to believe that the marginalised log-likelihood, which is a highly non-linear function of the (unbiased) estimate of the full log-likelihood, and with it the size of the errors on π_i are unbiased, even if one uses the unbiased $\hat{\mathsf{C}}^{-1}$. We demonstrate this by means of our simulated fitting procedure, where we now use in addition to the straight line model also a second simulation using a power-law model of the form $m_i = a\, x_i^b$. We marginalise over the intercept of the line and the power-law index, respectively. We record the sums over all pixels of the (one-dimensional) grid of the marginalised log-likelihood functions, which we compute using the true $\mathsf{\Sigma}^{-1}$ and the unbiased estimator $\hat{\mathsf{C}}^{-1}$. For $\mathsf{\Sigma}$, we choose the model (18). We average over $\approx 3 \times 10^4$ repetitions of these experiments. We plot the ratio of true to estimated log-likelihood sums in Fig. 2 (triangles and solid lines). The plot shows a bias of maximally $\approx$8% for the straight line and even less for the power-law, in a direction which would lead to an overestimation of the error bars on slope and amplitude. Although the effect is not very large, this is not guaranteed to remain so for models different from the ones considered here.

4.2. Measuring the size of confidence regions

For some applications, it is useful to have a simple measure of the size of the confidence regions. As an example, we make use of the Fisher information matrix F (Fisher 1935), which is defined by (Tegmark et al. 1997)

$$\mathsf{F} \equiv \left\langle \frac{\partial^2 \mathcal{L}}{\partial \pi_i\, \partial \pi_j} \right\rangle, \tag{25}$$

where the derivative is to be evaluated at the maximum-likelihood point $\boldsymbol{\pi}_0$. F can be interpreted as an estimate of the inverse covariance matrix of the parameter estimates, provided $\mathcal{L}$ is well approximated by a Gaussian around the maximum-likelihood point.

To demonstrate the bias in $\sqrt{\det \mathsf{F}^{-1}}$, we compute the Fisher matrix for the straight line and power-law fits using (Tegmark et al. 1997)

$$F_{ij} = \sum_{\alpha,\beta=1}^{p} \frac{\partial m_\alpha}{\partial \pi_i} \frac{\partial m_\beta}{\partial \pi_j}\, \mathsf{C}^{-1}_{\alpha\beta}, \tag{26}$$

which is valid if the covariance matrix does not depend on the parameters π_i; $\boldsymbol{m}$ is the model vector.

In Fig. 2, we give the ratio of $\sqrt{\det \mathsf{F}^{-1}}$, computed using the unbiased estimated covariance $\hat{\mathsf{C}}^{-1}$, to the value computed using the true covariance (boxes and dashed lines). One sees that in this case the size of the confidence regions is significantly overestimated, for p/n approaching unity by as much as $\approx$30% for the straight line case, and by a comparable, albeit slightly smaller factor for the power-law fit.

5. Bootstrapping and non-Gaussian statistics

The derivation of the unbiased estimator $\hat{\mathsf{C}}^{-1}$ rests on the assumptions of Gaussian noise and statistically independent data vectors. To test the performance of this estimator in real world situations, where one or both of these assumptions may be violated, we make use of an example from the domain of weak gravitational lensing. For an introduction to this field we refer the reader to Bartelmann & Schneider (2001).

We simulate a weak lensing survey consisting of one single field, containing N_g galaxies, which are assigned a random ellipticity ϵ. $\epsilon = \epsilon_1 + \mathrm{i}\epsilon_2$ is a complex number, which is related to the quadrupole moment of the light distribution of a galaxy (see Bartelmann & Schneider 2001). The two components of the ellipticity are drawn from a Gaussian distribution with dispersion $\sigma_\epsilon/\sqrt{2}$.

The goal of the survey is to measure the shear correlation function $\xi_+(\vartheta)$ and to fit a model prediction to it. An estimator for ξ_+ is given by (Schneider et al. 2002)

$$\hat{\xi}_+(\vartheta) = \frac{\sum_{ij}\left(\epsilon_1^{(i)}\epsilon_1^{(j)} + \epsilon_2^{(i)}\epsilon_2^{(j)}\right)\, \Delta_\vartheta\left(|\boldsymbol{\theta}_i - \boldsymbol{\theta}_j|\right)}{2n_\mathrm{p}(\vartheta)}, \tag{27}$$

where the galaxies are labelled with i and j and have the angular positions $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_j$. $\Delta_\vartheta(\phi)$ is unity if $\vartheta - \Delta\vartheta/2 < \phi \leq \vartheta + \Delta\vartheta/2$, where $\Delta\vartheta$ is the bin width, and zero otherwise. Finally, $n_\mathrm{p}(\vartheta)$ is the number of pairs of galaxies contributing to the correlation function in the bin centred on ϑ.

We also need the covariance matrix of $\xi_+(\vartheta)$, which, since we only have one measurement, is estimated using the bootstrapping algorithm (e.g. Efron & Tibshirani 1993): First, we create a catalogue of all $N_\mathrm{p} = N_\mathrm{g}(N_\mathrm{g}-1)/2$ possible pairs of galaxies in the field. We then create N_bs bootstrap realisations of the survey by repeatedly drawing N_p pairs with replacement from the catalogue. From these, we estimate the mean data vector and the

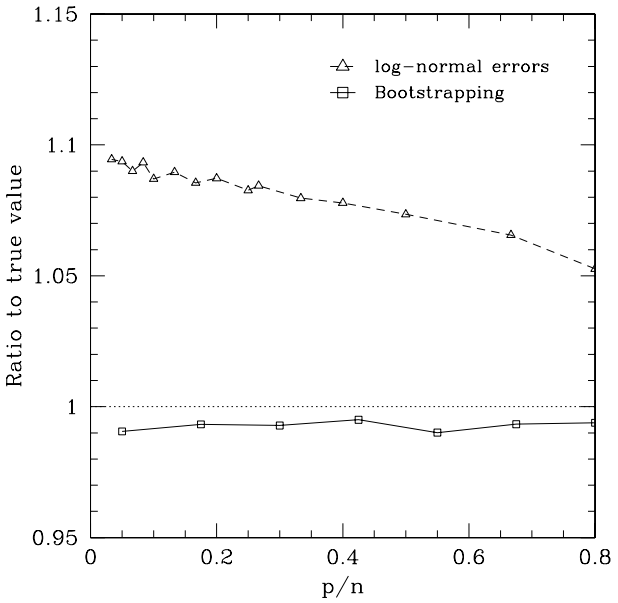

Fig. 3. Ratio of the traces of the unbiased estimator $\hat{\mathsf{C}}_*^{-1}$ and $\hat{\mathsf{C}}^{-1}$ to the trace of $\mathsf{\Sigma}^{-1}$; the covariances for the solid curve have been estimated using bootstrapping (see text), the dashed line shows the ratio of the traces for log-normal errors.

covariance matrix of the shear correlation function. As before, we do this for various numbers of bins, where we record the dependence of the traces of $\mathsf{\Sigma}^{-1}$, $\hat{\mathsf{C}}_*^{-1}$ and $\hat{\mathsf{C}}^{-1}$ on binning. For the simple case of pure shape noise, the population covariance is diagonal and can be easily computed using (Schneider et al. 2002)

$$\mathsf{\Sigma}_{ij} = \frac{\sigma_\epsilon^4}{2\, n_\mathrm{p}(\vartheta_i)}\, \delta_{ij}, \tag{28}$$

where ϑ_i is the angular separation corresponding to the centre of the ith bin. We precompute the function n_p numerically from a large set of independent data fields for all binning parameters we wish to use in the simulation.

In principle, both of the assumptions made for the derivation of Eq. (17) are violated: the noise in the shear correlation function is χ^2-distributed, because $\xi_+ \propto \epsilon\epsilon$, where ϵ is drawn from a Gaussian. However, the number of degrees of freedom of the χ^2-distribution, which equals the number of pairs, is very large, so that it is very well approximated by a Gaussian (central limit theorem). We therefore do not expect any significant influence on the performance of $\hat{\mathsf{C}}^{-1}$. We expect a larger impact by the fact that the data vectors resulting from the bootstrapping procedure are not statistically independent, since different bins necessarily contain the identical galaxy pairs. Strictly speaking, also the requirements for the application of the bootstrap procedure are not met, since the pairs of galaxies which we use to sample the distribution of the shear correlation function are not statistically independent. However, we argue that drawing individual galaxies instead of pairs is not correct, since this would sample the distribution of ϵ, not the one of ξ_+.

The outcome of $\approx 2 \times 10^4$ realisations of this experiment is given in Fig. 3 (solid line), with $N_\mathrm{g} = 500$ and $N_\mathrm{bs} = 40$. The figure shows that, in spite of the correlations among the pairs of galaxies and the data vectors, $\hat{\mathsf{C}}^{-1}$ is wrong by only $\approx 1\%$, and may well be used in bootstrap applications like this.

Finally, we explore the impact of non-Gaussian noise. For this purpose, we perform the same experiment as before, only replacing the Gaussian noise vectors ones, $\gamma^{(k)}$, with a log-normal distribution. These are computed using

$$\gamma_i^{(k)} = \exp\left(r_i^{(k)} - 1/2\right), \tag{29}$$

where $\boldsymbol{r}^{(k)}$ are vectors containing uncorrelated, Gaussian random variables with mean zero and variance $\sigma^2 = 1$. The result is shown in Fig. 3 (dashed line). Clearly, Eq. (17) is no longer applicable, although for $p/n > 0.2$, one still does much better with it than without it.

6. Summary and conclusions

We have given a proof for the fact that the standard estimator of the covariance matrix (3) is singular for $p > n$ if the mean vector used in (3) is known, and for $p > n - 1$ if the mean is estimated from the same data set as the covariance matrix. Furthermore, we noted that the inverse of the maximum-likelihood estimator of the covariance matrix is a biased estimator of the inverse population covariance matrix. This bias depends basically on the ratio of the number of bins to the number of independent observations and can be quite severe as these two numbers become comparable. If uncorrected for, it will lead to a significant underestimation of the size of confidence regions derived from maximum-likelihood fits. The bias can be corrected for $p < n - 2$ by using the estimator (17) instead, which was derived by Anderson (2003) under the assumption of Gaussian errors and statistically independent data vectors. We stress that there is no contradiction between the foregoing two statements: The singularity of $\hat{\mathsf{C}}_\mathrm{ML}$ for $p > n - 1$ derives from linear algebra alone, whereas the fact that $\hat{\mathsf{C}}^{-1}$ is zero for $p = n - 2$ is due to the statistical distribution of the covariance matrix. Going beyond $p = n - 1$, we find that it is not advisable to use the Singular Value Decomposition to invert the estimated covariance matrix, since the bias of the pseudo-inverse does not seem to be controllable and depends strongly on the population covariance matrix, which is a priori unknown.

Given the unbiased estimator $\hat{\mathsf{C}}^{-1}$, we argue that also the log-likelihood function is unbiased. However, great care has to be taken if one wishes to perform further analysis of $\mathcal{L}$: since it is a statistical variable, any nonlinear operation on it has the potential to cause a bias. We demonstrate this for the case of marginalisation over certain parameters, where the bias is relatively mild for the examples we chose. The situation is much worse if one tries to quantify the size of the confidence regions. The square root of the determinant of the inverse Fisher matrix shows a significant amount of bias, and therefore should not be used as absolute measures of measurement uncertainty.

The upshot of all this is the following: avoid to use more bins for your likelihood fit than you have realisations of your data. If your errors are Gaussian and the data vectors are statistically independent, use the estimator $\hat{\mathsf{C}}^{-1}$ to obtain an unbiased estimate of the inverse covariance matrix and the log-likelihood function. If one or both of these two requirements are not fulfilled, the estimator is not guaranteed to work satisfactorily; this should be checked from case to case.

Finally, we note that the estimates of the covariance matrix and its inverse can be quite noisy. If one has prior knowledge about the structure of the covariance matrix, one can develop estimators with a much lower noise level. Since this noise is responsible for most of the problems discussed in this paper, these improved estimators may also be useful in situations where the

requirements for the use of $\hat{\mathsf{C}}^{-1}$ are not fulfilled. We will explore these possibilities in a future paper.

Acknowledgements. We wish to thank Tim Eifler for useful suggestions, Jack Tomsky for pointing us to Anderson's book and Jens Rödiger, Martin Kilbinger, Ismael Tereno and Marco Hetterscheidt for careful reading of the manuscript.

References

Anderson, T. W. 2003, An introduction to multivariate statistical analysis, 3rd edn. (Wiley-Interscience)
Barlow, R. J. 1991, Statistics (John Wiley & Sons)
Bartelmann, M., & Schneider, P. 2001, Phys. Rep., 340, 291
Budavári, T., Connolly, A. J., Szalay, A. S., et al. 2003, ApJ, 595, 59
Efron, B., & Tibshirani, R. J. 1993, An Introduction to the bootstrap (Chapman & Hall/CRC)
Fisher, R. 1935, J. Roy. Stat. Soc., 98, 39
Hetterscheidt, M., Simon, P., Schirmer, M., et al. 2006 [arXiv:astro-ph/0606571]
Hirata, C. M., Mandelbaum, R., Seljak, U., et al. 2004, MNRAS, 353, 529
Mandelbaum, R., Hirata, C. M., Ishak, M., Seljak, U., & Brinkmann, J. 2006, MNRAS, 367, 611
Pan, J., & Szapudi, I. 2005, MNRAS, 362, 1363
Press, W., et al. 1992, Numerical Recipes in C (Cambridge University Press)
Schneider, P., van Waerbeke, L., Kilbinger, M., & Mellier, Y. 2002, A&A, 396, 1
Semboloni, E., Mellier, Y., van Waerbeke, L., et al. 2006, A&A, 452, 51
Tegmark, M., Taylor, A. N., & Heavens, A. F. 1997, ApJ, 480, 22