Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **456,** L132–L136 (2016) doi:10.1093/mnrasl/slv190

# Parameter inference with estimated covariance matrices

Elena Sellentin[1]★ and Alan F. Heavens[2]

[1]*Institut für Theoretische Physik, Ruprecht-Karls-Universität Heidelberg, Philosophenweg 16, D-69120 Heidelberg, Germany*
[2]*Imperial Centre for Inference and Cosmology (ICIC), Department of Physics, Imperial College, Blackett Laboratory, Prince Consort Road, London SW7 2AZ, UK*

Accepted 2015 November 26. Received 2015 November 18; in original form 2015 September 29

**ABSTRACT**

When inferring parameters from a Gaussian-distributed data set by computing a likelihood, a covariance matrix is needed that describes the data errors and their correlations. If the covariance matrix is not known a priori, it may be estimated and thereby becomes a random object with some intrinsic uncertainty itself. We show how to infer parameters in the presence of such an estimated covariance matrix, by marginalizing over the true covariance matrix, conditioned on its estimated value. This leads to a likelihood function that is no longer Gaussian, but rather an adapted version of a multivariate *t*-distribution, which has the same numerical complexity as the multivariate Gaussian. As expected, marginalization over the true covariance matrix improves inference when compared with Hartlap et al.'s method, which uses an unbiased estimate of the inverse covariance matrix but still assumes that the likelihood is Gaussian.

**Key words:** methods: data analysis – methods: statistical – cosmology: observations.

## 1 INTRODUCTION

A very common problem in statistical inference concerns data that are Gaussian-distributed. The likelihood of the observed data $\boldsymbol{X}_{\rm o}$ is a multivariate Gaussian, characterized only by a mean data vector $\boldsymbol{\mu}$ and a covariance matrix $\boldsymbol{\Sigma}$:

$$G(\boldsymbol{X}_{\rm o}|\boldsymbol{\mu}, \Sigma) = \frac{1}{\sqrt{|2\pi\Sigma|}} \exp\left[-\frac{1}{2}(\boldsymbol{X}_{\rm o}-\boldsymbol{\mu})^T \Sigma^{-1}(\boldsymbol{X}_{\rm o}-\boldsymbol{\mu})\right]. \quad (1)$$

The posterior probability of the parameters is proportional to the likelihood, now treated as a function of the parameters (through the dependence of the mean and the covariance matrix), multiplied by a suitable prior. Ideally one has analytic expressions for the mean and covariance in terms of the model parameters, but in many cases these are not available, and one or both may need to be estimated from simulated data which mimic the experiment that is to be analysed (e.g. Semboloni et al. 2006; Heymans et al. 2013), or from the data themselves (e.g. Budavári et al. 2003). However, although an unbiased simulated covariance matrix $\mathbf{S}$ can be constructed, its inverse is not an unbiased estimator of the inverse (or precision) matrix $\boldsymbol{\Sigma}^{-1}$, which is what is needed in the likelihood equation (1). One can construct an unbiased estimator of $\boldsymbol{\Sigma}^{-1}$ by a rescaling of $\mathbf{S}$ (Anderson 2003), as advocated by Hartlap, Simon & Schneider (2007). This widens up the credible intervals. If simulations are computationally cheap, then one can generate a large number $N$ of simulated data sets and obtain an accurate estimate of the covariance matrix. This asymptotic regime occurs only when $N$ far exceeds the size of the data vector, $p$. In many practical cases this is not possible, and the number of simulated data sets is small, with the consequence that statistical noise in the precision matrix propagates into errors in the parameters (Hamimeche & Lewis 2009; Dodelson & Schneider 2013; Taylor, Joachimi & Kitching 2013). However, there is a more fundamental difficulty with the approach adopted, as it assumes that the likelihood is still Gaussian, albeit with a different precision matrix, whereas in fact it is not.

A principled way to tackle the problem is to recognize that the simulated data provide *samples* of the covariance matrix, so $\mathbf{S}$ is itself a random object, based on a number of simulations. For Gaussian data, we have the advantage that the sample distribution of $\mathbf{S}$ is known, for a given true covariance matrix $\boldsymbol{\Sigma}$, and we can exploit this, with a suitable prior, by constructing the probability of $\boldsymbol{\Sigma}$ conditional on the sample $\mathbf{S}$, and then marginalizing over the unknown covariance matrix $\boldsymbol{\Sigma}$. This can be done analytically for our preferred choice of Jeffreys prior for $\boldsymbol{\Sigma}$. As a consequence, we properly propagate the uncertainty in the covariance matrix into the final inference, computing the quantity we want, i.e. the likelihood given the *simulated* covariance matrix $\mathbf{S}$ and the number of samples $N$ on which is it based: $P(\boldsymbol{X}_{\rm o}|\boldsymbol{\mu}, \mathbf{S}, N)$. This object, where we keep the dependence on the number of simulated data sets $N$ explicit to emphasize its importance, is the main result of this Letter. It is not Gaussian, but rather follows a modified version of the multivariate *t*-distribution. In practical terms, it is no more expensive to compute than the Hartlap-scaled Gaussian likelihood, but statistically sound, and can be retrospectively applied to many analyses that have used a

★ E-mail: sellentin@stud.uni-heidelberg.de

different likelihood function by appropriate re-weighting of points, provided that the chains adequately sample the parameter space that the $t$-distribution favours.

## 2 REPLACING A TRUE COVARIANCE MATRIX BY AN ESTIMATOR

When inferring cosmological model parameters $\boldsymbol{\theta}$ from a data set, we usually have just one $p$-dimensional observed data vector $\boldsymbol{X}_{\rm o}$, which is a single realization of a statistical process which we assume to be a multivariate Gaussian of which the mean $\boldsymbol{\mu}$, and the covariance matrix $\boldsymbol{\Sigma}$ may depend on the parameters $\boldsymbol{\theta}$

$$\boldsymbol{X}_{\rm o} \sim \mathcal{N}_{\rm p}\left[\boldsymbol{\mu}(\boldsymbol{\theta}), \Sigma(\boldsymbol{\theta})\right]. \tag{2}$$

In the following, we suppress this dependence on the parameters but it is still implied.

If $\boldsymbol{\Sigma}$ were known precisely, the likelihood would be the Gaussian, equation (1). However, if $\boldsymbol{\Sigma}$ is unknown, and all we have is an estimator $\mathbf{S}$, then the likelihood $G(\boldsymbol{X}_{\rm o}|\boldsymbol{\mu}, \Sigma)$ must be replaced by another likelihood of which we will show that it is not a Gaussian.

One method – viable for Frequentists – of estimating the covariance matrix, is to draw further independent data vectors from the distribution of $\boldsymbol{X}_{\rm o}$ and to calculate their sample covariance. Typically, such repeated independent measurements are however impossible in cosmology. None the less, if we can simulate the observation, then we are able to generate further samples $\boldsymbol{X}_i; \quad i = 1, \ldots N$, that are statistically equivalent to the single observation $\boldsymbol{X}_{\rm o}$. The covariance matrix $\mathbf{S}$ can then be estimated from these simulations, and the likelihood that we require is the probability of the data, given $\mathbf{S}$ and the number of simulations on which it is based, i.e. $P(\boldsymbol{X}_{\rm o}|\boldsymbol{\mu}, \mathbf{S}, N)$.

If we run $N$ independent simulations, then $\bar{\boldsymbol{X}} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{X}_i$ is the average, and an unbiased estimator of $\boldsymbol{\Sigma}$ is

$$\mathbf{S} = \frac{1}{N-1}\sum_{i=1}^{N}(\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})^T. \tag{3}$$

In the following subsection, we derive an analytical replacement for the Gaussian likelihood, equation (1), conditioned on an estimate $\mathbf{S}$, and from Section 4.1 onwards we study the effects of this replacement on parameter inference.

### 2.1 Derivation of the multivariate $t$-distribution

We now derive the likelihood $P(\boldsymbol{X}_{\rm o}|\boldsymbol{\mu}, \mathbf{S}, N)$ that depends on an estimator $\mathbf{S}$ instead of the true covariance $\boldsymbol{\Sigma}$.

Any matrix of the type $\mathbf{M} = \sum_{i=1}^{m} \boldsymbol{Y}_i \boldsymbol{Y}_i^T$ is by construction a Wishart matrix (Wishart 1928; Mardia, Kent & Bibby 1979; Anderson 2003), if $\boldsymbol{Y}$ is drawn from a multivariate Gaussian. When estimating a covariance matrix by averaging over random samples drawn from simulations, the estimated covariance matrix is a Wishart matrix, and has a Wishart distribution (Anderson 2003),

$$\mathcal{W}(\mathbf{S}|\Sigma/n, n) = \frac{|\mathbf{S}|^{\frac{n-p-1}{2}} \exp\left[-\frac{1}{2}n\mathrm{Tr}\left(\Sigma^{-1}\mathbf{S}\right)\right]}{2^{\frac{pn}{2}}|\Sigma/n|^{\frac{n}{2}}\Gamma_p\left(\frac{n}{2}\right)}, \tag{4}$$

where we call $n = N - 1$ the degrees of freedom and $\Gamma_p$ is the $p$-dimensional Gamma function. By the central limit theorem, this distribution is also asymptotically appropriate if the sampling distribution of $\boldsymbol{X}$ is non-Gaussian.

We can invert this distribution to yield the distribution $P(\boldsymbol{\Sigma}|\mathbf{S}, N)$ of the true covariance matrix $\boldsymbol{\Sigma}$ conditioned on the estimator $\mathbf{S}$, by using Bayes' Theorem

$$P(\Sigma|\mathbf{S}, N)\pi(\mathbf{S}) = \mathcal{W}(\mathbf{S}|\Sigma/n, n)\pi(\Sigma) \tag{5}$$

and adopting priors $\pi$. Since the determinant of the positive-definite covariance matrix is strictly positive, it is a scaling parameter, and we therefore assume the independence-Jeffreys prior (Jeffreys 1961; Sun & Berger 2006)

$$\pi(\Sigma) \propto |\Sigma|^{-\frac{p+1}{2}}. \tag{6}$$

This is by construction invariant under reparametrizations, and can therefore be regarded as uninformative, independent of the choice of parameters.[1] We then have

$$\begin{aligned} P(\Sigma|\mathbf{S}, N) &\propto \mathcal{W}(\mathbf{S}|\Sigma/n, n)\pi(\Sigma) \\ &\propto |\Sigma|^{-\frac{n+p+1}{2}} \exp\left[-\frac{1}{2}n\mathrm{Tr}\left(\Sigma^{-1}\mathbf{S}\right)\right] \\ &\propto \mathcal{W}^{-1}(\Sigma|n\mathbf{S}, n), \end{aligned} \tag{7}$$

showing that the uncertainty of the unknown true $\boldsymbol{\Sigma}$ can be described by an inverse Wishart distribution, conditioned on the sample estimate,

$$\mathcal{W}^{-1}(\Sigma|\mathbf{C}, n) = \frac{|\mathbf{C}|^{\frac{n}{2}}|\Sigma|^{-\frac{n+p+1}{2}} \exp\left(-\frac{1}{2}\mathrm{Tr}\left(\Sigma^{-1}\mathbf{C}\right)\right)}{2^{\frac{np}{2}}\Gamma_p\left(\frac{n}{2}\right)}, \tag{8}$$

where we used $\mathbf{C} = n\mathbf{S}$. Increasing the estimates, $N = n + 1$, of the covariance matrix, will make this distribution more sharply peaked, reflecting the improvement of the estimation.

Given the distribution equation (8), we can now marginalize the Gaussian likelihood over the unknown covariance, to find what we are after, which is the likelihood of the data $\boldsymbol{X}_{\rm o}$, given a mean $\boldsymbol{\mu}$ and an estimate $\mathbf{S}$ of the covariance matrix from $N$ simulations:

$$\begin{aligned} P(\boldsymbol{X}_{\rm o}|\boldsymbol{\mu}, \mathbf{S}, N) &= \int \mathrm{d}\Sigma\, G(\boldsymbol{X}_{\rm o}|\boldsymbol{\mu}, \Sigma)P(\Sigma|\mathbf{S}, N) \\ &\propto \int \mathrm{d}\Sigma\, |\Sigma|^{-\frac{N+p+1}{2}} \exp\left[-\frac{1}{2}\mathrm{Tr}\left(\Sigma^{-1}\mathbf{Q}\right)\right], \end{aligned} \tag{9}$$

where we have defined $\mathbf{Q} = n\mathbf{S} + (\boldsymbol{X}_{\rm o} - \boldsymbol{\mu})(\boldsymbol{X}_{\rm o} - \boldsymbol{\mu})^T$. The last line is structurally the integration over an unnormalized inverted Wishart distribution $\mathcal{W}^{-1}(\Sigma|\mathbf{Q}, N)$, so the result is the normalization constant as in equation (8), leading to

$$P(\boldsymbol{X}_{\rm o}|\boldsymbol{\mu}, \mathbf{S}, N) \propto |\mathbf{Q}|^{-\frac{N}{2}}. \tag{10}$$

Resubstituting $\mathbf{Q}$, using the matrix identity

$$|\mathbf{A} + \boldsymbol{b}\boldsymbol{b}^T| = |\mathbf{A}|(1 + \boldsymbol{b}^T\mathbf{A}^{-1}\boldsymbol{b}) \tag{11}$$

and normalizing, we arrive at the likelihood for the $p$-dimensional data set $\boldsymbol{X}_{\rm o}$, conditioned on the mean $\boldsymbol{\mu}$ and a sample of the covariance matrix $\mathbf{S}$ from $N$ simulations:

$$P(\boldsymbol{X}_{\rm o}|\boldsymbol{\mu}, \mathbf{S}, N) = \frac{\bar{c}_{\rm p}|\mathbf{S}|^{-1/2}}{\left[1 + \frac{(\boldsymbol{X}_{\rm o}-\boldsymbol{\mu})^T\mathbf{S}^{-1}(\boldsymbol{X}_{\rm o}-\boldsymbol{\mu})}{N-1}\right]^{\frac{N}{2}}}. \tag{12}$$

This is a cosmologist's version of a multivariate $t$-distribution. It is not the standard (Frequentist) multivariate $t$-distribution, which

[1] The power $(p + 1)/2$ also leads to $N - 1$ degrees of freedom in the inverse Wishart distribution, which is an intuitive result. Another power would only change the degrees of freedom, showing that the influence of the prior can be lessened by increasing the number of simulations $N$.

jointly estimates the mean and its covariance from a data set of $N$ data vectors. In contrast, we have assumed exactly one data vector that determines where the likelihood will peak – and $N$ simulated vectors from which we estimate the covariance. The normalization is

$$\bar{c}_p = \frac{\Gamma\left(\frac{N}{2}\right)}{[\pi(N-1)]^{p/2}\,\Gamma\left(\frac{N-p}{2}\right)}, \tag{13}$$

where $\Gamma$ is the usual Gamma function and we require $N > p$. For expensive simulations, when a feasible $N$ is still comparable to $p$, the differences between a Gaussian and the $t$-distribution become important. So if a covariance matrix must be replaced by an estimator from simulations, the modified $t$-distribution equation (12) replaces the multivariate Gaussian equation (1). This is the main result of the Letter.

## 3 ATTEMPTING TO DEBIAS A GAUSSIAN LIKELIHOOD

Instead of using the $t$-distribution equation (12) it has become standard in cosmology to follow a procedure outlined by Hartlap et al. (2007), where the authors propose to stick with a Gaussian likelihood, and only to replace the true inverse covariance matrix by a scaled inverse sample covariance matrix

$$\Sigma^{-1} \rightarrow \alpha \mathbf{S}^{-1} \tag{14}$$

with

$$\alpha = \frac{N-p-2}{N-1}. \tag{15}$$

This is motivated by the fact that $\mathbf{S}^{-1}$ follows an inverse Wishart distribution, which has a biased expectation value $\langle \mathbf{S}^{-1} \rangle = \alpha^{-1}\mathbf{\Sigma}^{-1}$ as shown in Anderson (2003). Here, the angular brackets denote averaging over the inverse Wishart distribution.

Hartlap et al. (2007) argue that this debiased inverse covariance matrix will remove all biases from parameter inference. However, the situation is more complex. In a Bayesian analysis one would not necessarily define an estimator $\hat{\boldsymbol{\theta}}$, but if one does, the bias is $b_\theta = \langle \hat{\boldsymbol{\theta}} \rangle - \boldsymbol{\theta}$, where the angular brackets now denote the average over the likelihood of the parameters. Adopting the wrong sampling distribution will yield incorrect posterior distributions, with biased parameter estimates (should they be made) and incorrect errors, even if the inverse covariance matrix itself has been debiased.

We compare univariate examples of the likelihoods and the modified $t$-distribution equation (12) in Fig. 1: the Hartlap-scaled and the unscaled Gaussian only differ in width, whereas the $t$-distribution has a more sharply peaked central region but broader extreme wings than a Gaussian, allowing for more scatter away from the peak.

Additionally, the scaling in equation (14) implies a sharp mapping between the estimator $\mathbf{S}^{-1}$ and $\mathbf{\Sigma}^{-1}$, which does not account for the randomness of $\mathbf{S}^{-1}$, due to the finite width of the inverse Wishart distribution. Therefore, $\alpha\mathbf{S}^{-1}$ applied to a *single* given $\mathbf{S}^{-1}$ should not be interpreted as a reliable 'debiasing' but rather a scaling that widens up the Gaussian likelihood equation (1) in an essentially random way. This randomness will propagate through the parameter inference and introduce a scatter of the likelihood contours of which we show a simple example in Fig. 2. This scatter can only be reduced by estimating the inverse covariance matrix more precisely, see also (Dodelson & Schneider 2013; Taylor et al. 2013).

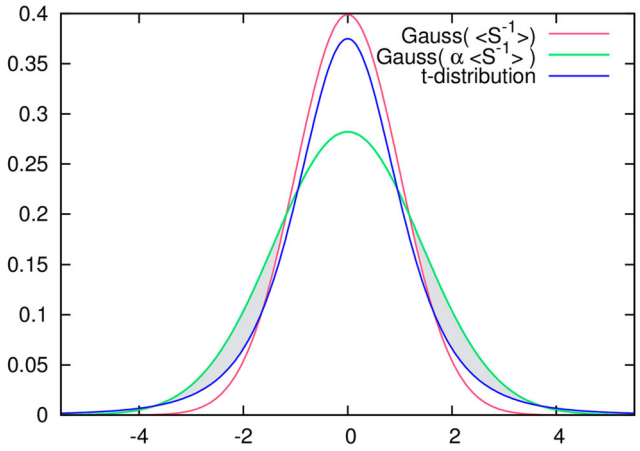


**Figure 1.** Comparison of the two Gaussian likelihoods and the $t$-distribution for a *particular* estimated $\mathbf{S}$, using $N = 5, p = 1, \alpha = 0.5$ which are examples. The grey shaded areas indicate the heavy and short wings of the Hartlap-scaled likelihood.

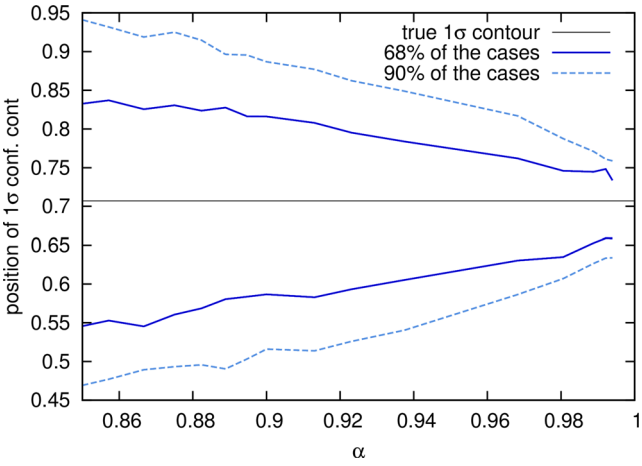


**Figure 2.** The $1\sigma$-confidence contour of a one-dimensional normal distribution lies at $1/\sqrt{2} \approx 0.707$. However, if the covariance is estimated from simulations, its random scatter will make the estimated likelihood randomly too narrow or too broad. In 68 per cent (90 per cent) of the estimated covariances, then deduced $1\sigma$-contour falls into the area bordered by the dark blue (dashed blue) lines. The number of simulations increases with $\alpha$ from equation (15).

## 4 COMPARISON OF THE DISTRIBUTIONS

### 4.1 Illustrative univariate example

We illustrate with a univariate frequentist example that the Hartlap-scaled Gaussian introduces errors into the parameter inference, whereas equation (12) does not. We choose a *true* mean $\mu_t$, which we want to estimate in the following. We then produce 10 000 Gaussian data sets with this mean, and produce 150 estimates of the covariance matrix from $N$ further samples (where $N$ determines $\alpha$). For each data set and each covariance matrix, we then calculate the Hartlap-scaled likelihood and the modified $t$-distribution. Both the Hartlap-scaled Gaussian and the $t$-distribution of $\mu_t$ make quantitative predictions such as stating that $\mu_t$ will fall 5 per cent of the time into the lower 5 per cent tail of the likelihood, or 68 per cent of the time into the 68 per cent likelihood contour, given some data sets. But since the two likelihoods differ in shape, their lower-tail probabilities and likelihood contours will also differ, and only one will

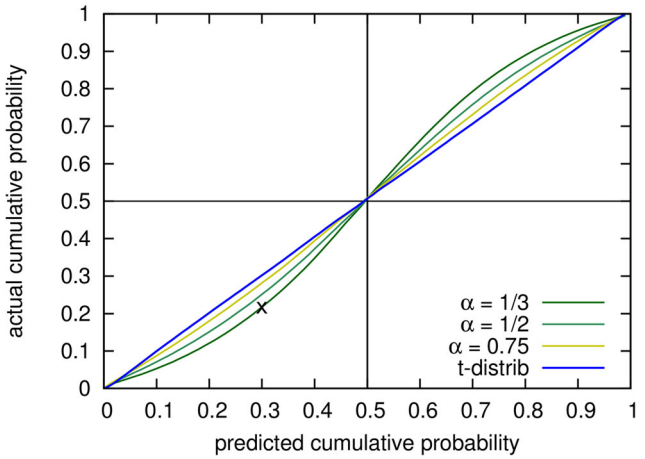


**Figure 3.** Predicted versus true cumulative probability for an illustrative univariate estimation of a mean. The $t$-distribution follows the diagonal line of unit slope, meaning it predicts correctly the shape of the likelihood, whereas the Hartlap-scaled Gaussian is too broad. For example, the marked point is the lower 30 per cent-tail of the Hartlap-scaled Gaussian – but in reality the true mean falls into this tail only with a probability of 0.2.

make the correct quantitative predictions. Since we know the true mean, we can test this. Likelihood contours and tail-probabilities can be converted into each other, so it is sufficient to test only one. We choose the lower $x$-per cent tail probability, i.e. the cumulative probability function and check whether the $x$-per cent cumulative distribution does indeed cover the true mean $x$-per cent of the times. In Fig. 3, we find that only the $t$-distribution correctly reproduces the cumulative distribution – the line is straight with a slope of unity. The Hartlap-scaled Gaussian does not capture the scatter around the peak correctly, which will lead to a mis-estimate of the parameter errors, even on average. As expected, the discrepancy decreases as more simulations are included in the estimation of $\mathbf{S}$ (i.e. as $\alpha \to 1$).

### 4.2 Assessment of confidence in higher dimensions

The issue at hand can be studied in higher dimensions by investigating the distribution of the following quantities:

$$\chi^2 = (\boldsymbol{X}_\mathrm{o} - \boldsymbol{\mu})^T\, \Sigma^{-1}\, (\boldsymbol{X}_\mathrm{o} - \boldsymbol{\mu}) \tag{16}$$

which is the true $\chi^2$; the same quantity but with the estimated $\mathbf{S}$ replacing $\boldsymbol{\Sigma}$,

$$T^2 = (\boldsymbol{X}_\mathrm{o} - \boldsymbol{\mu})^T\, \mathbf{S}^{-1}\, (\boldsymbol{X}_\mathrm{o} - \boldsymbol{\mu})\,; \tag{17}$$

and the Hartlap-scaled version

$$H^2 = (\boldsymbol{X}_\mathrm{o} - \boldsymbol{\mu})^T\, \alpha\mathbf{S}^{-1}\, (\boldsymbol{X}_\mathrm{o} - \boldsymbol{\mu})\,. \tag{18}$$

By construction, we have $\langle H^2 \rangle = \langle \chi^2 \rangle$, meaning the Hartlap-scaling does indeed debias the expectation value. It does however underestimate statistical scatter, as we shall show in the following.

$\chi^2$ follows the $\chi^2_p$-distribution, which only arises if the covariance is precisely known and indeed the correct covariance of $\boldsymbol{X}_\mathrm{o}$. The quantity $T^2$ will not follow the $\chi^2_p$-distribution, because it contains not only a random vector $\boldsymbol{X}_\mathrm{o} \sim \mathcal{N}_p(\boldsymbol{\mu}, \Sigma)$, but additionally the random estimate of the covariance matrix that follows the Wishart distribution $\mathcal{W}(\Sigma/n, n)$. $T^2$ therefore follows:

$$\frac{T^2(n-p+1)}{pn} \sim F_{p,n-p+1}, \tag{19}$$

where $n = N - 1$, and the $F_{p,\,n-p+1}$ is the $F$-distribution of $p$ and $n - p + 1$ degrees of freedom (Anderson 2003). Consequently, a change of variables shows that,

$$T^2 \sim \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma(p/2)\Gamma(n-p+1/2)} \frac{n^{-p/2}(T^2)^{p/2-1}}{(T^2/n+1)^{\frac{n+1}{2}}} \tag{20}$$

instead of $T^2 \sim \chi^2_p$, see Fig. 4. Only for $N \to \infty$ will the Wishart distribution tend towards a delta-function, and the distribution of $T^2$ will then tend towards a $\chi^2_p$-distribution.

The distribution of the Hartlap-scaled $H^2$ is more sharply peaked than that of $\chi^2$, thereby suggesting that the experiment has less statistical scatter than the $\chi^2_p$ distribution on average. This is impossible since the $\chi^2_p$ distribution is subject to scatter of the random vector $\boldsymbol{X}_\mathrm{o}$ only.

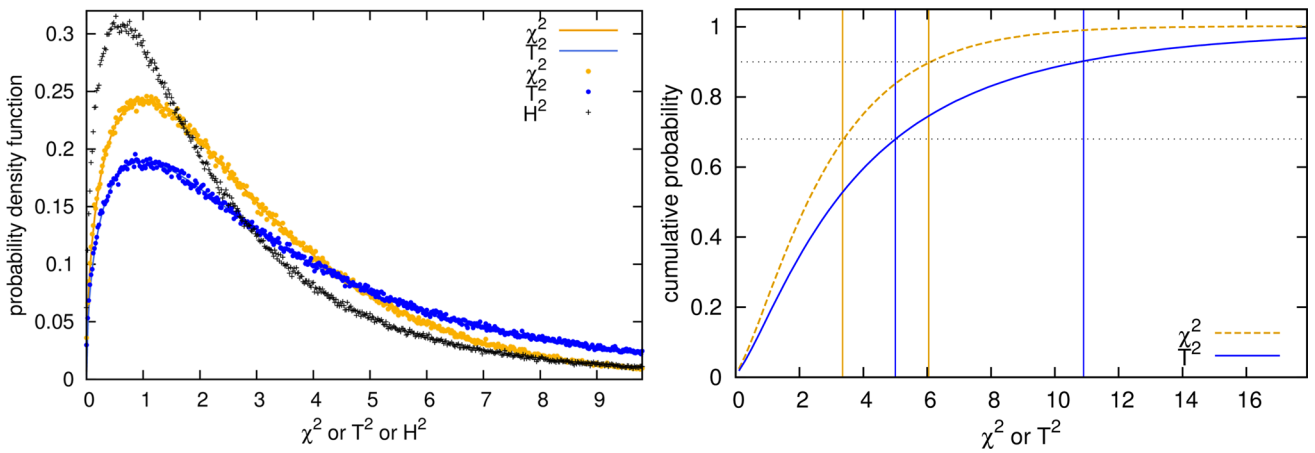


**Figure 4.** Left: the distribution of different interpretations for $(\boldsymbol{X}_\mathrm{o} - \boldsymbol{\mu})^\mathrm{T}\mathbf{S}^{-1}(\boldsymbol{X}_\mathrm{o} - \boldsymbol{\mu})$, using $p = 3$, $N = 10$. Dots represent simulations, solid lines are the analytical $\chi^2_p$- and $T^2$-distribution. For $N \gg p$, the $T^2$-distribution approximates the $\chi^2_p$-distribution. The closer $N$ is to $p$, the more differs the $T^2$-distribution from the $\chi^2_p$-distribution, being broader than $\chi^2_p$, leading to a cumulative distribution that rises more slowly. The Hartlap-scaled $H^2$ follows the black distribution which is more sharply peaked than the $\chi^2_p$, although the $\chi^2_p$-distribution is the minimal scatter that one can achieve; this means the Hartlap-scaled $H^2$ underestimates the joint scatter of $\boldsymbol{X}_\mathrm{o}$ and $\mathbf{S}^{-1}$. Right: the cumulative distributions of $\chi^2$ and $T^2$ from the left. The vertical lines mark the 68 per cent and 90 per cent confidence limits.

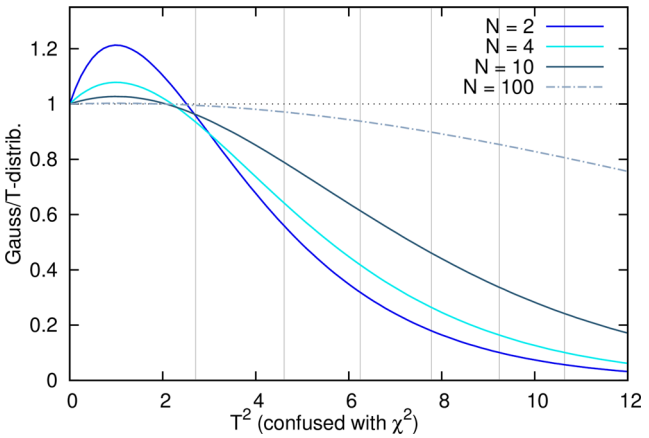


**Figure 5.** Unnormalized weights $G(\boldsymbol{X}_\mathrm{o}, \boldsymbol{\mu}, \mathbf{S}^{-1})/P(\boldsymbol{X}_\mathrm{o}, \boldsymbol{\mu}, \mathbf{S}^{-1}, n)$ for mapping between a Gaussian likelihood and a $t$-distribution. The normalization depends on the dimensionality of the data set, and leads to an offset along the $y$-axis, that is however independent of theoretical parameters. The number of simulations in the covariance matrix is $N$. The vertical lines depict the $\chi^2$ values (2.71, 4.61, 6.25, 7.78, 9.24 and 10.64) that enclose 90 per cent confidence for a multivariate Gaussian.

The cumulative probabilities $P_\mathrm{c}(\chi^2_\mathrm{c})$ or $P_\mathrm{c}(T^2_\mathrm{c})$ give our confidence that the mean $\boldsymbol{\mu}$ of the multivariate vector $\boldsymbol{X}_\mathrm{o}$ is enclosed within an ellipsoid bounded by $\chi^2_\mathrm{c}$ or $T^2_\mathrm{c}$. The more slowly rising cumulative distribution function of $T^2$, therefore shows that we need $T^2 > \chi^2$ in order to achieve the same confidence that the mean is captured within the confidence contours. In parameter space, this will lead to an increase of the Bayesian confidence intervals.

### 4.3 Reweighting an MCMC chain that sampled from a Gaussian likelihood

We have shown above that $T^2$, $\chi^2$ and $H^2$ follow different distributions, which will affect parameter inference. Often, the error of confusing a $T^2$ with a $\chi^2$ or $H^2$ can retrospectively be undone with very little numerical effort by reweighting an existing Monte Carlo Markov Chain (MCMC) chain.

In Fig. 5, we plot weights for reweighting a chain that sampled from $\exp(-\chi^2/2)$. If a Hartlap-scaling has been applied, it would additionally need to be removed.

We note that the maximum of the $t$-distribution in the full parameter space coincides with the maximum of $\chi^2$ (and also of $H^2$), but once any parameters are marginalized over, the resulting parameter posteriors will not in general peak in the same place.

## 5 CONCLUSIONS

We have studied how statistical uncertainties in an estimated covariance matrix affect parameter inference. We summarize our findings as follows.

For data $\boldsymbol{X}_\mathrm{o}$ drawn from a multivariate Gaussian, the likelihood will be Gaussian if the data covariance $\boldsymbol{\Sigma}$ is exactly known. If however the covariance is estimated from $N$ simulations, $\mathbf{S} = 1/(N-1)\sum_{i=0}^{N}(\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})^T$, the estimator $\mathbf{S}$ is unbiased, but $\mathbf{S}^{-1}$ is not an unbiased estimator of $\boldsymbol{\Sigma}^{-1}$. An unbiased estimator is $\alpha\langle\mathbf{S}^{-1}\rangle = \boldsymbol{\Sigma}^{-1}$ where $\alpha = (N-p-2)/(N-1)$ (Anderson 2003). An earlier proposal, by Hartlap et al. (2007), uses the unbiased estimate $\alpha\mathbf{S}^{-1}$ of the inverse covariance matrix, but keeping a Gaussian likelihood. The statistical scatter of the estimator $\mathbf{S}^{-1}$ is not fully accounted for, and this yields posteriors that are on average simultaneously too broad in their centres, yet not broad enough in the extremes.

The principled approach is to recognize that we have a *sample* of the covariance matrix $\mathbf{S}$, and compute the likelihood by marginalizing over the inverse-Wishart distribution of the true covariance matrix $\boldsymbol{\Sigma}$, conditioned on $\mathbf{S}$. This gives a modified multivariate $t$-distribution $P(\boldsymbol{X}_\mathrm{o}|\boldsymbol{\mu}, \mathbf{S}, N)$, given by equation (12). This is what we require for parameter inference and is the main result of this Letter.

For parameter inference in the presence of a covariance matrix estimated from a finite number of simulations, our results imply that MCMC chains should evaluate the modified $t$-distribution equation (12) at each sample point, instead of a Gaussian distribution. The numerical complexity will not be increased by this. It stays constant since both distributions must evaluate the quantity $(\boldsymbol{X}_\mathrm{o} - \boldsymbol{\mu})^T\mathbf{S}^{-1}(\boldsymbol{X}_\mathrm{o} - \boldsymbol{\mu})$. Consequently, a reweighting of existing MCMC chains is possible without much effort if the chains record $(\boldsymbol{X}_\mathrm{o} - \boldsymbol{\mu})^T\mathbf{S}^{-1}(\boldsymbol{X}_\mathrm{o} - \boldsymbol{\mu})$.

## ACKNOWLEDGEMENTS

ES acknowledges financial support from the RTG *Particle Physics beyond the Standard Model*, through the DFG fund 1940 and the transregional collaborative research centre TR 33 '*The Dark Universe*' of the DFG. We thank Andrew Jaffe, Justin Alsing and Ewan Cameron for useful discussions and comments.

## REFERENCES

Anderson T. W., 2003, An Introduction to Multivariate Statistical Analysis, 3rd edn. Wiley, New York
Budavári T. et al., 2003, ApJ, 595, 59
Dodelson S., Schneider M. D., 2013, Phys. Rev. D, 88, 063537
Hamimeche S., Lewis A., 2009, Phys. Rev. D, 79, 083012
Hartlap J., Simon P., Schneider P., 2007, A&A, 464, 399
Heymans C. et al., 2013, MNRAS, 432, 2433
Jeffreys H., 1961, Theory of Probability. Oxford Univ. Press, Oxford
Mardia K. V., Kent J. T., Bibby J. M., 1979, Multivariate Analysis. Probability and Mathematical Statistics. Academic Press, London
Semboloni E. et al., 2006, A&A, 452, 51
Sun D., Berger J., 2006, in Bernardo J. M. et al., eds, Proc. Valencia/ISBA 8th World Meeting on Bayesian Statistics. Oxford Univ. Press, Oxford
Taylor A., Joachimi B., Kitching T., 2013, MNRAS, 432, 1928
Wishart J., 1928, Biometrika, 20, 32

This paper has been typeset from a TEX/LATEX file prepared by the author.