Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **459,** 1916–1928 (2016)
Advance Access publication 2016 April 3

doi:10.1093/mnras/stw738

Gaussianization for fast and accurate inference from cosmological data

Robert L. Schuhmann,* Benjamin Joachimi and Hiranya V. Peiris
Department of Physics and Astronomy, University College London, Gower Place, London WCE1 6BT, UK

Accepted 2016 March 28. Received 2016 March 4; in original form 2015 August 27

ABSTRACT

We present a method to transform multivariate unimodal non-Gaussian posterior probability densities into approximately Gaussian ones via non-linear mappings, such as Box–Cox transformations and generalizations thereof. This permits an analytical reconstruction of the posterior from a point sample, like a Markov chain, and simplifies the subsequent joint analysis with other experiments. This way, a multivariate posterior density can be reported efficiently, by compressing the information contained in Markov Chain Monte Carlo samples. Further, the model evidence integral (i.e. the marginal likelihood) can be computed analytically. This method is analogous to the search for normal parameters in the cosmic microwave background, but is more general. The search for the optimally Gaussianizing transformation is performed computationally through a maximum-likelihood formalism; its quality can be judged by how well the credible regions of the posterior are reproduced. We demonstrate that our method outperforms kernel density estimates in this objective. Further, we select marginal posterior samples from *Planck* data with several distinct strongly non-Gaussian features, and verify the reproduction of the marginal contours. To demonstrate evidence computation, we Gaussianize the joint distribution of data from weak lensing and baryon acoustic oscillations, for different cosmological models, and find a preference for flat Λcold dark matter. Comparing to values computed with the Savage–Dickey density ratio, and Population Monte Carlo, we find good agreement of our method within the spread of the other two.

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

1 INTRODUCTION

According to the Bayesian paradigm, inference on any data set will yield a posterior probability distribution on the space of model parameters (MPs). This density function represents, in its entirety, the full knowledge gained in the attempt to infer the underlying parameters. Such distributions often depart significantly from a Gaussian form. This led to the widespread use of Monte Carlo sampling methods to report the typically non-Gaussian posterior constraints obtained from experiments, such as *Planck*[1]. Reconstructing the posterior density from such a Markov Chain Monte Carlo (MCMC) sample, e.g. to visualize the multivariate parameter constraints, or to combine the constraints from multiple data sets, can be non-trivial due to the large sample size necessary to appropriately map the distribution; in addition, the contours often need further smoothing for stylistic reasons.

Instead, we propose to redefine the underlying MPs, so that the new posterior density approximately takes a Gaussian shape after the transformation from old to new parameters; this presupposes that we begin with a unimodal posterior density. Such a transformation would allow for enormous data compression: instead of a full MCMC sample from the posterior distribution, we only need to report the Gaussianizing transformation, and the first and second moments of the resulting Gaussian distribution. From these alone, we can reconstruct an analytic expression for the full non-Gaussian posterior density, and subsequently combine it with other datax sets.

Further, it becomes possible to display and compare non-ellipsoidally shaped contours of non-Gaussian parameter constraints – whether joint or marginalized – without any smoothing. Thus, this method allows for summarizing posterior densities in a versatile and efficient way, which faithfully reproduces the information contained in the full probability density.

The idea of transforming a function to a Gaussian shape is, in principle, not limited to reproducing probability densities. As the integral over a Gaussian can be performed analytically, this opens up a strategy to feasibly compute high-dimensional integrals, such as the model evidence (i.e. the marginal likelihood).

* E-mail: robert.schuhmann.13@ucl.ac.uk

[1] See Planck Collaboration XVI (2014) and Planck Collaboration XIII (2015). For the Markov chains see http://www.cosmos.esa.int/web/planck/pla; consult http://lambda.gsfc.nasa.gov for an eclectic list of data combinations in various cosmological models, compiled by NASA's High Energy Astrophysics Science Archive Research Center (HEASARC).

The transformed MPs are analogous to the normal parameters of the cosmic microwave background (CMB): it has been highly advantageous for rapid likelihood calculation (such as CMBfit, CMBwarp, and Parameters for the Impatient COsmologist, see Kosowsky, Milosavljevic & Jimenez 2002; Chu, Kaplinghat & Knox 2003; Jimenez, Verde & Peiris 2004; Sandvik et al. 2004; Fendt & Wandelt 2007), to redefine the cosmological MP such that the model is approximately linear in these newly defined *normal* parameters. Thus, the likelihood approximately takes the form of a multivariate Gaussian density. For most observables, we would be at a loss to search for a linearizing redefinition of the MP space directly motivated by the structure of the model itself. Instead, is it possible to computationally find suitable parameters, i.e. a suitable bijective transformation which approximately Gaussianizes the posterior in question?

Extending the work of Joachimi & Taylor (2011), we present an algorithm to find and test such a non-linear Gaussianizing transformation from a Markov chain sampling the posterior distribution of the original parameters. In principle, this distribution could stem from any experiment or data type. In Section 2, we describe the details of the algorithm, verification of the reconstructed posterior distribution, and the specific transformations employed. Following an illustration of these on a toy example in Section 3, Section 4 demonstrates the performance of our implementation, using Markov chains from the *Planck* satellite constraints on cosmological models (Planck Collaboration XVI 2014; Planck Collaboration XIII 2015). In Section 5, we present an efficient way to calculate the model evidence, a quantity needed to judge the predictivity of different MP spaces, via Gaussianizing transformations. Section 6 offers conclusions and suggests future directions.

2 GAUSSIANIZATION

To find the right multivariate transformation, we will at first adopt the strategy of redefining each MP separately, i.e. the first new MP will only depend on the first old MP, etc. In Section 2.4, we will drop this assumption and consider transformations which can correlate the MPs.

The set of all multivariate Gaussianization transformations, from which we are to pick the optimal one, will be constructed in the following way: assume a family of bijective real-valued functions $F_\Delta : \mathbf{R} \to \mathbf{R}$ indexed by n real transformation parameters (TPs) $\Delta = (\delta^1, \ldots, \delta^n)$. Given the d-dimensional vector of MPs $\boldsymbol{X} = (X_1, \ldots, X_d)$, we transform to the new (Gaussian-distributed) parameters $\boldsymbol{Y}$ via

$$\boldsymbol{Y} = (Y_1, \ldots Y_d) = \left[F_{\Delta_1}(X_1), \ldots, F_{\Delta_d}(X_d)\right], \tag{1}$$

where the full multivariate transformation is now specified by all d TP n-tuples $(\Delta_1, \ldots, \Delta_d)$, i.e. one $\Delta_i = (\delta_i^1, \ldots, \delta_i^n)$ for each MP. To avoid confusion, we shall from now on distinguish between MPs, which the posterior probability density depends on, and TPs, which specify one Gaussianizing transformation. The algorithm can be applied to arbitrary parametrized transformation families, suitable for various forms of non-Gaussianity – in principle, we could even choose different transformations for each MP, instead of using the same shape $F_{\Delta_i}(X_i)$ for all of them.

Assuming such a bijective transformation $\boldsymbol{X} \mapsto \boldsymbol{Y}$, we immediately have an analytic form for the posterior density

$$\begin{aligned}\Pi(\boldsymbol{X}) &= \left|\frac{\mathrm{d}\boldsymbol{Y}}{\mathrm{d}\boldsymbol{X}}\right| \tilde{\Pi}(\boldsymbol{Y}) \\ &= \left(\prod_{i=1}^{d}\left|\frac{\mathrm{d}F_{\Delta_i}}{\mathrm{d}X}(X_i)\right|\right)\frac{1}{\sqrt{(2\pi)^d \det\tilde{\boldsymbol{\Sigma}}}} \\ &\quad\times\exp\left\{-\frac{1}{2}\left[\boldsymbol{Y}(\boldsymbol{X})-\tilde{\boldsymbol{\mu}}\right]^T\tilde{\boldsymbol{\Sigma}}^{-1}\left[\boldsymbol{Y}(\boldsymbol{X})-\tilde{\boldsymbol{\mu}}\right]\right\}.\end{aligned} \tag{2}$$

One still needs to find the mean vector $\tilde{\boldsymbol{\mu}}$ and the covariance matrix $\tilde{\boldsymbol{\Sigma}}$ of the transformed posterior density $\tilde{\Pi}$. These are estimated from the transformed sample (see Section 2.1).

2.1 Finding the optimal transformation

Given a weighted point sample $\mathcal{D} = \{(\boldsymbol{X}^a, w^a)\}_{a=1}^{\mathcal{N}}$, containing $\mathcal{N}$ points in $\mathbf{R}^d$ and probability weights w^a, which has been sampled from the posterior distribution in question, we wish to quantify the Gaussianization properties of different transformations applied to this sample. To this end, we follow (Box & Cox 1964, see also Velilla 1993 and Joachimi & Taylor 2011) in maximizing the profile likelihood over TP space, i.e. depending only on the $n \times d$ real TPs contained in $\boldsymbol{\Delta} = (\Delta_1, \ldots, \Delta_d)$. This likelihood is a function of the TPs Δ, quantifying how well each transformation Gaussianizes the distribution of data set $\mathcal{D}$; however, it does not pertain to the posterior density in equation (2), which is a function of the MPs $\boldsymbol{X}$.

For the Gaussian parameters $\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}$ in equation (2), we insert their standard debiased weighted maximum-likelihood estimators, which depend on the transformed sample $\{(\boldsymbol{Y}^a, w^a)\}_{a=1}^{\mathcal{N}}$

$$\tilde{\boldsymbol{\mu}} = \frac{1}{W_1}\sum_{a=1}^{\mathcal{N}} w^a \boldsymbol{Y}^a; \tag{3}$$

$$\tilde{\boldsymbol{\Sigma}} = \frac{W_1}{(W_1)^2 - W_2}\sum_{a=1}^{\mathcal{N}} w^a(\boldsymbol{Y}^a - \tilde{\boldsymbol{\mu}})(\boldsymbol{Y}^a - \tilde{\boldsymbol{\mu}})^T, \tag{4}$$

with $W_1 = \sum w^a$ and $W_2 = \sum (w^a)^2$. These estimators depend on $\boldsymbol{\Delta}$ indirectly, as they are computed after $\mathcal{D}$ has been transformed with $\boldsymbol{\Delta}$. We arrive at the profile weighted log-likelihood

$$\begin{aligned}\mathcal{L}(\boldsymbol{\Delta}|\mathcal{D}) = &-\frac{W_1}{2}\ln\det\tilde{\boldsymbol{\Sigma}}(\boldsymbol{\Delta}, \mathcal{D}) \\ &+\sum_{a=1}^{\mathcal{N}} w^a \sum_{i=1}^{d}\ln\left|\frac{\mathrm{d}F_{\Delta_i}}{\mathrm{d}X}(X^a{}_i)\right|,\end{aligned} \tag{5}$$

where several terms independent of $\boldsymbol{\Delta}$ have been discarded. In general, both the covariance matrix of the transformed sample and the Jacobian term will depend on the TPs $\boldsymbol{\Delta}$ in a non-linear way, hence finding the maximum-likelihood values for the TPs will require numerical optimization. For this purpose, we have employed the Gnu Scientific Library implementation of the well-known Nelder–Mead simplex algorithm (Nelder & Mead 1965).

As already noted by Joachimi & Taylor (2011), log-likelihood degeneracies in TP space are common. These may jeopardize the numerical stability of the calculation of $\mathcal{L}$. There are generic cases where a moderately large value for one TP may already result in unmanageably large numerical values for the transformed sample, such as e.g. the power transformation $X_i \mapsto (X_i)^{\lambda_i}$ with $\lambda_i \sim 50$. Generically, the optimization algorithm tends to slide into these TP

space regions quite easily. Hence, we include a penalty term of the form

$$\mathcal{P}(\mathbf{\Delta}) = \epsilon \sum_{i=1}^{d} \sum_{s=1}^{n} (\delta_i^s - \delta^{s,U})^p, \tag{6}$$

where $\delta^{s,U}$ are the parameter values corresponding to the identity transformation. We minimize the function $-\mathcal{L}(\mathbf{\Delta}; \mathcal{D}) + \mathcal{P}(\mathbf{\Delta})$ over the $n \times d$ real numbers in $\mathbf{\Delta}$. Values of $p = 4$ and $\epsilon = 10^{-4}$ have proven to be highly stabilizing, and at the same time do not distort the shape of the resulting analytic posterior distribution.

In this work, we employ the Nelder–Mead algorithm just for illustrating the method – faster and more reliable algorithms to find the global minimum of the likelihood function exist [such as Bound Optimization BY Quadratic Approximation (BOBYQA); see Powell 2007] and can readily be applied here.

2.2 Box–Cox transformations and their kin

The Box–Cox transformation (Box & Cox 1964) is a generalization of the power map. This transformation family is widely used in statistics and econometrics, e.g. to make data approximately homoscedastic and normal. Our usage is different in that we use it to alter the distribution of MPs, rather than the distribution of data. Including a shift parameter a, the 1D version is defined as

$$x \mapsto BC_{(a,\lambda)}(x) = \begin{cases} \lambda^{-1}[(x+a)^\lambda - 1] & (\lambda \neq 0) \\ \log(x+a) & (\lambda = 0) \end{cases} \tag{7}$$

for a single MP x, i.e. $(\delta^1, \delta^2) = (a, \lambda)$. Note that the family is continuous at $\lambda = 0$ and that the mapping requires $a < x$. Typically, an MP with a skewed distribution can be transformed to an MP with symmetric, Gaussian distribution upon the appropriate choice of the power TP λ, e.g. a log-normal distribution can be analytically transformed to a Gaussian with $a = \lambda = 0$. The identity transformation corresponds to $\delta^{1,U} = a = 1$ and $\delta^{2,U} = \lambda = 1$. Inserting this transformation family into equation (2), we recover the formula given in Joachimi & Taylor (2011).

As an extension of the Box–Cox family, we propose the Arcsinh–Box–Cox transformation ('ABC transformation' hereafter):

$$x \mapsto \mathrm{ABC}_{(a,\lambda,t)}(x) = \begin{cases} t^{-1} \sinh[t\, BC_{(a,\lambda)}(x)] & (t > 0) \\ BC_{(a,\lambda)}(x) & (t = 0) \\ t^{-1} \operatorname{arcsinh}[t\, BC_{(a,\lambda)}(x)] & (t < 0). \end{cases} \tag{8}$$

The inclusion of the TP t will prove particularly useful to remove residual kurtosis from an MP distribution. The identity transformation reads $\delta^{1,U} = a = 1$, $\delta^{2,U} = \lambda = 1$, $\delta^{3,U} = t = 0$.

The Box–Cox family does not form a group, because two subsequent transformations cannot be expressed as another Box–Cox transformation; the same holds for the ABC family. This will be of importance for Section 2.4.

Box–Cox transformations demonstrate that the domain of the function F_Δ – in particular its dependence on Δ – requires special attention: for given a, it is defined only for $x \in (-a, \infty)$, the same holds for ABC transformations. Thus, the optimization procedure for the sample $\mathcal{D} = \{\boldsymbol{X}^a\}_{a=1}^{\mathcal{N}}$ requires that a_i, the shift parameter for the MP X_i, is bounded from below, i.e. $a_i > \min_a(-X_i^a)$. Conversely, this means that, once the optimal TPs $\mathbf{\Delta}^{\mathrm{opt}}$ are found and inserted into the analytic expression for the original posterior density, equation (2), it is not defined for every value possible value of the MP $\boldsymbol{X}$, but only for $X_i > a_i^{\mathrm{opt}}$. This also necessitates that the normalization needs to be adjusted, which can be done analytically. However, if the sample is large enough so that the tails of the distributions are properly represented, this truncation of the domain is not problematic.

2.3 Verifying the optimal transformation

Once the optimal transformation within its family is found, how do we judge the effectiveness of the resulting Gaussianization? We adopt the following pragmatic standpoint: if the analytic posterior manages to reproduce the 1D and 2D marginalized contours of the sample, it is deemed acceptable. To this end, we propose the test via a cross-contour (CC) plot. The idea is to characterize a probability density by the location of its contours – the surfaces of constant density – and the probability mass stored inside, i.e. the integral of the density over the interior of a contour. If two densities $p(\boldsymbol{X})$ and $q(\boldsymbol{X})$ are identical, then they will store the same mass in any region of the parameter space; if they are different, we expect to find different probabilities for the same regions (e.g. the regions bounded by contours of p). Thus, looking at the family of contour-bounded regions of p, we can ask: does the probability for these, assigned via q, agree with the probability for them assigned via p?

To formalise this, consider the following: given a probability density p in d dimensions, which takes function values between 0 and $p_{\max}$, we define the contour-bounded region assigned to the density value $r \in [0, p_{\max}]$ as

$$\Omega_p(r) = \{\boldsymbol{X} \in \mathbf{R}^d : p(\boldsymbol{X}) \geq r\}.$$

The probability mass enclosed in any of these is

$$\int_{\Omega_p(r)} p(\boldsymbol{X})\, \mathrm{d}\boldsymbol{X} \in [0, 1].$$

Now, assuming we have two probability densities p and q in d dimensions, do the contours of q reproduce those of p? They do in the relevant sense if for every $r \in [0, p_{\max}]$, the q-mass enclosed in the r-contour of p equals the p-mass in this contour, i.e.

$$\int_{\Omega_p(r)} q(\boldsymbol{X})\, \mathrm{d}\boldsymbol{X} = \int_{\Omega_p(r)} p(\boldsymbol{X})\, \mathrm{d}\boldsymbol{X}. \tag{9}$$

It should be noted that this alone is not a sufficient condition for $p \equiv q$, but the counterexamples, which can be constructed mathematically, are non-generic and can be neglected for our purposes.

To detect deviations of the contours of p and q, we could simply plot the left and the right side of equation (9) for a grid of r-values between 0 and $p_{\max}$, and plot the points with respect to the line $y = x$. For concrete problems, it is often more instructive to subtract the right side from the left side, and plot the excess (or deficit) probability mass of q inside the contours of p. If, in this plot, the excess for every contour is consistent with zero, we have succeeded.

In our situation, we compare a point sample $\mathcal{D}$ with a probability density function p – the analytic posterior density as reconstructed via Gaussianization. The right side of equation (9) is the probability mass in the region where the density is greater or equal to r; the left side is the fraction of the point sample which lies in the same region. Therefore, for every value r in the range of p, we find the probability mass enclosed in $\Omega_p(r)$ by gridding $p(\boldsymbol{X})$ over a region containing the sample. Similarly, we count the number of points in $\mathcal{D}$ where the value of p is above r, to compute the fraction of points that lie within $\Omega_p(r)$. This fraction is an estimator of the actual probability mass enclosed, because $\mathcal{D}$ is a discrete sample from the actual posterior distribution. To find the variance of this estimator, we calculate the fraction on 2000 bootstrap realizations of $\mathcal{D}$, and determine the 95 per cent-confidence intervals from these.

If, for every r, the analytic posterior probability mass inside $\Omega_p(r)$ is within this confidence interval for the sample point fraction within $\Omega_p(r)$, we judge our reconstruction attempt to be successful.

It should be noted that poor MCMC sampling of the original target density will yield a poor representation of this density by our reconstructed density (2). Any information about the distribution lost by undersampling cannot be regained. However, as demonstrated in Section 3, our method reproduces less biased contours than other standard methods of density estimation even in regions of low-point density, i.e. where any density estimate must be an extrapolation. Hence, it can be used when the length of the input Markov chains is restricted by computational cost or file size.

2.4 Multipass transformations

If even the optimal Gaussianizing transformation amongst a given family does not bring the posterior density sufficiently close to a Gaussian shape (e.g. as determined via a CC plot), we have two options. We can provide a different family of transformations and redo the optimization; or we can repeat the process on the sample after the first transformation. As already mentioned in Section 2.2, the transformation families employed in this work do not form groups. Hence, two subsequent transformations do not result in another transformation from that family, and transforming twice potentially provides a better Gaussianization than transforming once. In principle, it is possible to apply multiple subsequent transformations, should the quality of the result necessitate it.

In this spirit, we have implemented the following two-pass transformation protocol:

(i) Step 1: Optimize the TPs of the first transformation.
(ii) Step 2: Linear reshaping: centring, rescaling, rotating.
(iii) Step 3: Optimize the TPs of the second transformation.

Strictly speaking, this transformation, whilst being bijective, no longer falls into the class as set up in equation (1), as different MPs are mixed. None the less, equation (2) for the analytic posterior density generalizes in a straightforward way.

In the second step, the sample after the first Gaussianizing transformation is subjected to the following maps (in this order): subtract the sample mean from every parameter, so that the sample is centred on the origin. Then, rescale every parameter such that the standard deviation is unity. Finally, rotate into the eigenbasis of the covariance matrix – this procedure is generally known as principal component analysis (PCA). These reshaping operations not only help to avoid numerical instabilities (centring, rescaling), but also open up new directions for Gaussianization by presenting uncorrelated parameters to the second Gaussianizing transformation, since the transformations defined in equation (1) cannot mix parameters. If two parameters have substantial covariance after step 1, it can be crucial to decorrelate them.

Nevertheless, a price is to be paid for the Gaussianizing power added with Step 2: it sacrifices a decisive property of the simple one-step transformation routine, namely that every transformed MP Y_i only depends on a single untransformed MP X_i. This property allows for easy marginalization of the analytic posterior: to compute this, we can marginalize the Gaussianized sample by dropping all coordinates we wish to marginalize out and determining the mean vector and covariance matrix of the remaining ones. Transforming this marginalized Gaussian density back will then yield the marginalized posterior density on the untransformed MPs. However, with linear reshaping included, this is no longer possible.

Table 1. Optimally Gaussianizing parameters for the distribution in Fig. 1, as found with one-pass Box–Cox transformation.

Parameter	Input value	Recovered value
a_1	2	2.4
λ_1	0.4	0.1
a_2	3	2.6
λ_2	4	2.9

This may be problematic for some applications (such as visualization of 1D or 2D marginal distributions, or creating a CC plot), but not for others – as long as we need only the marginal distribution of a single combination of parameters, we can marginalize by discarding all MP columns of the sample except the ones in question, prior to Gaussianizing.

3 A TOY EXAMPLE

We illustrate these ideas on a 2D example. We draw a sample of 10 000 points from a bivariate Gaussian distribution, and map it through an inverse Box–Cox transformation with known input TP values (see Table 1). All weights are set to unity.

This mock data sample has the advantage that there is at least one Box–Cox transformation which precisely Gaussianizes the underlying probability distribution. Fig. 1 shows the original sample, and the one transformed with the one-pass Box–Cox transformation which was found to be optimally Gaussianizing, i.e. maximizing the log-likelihood in equation (5). As this is a comparably simple problem, we have set the penalty term in equation (6) to zero. The Nelder–Mead algorithm was started 16 times independently with randomized initial conditions. The values of the recovered optimal TPs are shown in Table 1; the standard deviation amongst these 16 values is of order 10^{-7} at worst, so multiple Nelder–Mead runs are not necessary in this low-dimensional example: all of them find the same maximum of the log-likelihood. In high-dimensional cases, however, this strategy can increase the robustness of the procedure. The apparent difference between the parameters of the single inverse Box–Cox transformation and the values found for Box–Cox optimization is due to degeneracies in parameter space. To illustrate these, we show the profile likelihood for (a_1, λ_1) where (a_2, λ_2) are held fixed at their input values, and vice versa, in Fig. 2. The TPs found by the optimization algorithm (black crosses – note that they are projected on to the plane for which the profile likelihood is shown) are degenerate with the input ones (red star). Both Box–Cox transformations map the distribution to sufficiently Gaussian form. We compare our method of reconstructing an analytic posterior density from an MCMC sample with the standard non-parametric method, kernel density estimation (KDE), which also aims to find a functional form for the probability density. The $1-3\sigma$ contours of the posterior density from Gaussianization are shown jointly with those from KDE: these employ a Gaussian kernel, whose covariance matrix is estimated from the sample, and Silverman's rule (Silverman 1998) has been used to determine the bandwidth parameter. No additional smoothing has been applied in Fig. 3, top panel. The bottom panel shows the excess CC probability masses between analytic posterior and sample, and KDE and sample, respectively, as detailed in Section 2.3. Whereas the Box–Cox posterior is consistent with the sample distribution for every single contour, the KDE contours show a strong bias – the contours are wider than they should be. Given that the precision of the contour

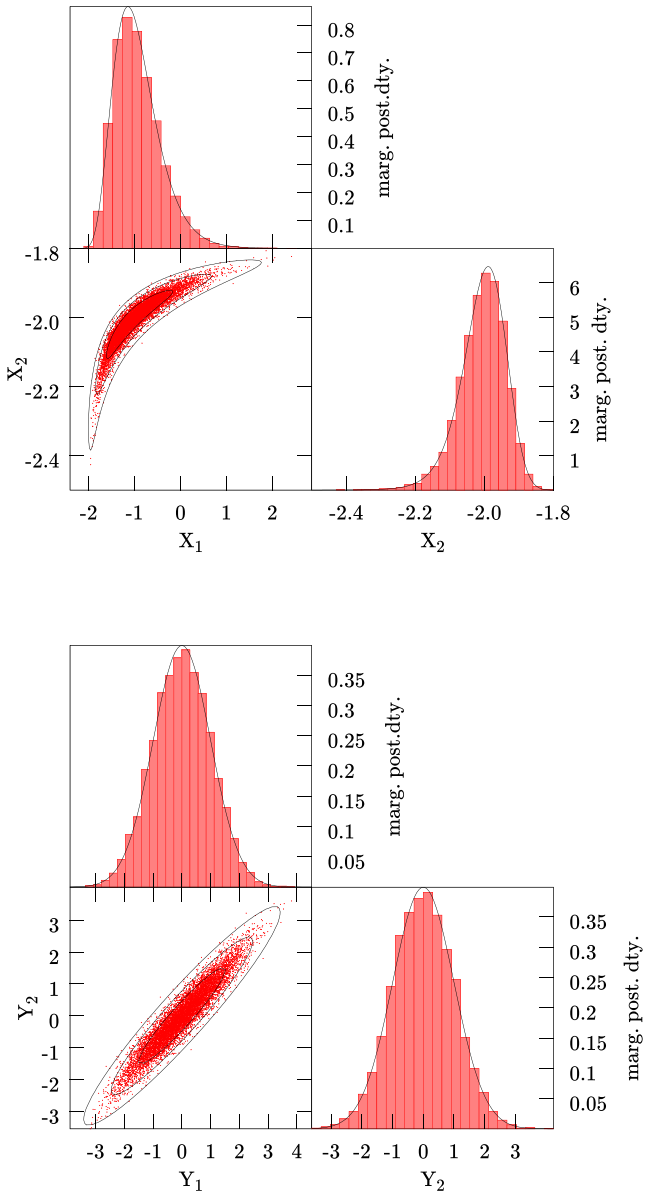

Figure 1. Bivariate sample before (top) and after Gaussianization (bottom). We show the 2D sample and its 1D marginals (red), and compare to the reconstructed analytic posterior density (black): the full 2D contours, and its 1D marginal distributions.

reconstruction is, for the Box–Cox method, limited only by the finite size of the sample, it has the potential to perform better than the (biased) kernel density method. Additionally, for applications in which frequent calls of the posterior density are a bottleneck for computation speed, our method of density reconstruction can be advantageous: the additional initial cost for finding the TPs can be outweighed by the subsequent evaluation speedup.

4 PERFORMANCE RESULTS: *Planck* DATA

To demonstrate how the algorithm works on real data, we have employed MCMC samples from the first data release of the *Planck* mission (see Planck Collaboration XVI 2014). This satellite has measured the temperature and polarization anisotropies in the CMB, whose power spectra are sensitive measures of the underlying cosmology. The *Planck* Collaboration has published several data products[2], including MCMC samples from the posterior probability densities of various cosmological models, generated with CosmoMC (see Lewis & Bridle 2002, also: http://cosmologist.info/cosmomc).

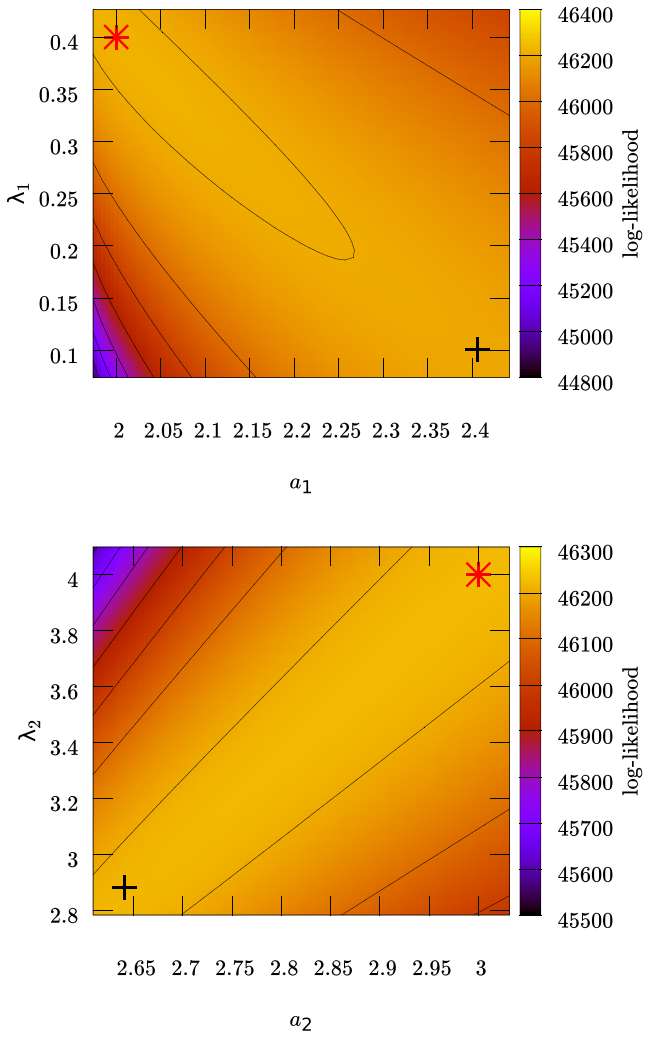

Figure 2. Profile log-likelihood for the TPs relating to X_1 (top), and to X_2 (bottom). The red star shows the input values, the black cross the recovered values, as projected on to the plane. The degeneracies between different TPs are apparent.

The baseline cosmology is the standard model of a flat Universe with cold dark matter (CDM) and a cosmological constant, commonly known as ΛCDM. It contains six parameters: $\Omega_b h^2$ (today's baryon density), $\Omega_c h^2$ (today's CDM density), $100\,\theta_{MC}$ (scaled sound horizon), τ (reionization optical depth), n_s (spectral index of primordial scalar perturbations), and $\ln(10^{10} A_s)$ (log power amplitude of primordial scalar perturbations). Several extensions of this baseline model are also listed, including those by adding either of the following parameters: Ω_K (curvature parameter), w (dark energy equation of state), r (primordial tensor-to-scalar amplitude ratio), and N_{eff} (effective number of relativistic degrees of freedom). Further, these chains list derived quantities, e.g. today's Hubble parameter H_0, the age of the Universe, and a variety of foreground modelling parameters, such as A_ν^{PS} and A_ν^{CIB}, modelling the

[2] See http://www.cosmos.esa.int/web/planck/pla.

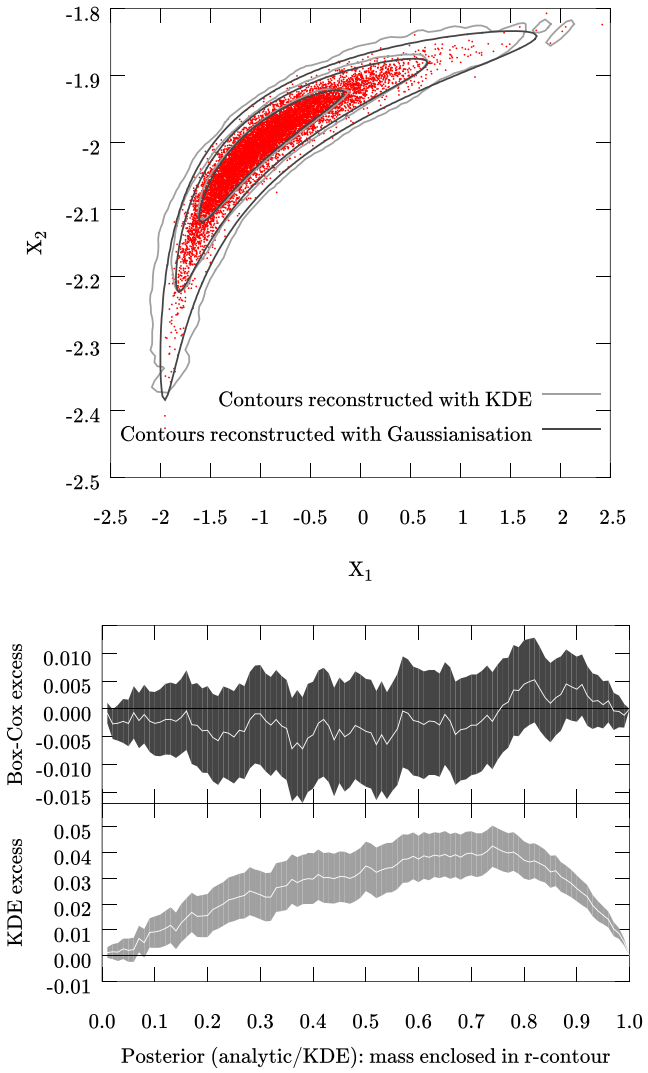

Figure 3. Comparison of the analytic posterior density, as found via Gaussianization (dark grey), to KDE (light grey). The top panel shows the 2D contours of each density estimation method in relation to the original sample (red dots). The bottom panel (CC plots; top: Box–Cox, bottom: KDE) compares the respective contours of each function to the original sample: We determine the fraction of the point sample located inside one probability contour, and plot the excess of this fraction over the probability density mass for that same contour. The band shows the 95 per cent variance in the point fraction due to sampling.

amplitudes of Poisson point sources and the cosmic infrared background in the frequency bands $\nu = 100$, 143, and 217 GHz. These are of particular interest to us, as the most prominent non-Gaussian features of the posterior densities can be seen in them.

The chains, as presented, are not decorrelated, so we thin them by using every 20th sample. We employ the `'..._planck_lowl_...'` chains, which use only the temperature–temperature correlations. The plots in this section are created using the seven-parameter model including Ω_K; the sample contains 11 546 points after thinning.

All these Markov chains assume uniform proper prior densities (i.e. being supported on compact rectangular boxes) and list the log-likelihood for every point (for further details, see Planck Collaboration XVI 2014).

We show several 2D marginalized posterior samples exhibiting different non-Gaussian features, and how well they are reproduced by the analytic posterior (equation 2), such as triangular shapes (see Fig. 4), pronounced non-linear degeneracies (see Fig. 5), and sharp boundaries (walls) arising from MP space boundaries (see Fig. 6).

Fig. 5 demonstrates the usefulness of the intermediate PCA in between the ABC transformations: the first transformation has straightened out the curved shape of the maximum, but the distribution still appears skewed towards the upper-left direction (see top-right panel). This is remedied by reshaping, PCA and another ABC transformation (bottom-right panel) – the second Gaussianizing transformation having only little effect compared to the PCA.

The Gaussianization of the distribution in Fig. 6 (top left) shows how two concatenated transformations can be more powerful than a single one. The once-transformed sample still exhibits negative excess kurtosis, which is removed by the second transformation (bottom-left to bottom-right panel).

Further, we compare the CC plots of this two-pass transformation and the one-pass transformation in Fig. 7, which also shows the resulting contours (left-hand panel). The associated one-pass CC plot (bottom right) shows a significant deficit of point sample mass compared to the analytic posterior mass, for the posterior contours between ∼0.1 and ∼0.3, as well as for ∼0.8, and between ∼0.95 and 1. The latter is visible between the 2σ- and 3σ-contours close to the wall-like constraint at $\ln(10^{10}A_S) \simeq 2.92$. By contrast, the CC plot for the two-pass transformation (Fig. 7, top right) demonstrates good agreement between the contours of analytic posterior and point samples.

To demonstrate the algorithm working on a high-dimensional example, we Gaussianize a 7D *Planck* MCMC sample with an ABC transformation. In order to visualize the result, we show all 1D and 2D marginal distributions of the point sample and the full analytic posterior density (see Fig. 8). We employ one-pass transformations, because, as discussed in Section 2.4, the marginalization of the analytic posterior from 7D down to 2D or 1D would not be possible without explicit integration or sampling, had we chosen to use the two-pass protocol.

5 APPLICATION: FAST EVIDENCE COMPUTATION

To decide which of two models $\mathcal{M}_1$ and $\mathcal{M}_2$, each with their associated MP space, is more predictive on a common set of data $\mathcal{D}$, the essential quantity to compute is the model evidence $E_i = \mathcal{P}(\mathcal{D}|\mathcal{M}_i)$, see Jaynes (2003), MacKay (2003), Kass & Raftery (1995), and Skilling (2006). The ratio of the evidences (called Bayes factor) is then used for updating the prior model odds $\mathcal{P}(\mathcal{M}_1) : \mathcal{P}(\mathcal{M}_2)$ to posterior model odds $\mathcal{P}(\mathcal{M}_1|\mathcal{D}) : \mathcal{P}(\mathcal{M}_2|\mathcal{D})$ in the Bayesian sense:

$$\frac{\mathcal{P}(\mathcal{M}_1|\mathcal{D})}{\mathcal{P}(\mathcal{M}_2|\mathcal{D})} = \frac{\mathcal{P}(\mathcal{D}|\mathcal{M}_1)}{\mathcal{P}(\mathcal{D}|\mathcal{M}_2)} \frac{\mathcal{P}(\mathcal{M}_1)}{\mathcal{P}(\mathcal{M}_2)}. \tag{10}$$

The evidence itself, for each model, can be computed via

$$E_i = \mathcal{P}(\mathcal{D}|\mathcal{M}_i) = \int \mathrm{d}\boldsymbol{X}\, \mathcal{P}(\mathcal{D}|\boldsymbol{X}, \mathcal{M}_i)\mathcal{P}(\boldsymbol{X}|\mathcal{M}_i) \tag{11}$$

i.e. via integration of the (unnormalized) posterior density $\Pi(\boldsymbol{X}) = \mathcal{P}(\mathcal{D}|\boldsymbol{X}, \mathcal{M}_i)\mathcal{P}(\boldsymbol{X}|\mathcal{M}_i)$ over the respective parameter space of model $\mathcal{M}_i$ – hence the term 'marginal likelihood' for E. If Π takes a form with non-Gaussian features, and if the MP space is high-dimensional, this integral itself is often difficult to calculate. However, with a bijective transformation $T : \boldsymbol{X} \mapsto \boldsymbol{Y}$ that

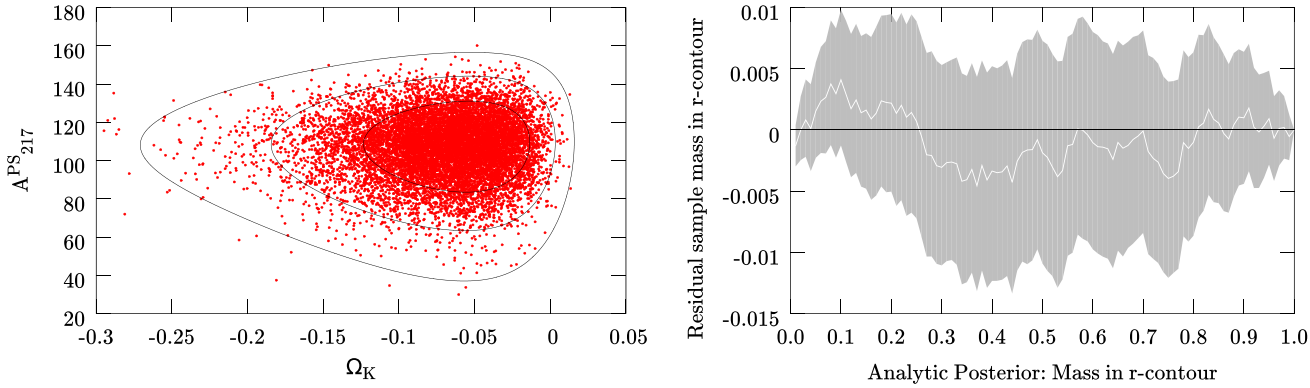

Figure 4. One-pass Gaussianization of a triangular-shaped non-Gaussian feature in a 2D marginal *Planck* posterior via ABC transformation. Left: original sample (red dots) and reconstructed analytic posterior (black contours). Right: the CC plot shows that for every contour of the analytic posterior, the probability mass inside (white line) equals the fraction of the point sample inside, within its 95 per cent-confidence interval (green band).

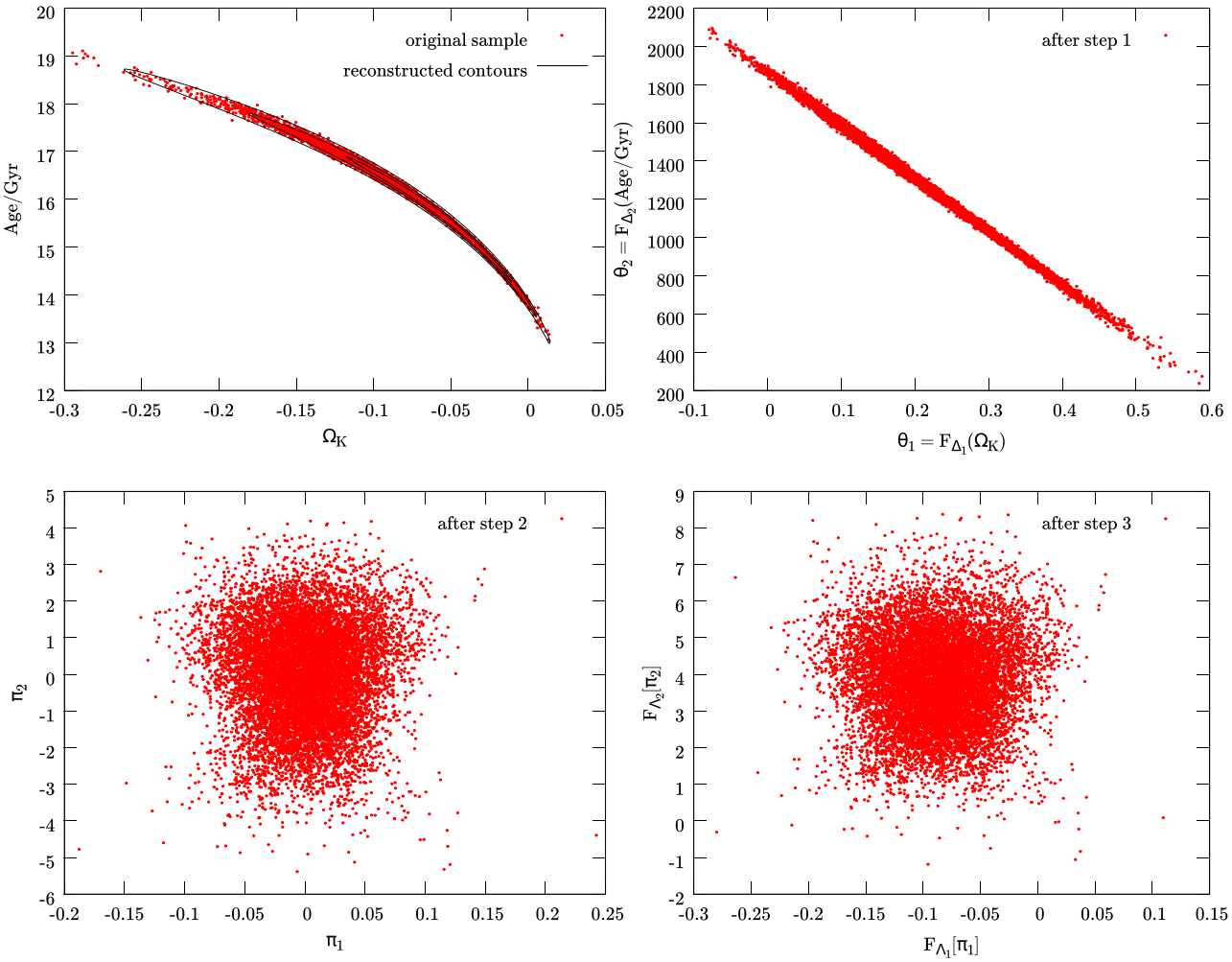

Figure 5. Two-pass Gaussianization of a non-Gaussian MP degeneracy in a 2D marginal *Planck* posterior via ABC transformation, explicitly showing the protocol described in Section 2.4: (θ_1, θ_2) are the parameters after the first transformation; (π_1, π_2) are the coordinates after rotation into the PCA eigenbasis of the centred and rescaled (θ_1, θ_2)-sample, which are finally transformed again. (Δ_1, Δ_2) designate the TPs of the first, (Λ_1, Λ_2) those of the second ABC transformation. Note, how crucial the intermediate PCA step is to achieve Gaussianity.

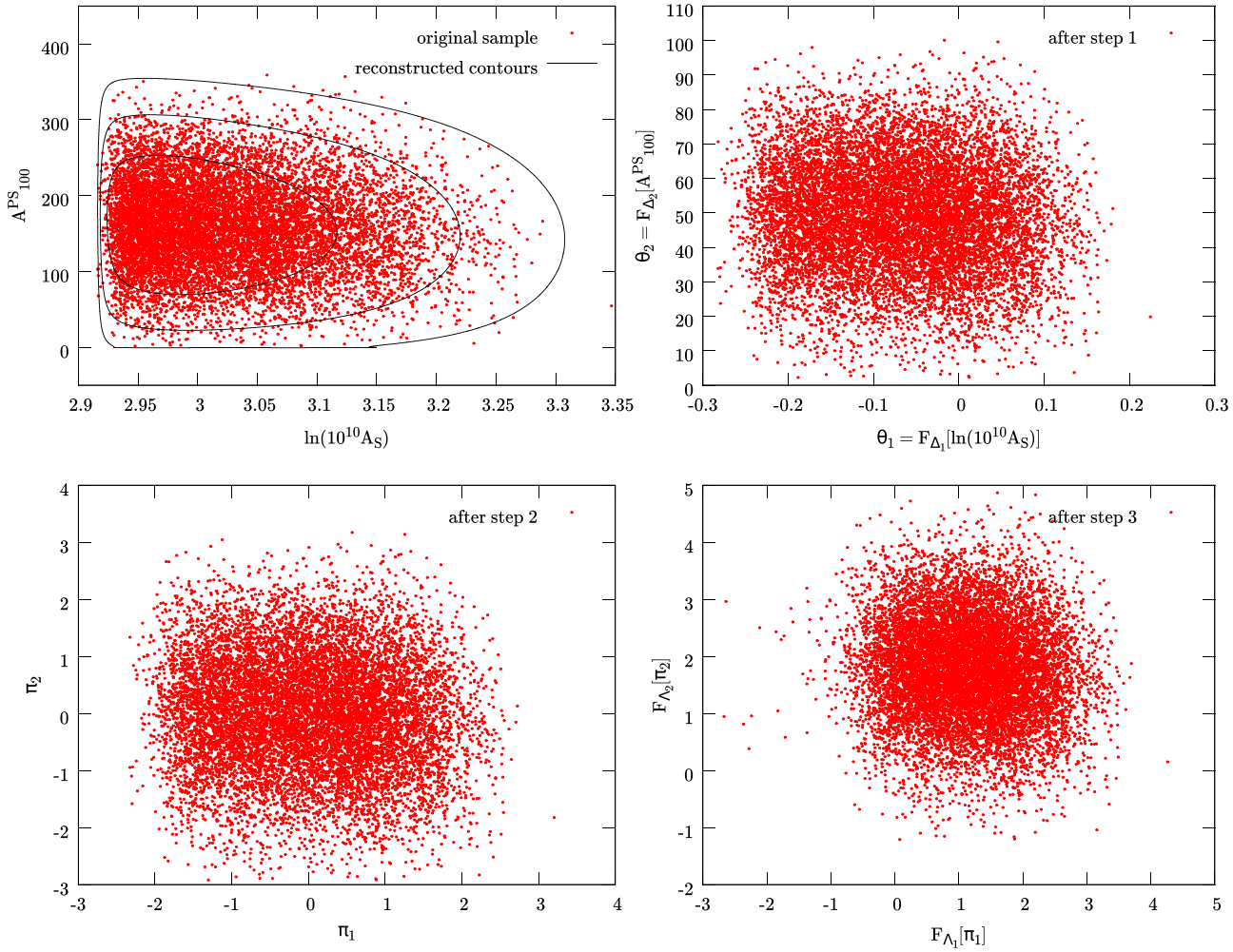

Figure 6. Two-pass Gaussianization of a wall-like non-Gaussian feature in a 2D marginal *Planck* posterior via ABC transformation; in the same format as Fig. 5, but with the TPs (Δ_1, Δ_2) and (Λ_1, Λ_2) that have been found for this sample. It is apparent that after the first transformation, the parameter θ_1 still exhibits residual kurtosis.

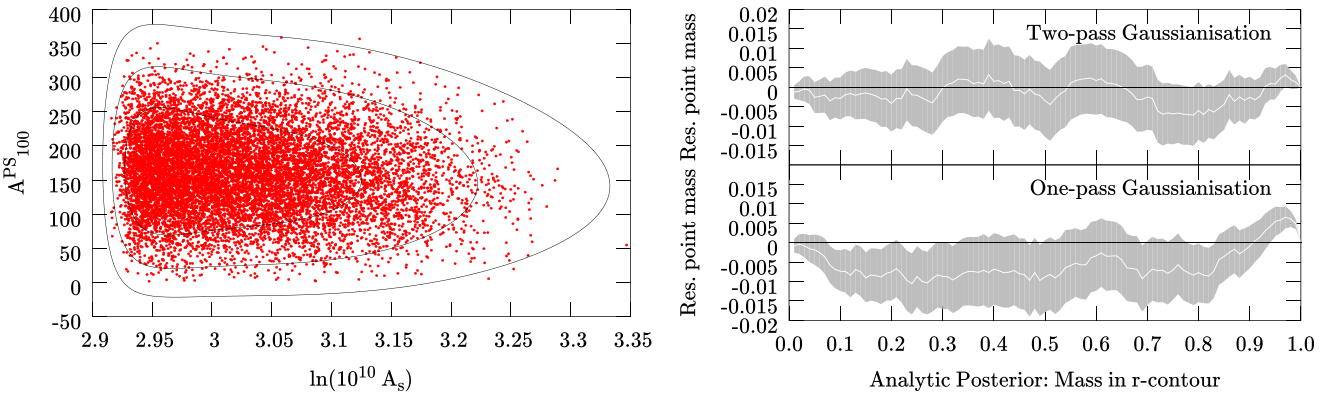

Figure 7. One-pass Gaussianization of a 2D marginal *Planck* posterior via ABC transformation. Left: original sample and contours of the analytic posterior. Top right: CC plot for the two-pass Gaussianization in Fig. 6. Bottom right: CC plot for the one-pass Gaussianization (see left-hand panel), showing deviations of the CC masses.

Gaussianizes $\widetilde{\Pi}(\boldsymbol{Y}) = \Pi[T^{-1}(\boldsymbol{Y})]\,|\mathrm{d}\boldsymbol{X}/\mathrm{d}\boldsymbol{Y}|$, we can compute the evidence integral analytically. If $\widetilde{\Pi}$ has the shape of an (unnormalized) multivariate Gaussian

$$\widetilde{\Pi}(\boldsymbol{Y}) = \widehat{\Pi}\ \exp\left[-\frac{1}{2}(\boldsymbol{Y}-\tilde{\boldsymbol{\mu}})^T\,\tilde{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{Y}-\tilde{\boldsymbol{\mu}})\right] \tag{12}$$

with means $\tilde{\boldsymbol{\mu}}$, covariance matrix $\tilde{\boldsymbol{\Sigma}}$, and maximum $\widehat{\Pi}$, the log-evidence reads

$$\ln E = \ln\widehat{\Pi} + \frac{1}{2}\ln\det\tilde{\boldsymbol{\Sigma}} + \frac{d}{2}\ln(2\pi). \tag{13}$$

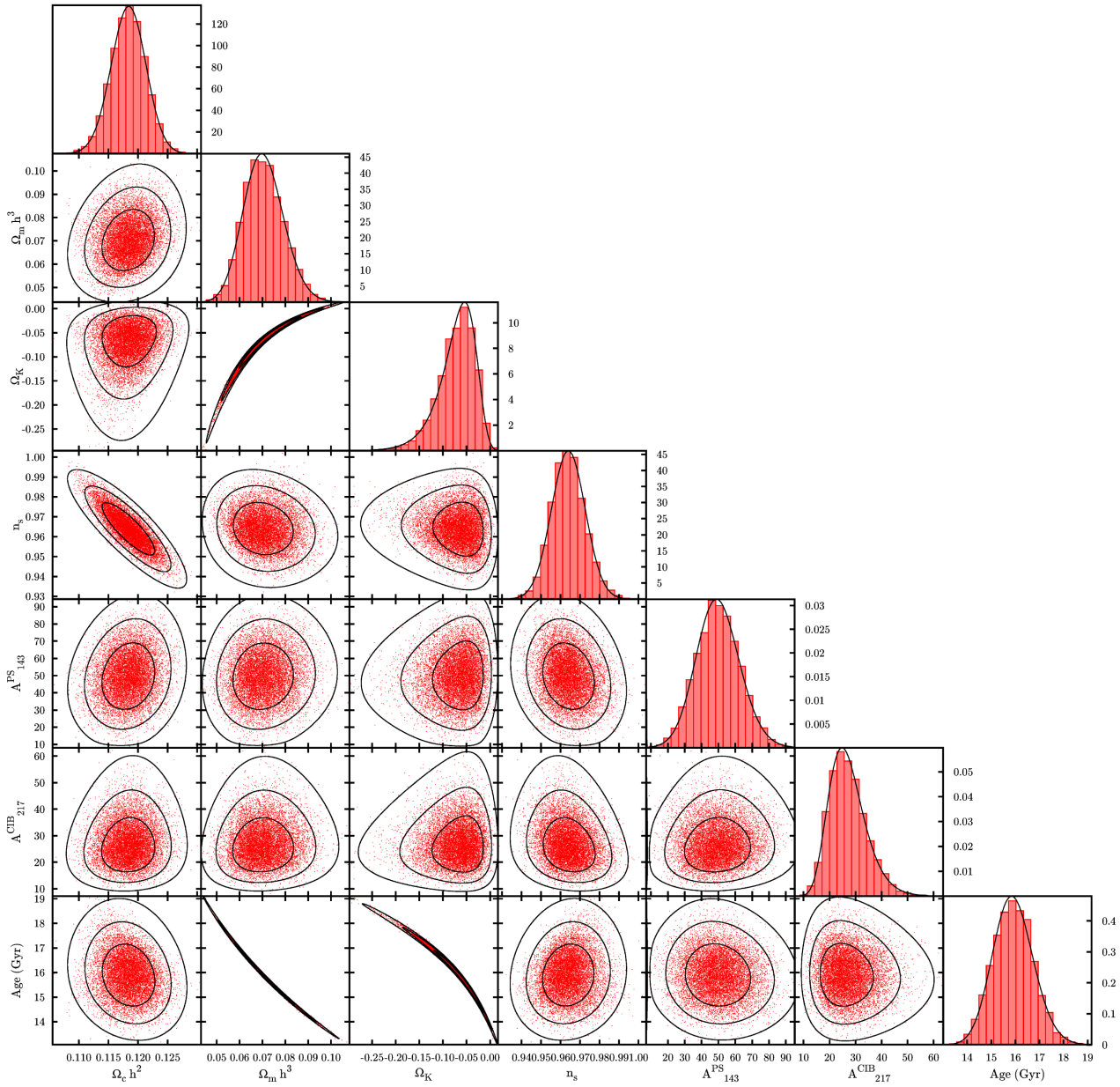

Figure 8. Reconstruction of a 7D *Planck* posterior density via a one-pass ABC transformation: 1D and 2D marginals. Black: marginal analytic posterior density (1D) or $1\sigma, 2\sigma, 3\sigma$ contours. Red: marginal point sample distributions. For the 1D cases, the histograms have renormalized bar heights, to demonstrate the agreement with the value of the probability density.

Similar expressions for Gaussian posterior densities can be found in Taylor & Kitching (2010). To estimate $\widehat{\Pi}$, we need the absolute normalization of $\widetilde{\Pi}$; hence this method can only be applied to samples which provide the values for Π (possibly also in the form of log-likelihood and log-prior). From these, we compute the values of $\ln \widetilde{\Pi}(\boldsymbol{Y})$ on the optimally Gaussianized sample by adding the logarithm of the transformation Jacobian, and then fit the parameters $\tilde{\boldsymbol{\mu}}$, $\tilde{\boldsymbol{\Sigma}}$, and $\widehat{\Pi}$ of the Gaussian via least-squares regression. This can be performed analytically, and even be used to compute an error bar on the value of $\ln E$ – see Appendix A for details.

If the prior distribution for one MP, and hence the posterior, is supported only on a finite interval, the same will hold true for the transformed MP if we restrict ourselves to one-pass transformations. If the sample size is large enough to properly represent the cut-off, the Gaussianization transformation will alleviate this feature, but may not fully remove it. Assuming the marginal distribution to be Gaussian, when in reality we may deal with a truncated Gaussian, will lead to a systematic error in the evidence, so it is advantageous to remove these features before starting the search for optimally Gaussianizing TPs. Appendix B details 'unboxing transformations', which redefine the MPs, mapping a finite open interval to the entire real line. In fact, it is also possible to use them for posterior density reconstruction, before Step 1 in Section 2.4.

To demonstrate this idea, we compute the evidence integral first on a mock data set, and subsequently on real data from cosmology. For the former, we draw a random sample of length 10 000 from a 10D log-normal probability distribution, and assign to each point the value of the probability density function, multiplied with a factor of $E = \exp(5)$. All weights are set to unity. This mock sample is subjected first to the Gaussianization procedure with

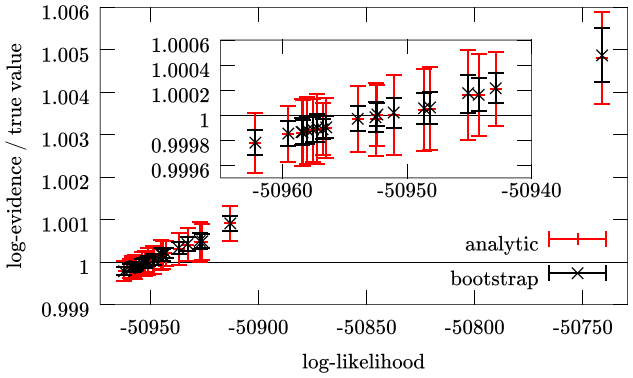

Figure 9. Ratio of log-evidence values for mock data set drawn from a 10D log-normal distribution, computed after 24 independent Gaussianization runs to the true value, versus the negative log-likelihood for each run (including the penalty term). Values and error bars shown are from analytic regression, as detailed in Appendix A (red), and the means and 1σ confidence intervals from the bootstrapped distributions (black). The black line indicates the true value. The inset is a zoomed-in version of the lower left corner.

one-pass ABC transformations (no unboxing), and then to the regression outlined in Appendix A to retrieve the best-fitting estimate for $\ln E$ and its error bar. To verify these, we determine the distribution of the estimator in equation (13) by producing 1000 bootstrap samples from the transformed sample, and computing $\ln E$ on each of them, together with its 1σ confidence intervals. To increase the reliability of the optimally Gaussianizing transformation, 24 independent Gaussianization runs are started with randomized initial conditions; Fig. 9 shows the results for these. Relative to the 'true' value of 5, these agree to subpercent accuracy. It is also noteworthy that a lower local maximum of the likelihood, i.e. one further to the right, will depart more from the true value. Hence, a location in TP space that is close to the exact optimum will yield a biased value for the log-evidence. This indicates that the evidence is a sensitive indicator for departures from Gaussianity. Note that a log-normal distribution can be precisely Gaussianized with Box–Cox transformations – and thus also by ABC transformations, which are a superset of these. Further, it is noteworthy that our analytic procedure yields error bars of the right magnitude, yet somewhat more conservative, compared to their bootstrapped counterparts. This discrepancy arises because the bootstrapped distributions for $\ln\det\widetilde{\mathbf{\Sigma}}$ and $\ln\widetilde{\Pi}$ deviate slightly from Gaussianity, whereas the analytic error bars assume Gaussian error propagation (see Appendix A for details).

For a demonstration on real-world data, we Gaussianize the joint posterior distribution of MPs of data from weak lensing and baryon acoustic oscillations (BAO). The weak lensing data set is the 2D cosmic shear data taken by the Canada–France–Hawaii Telescope Lensing Survey (CFHTLenS; see Heymans et al. 2012; Kilbinger et al. 2013). The CFHTLenS survey analysis combined weak lensing data processing with THELI (Erben et al. 2013), shear measurement with lensfit (Miller et al. 2013), and photometric redshift measurement with point spread function-matched photometry (Hildebrandt et al. 2012). A full systematic error analysis of the shear measurements in combination with the photometric redshifts is presented in Heymans et al. (2012), with additional error analyses of the photometric redshift measurements presented in Benjamin et al. (2013).

The BAO data set is the Data Release 9 (DR9) constant mass (CMASS) sample from the Baryon Oscillation Spectroscopic Survey (BOSS), which is part of the Sloan Digital Sky Survey III (see Anderson et al. (2012). This contains 264 283 massive galaxies in a redshift range $0.43 < z < 0.7$, whose correlation function and power spectrum both exhibit the features of BAO. The quantity $d(z) = r_S(z_d)/D_V(z)$, i.e. the ratio of the comoving sound horizon r_S at the baryon drag epoch z_d and the spherically volume-averaged distance $D_V(z)$, is a probe of the underlying cosmological parameters – see Percival et al. (2007) for details.

To draw samples from the posterior distribution, we use the COSMOPMC software package[3], which uses Population Monte Carlo (PMC), an algorithm to approximate the target distribution by a Gaussian mixture model. We compare three cosmological models: standard flat ΛCDM, curved ΛCDM, flat wCDM, and curved wCDM. The first has a 4D parameter space spanned by matter density Ω_m, power spectrum normalization σ_8, baryon density Ω_b, and the normalized Hubble parameter h_{100} – all other parameters are set to their best-fitting values for flat ΛCDM, see Planck Collaboration XIII (2015). The latter two contain a fifth model variable each – curvature parameter Ω_K and constant dark energy equation-of-state parameter w, respectively. For all of these parameters, flat proper priors were chosen. The baseline model – flat ΛCDM is always referred to as model 1, whereas model 2 is one of the two extensions. As a byproduct of the sampling process, PMC provides the model evidence for the data set used – see Kilbinger et al. (2010) for further details.

In the special situation where one model is nested inside the other, the evidence ratio $B_{12} = E_1/E_2$ can be computed via the Savage–Dickey density ratio (SDDR) – see Dickey (1971); Verde et al. (2013), and citations therein. Under mild conditions on prior and posterior densities for the full model $\mathcal{M}_2$ and the submodel $\mathcal{M}_1$, the ratio can be derived to be

$$B_{12} = \frac{\mathcal{P}(\psi = \psi_\mathrm{sub}|\mathcal{D}, \mathcal{M}_2)}{\mathcal{P}(\psi = \psi_\mathrm{sub}|\mathcal{M}_2)}, \tag{14}$$

where ψ denotes the extra parameter (or parameters) contained in $\mathcal{M}_2$ but not in $\mathcal{M}_1$, ψ_sub is the value of ψ that specifies the submodel $\mathcal{M}_1$, and $\mathcal{P}(\psi|\mathcal{D}, \mathcal{M}_2)$ and $\mathcal{P}(\psi|\mathcal{M}_2)$ are posterior and prior densities of the full model, marginalized over all MPs but ψ (see Verde et al. 2013 for details).

We find that the log-evidence values computed by COSMOPMC need to be offset by a factor of $n - 1$ times the log-prior density, where n is the number of data sets used. This is due to a non-standard interpretation of the prior density within COSMOPMC. Throughout this work, we apply this correction to the log-evidence values produced by COSMOPMC as well as the log-posterior values extracted from the COSMOPMC output. We follow the practice of Kilbinger et al. (2010, 2013) of accepting a COSMOPMC run as soon as the built-in convergence diagnostic, called perplexity, exceeds a value of $p > 0.7$. Sampling to even higher values for the perplexity, up to $p \sim 0.95$, still changes the COSMOPMC value for $\ln E$ by as much as ~ 0.1 – this indicates a residual bias in the statistic. However, since the exact same offset has to appear in the COSMOPMC output values for the log-evidence $\ln E$ and for the non-normalized log-posterior $\ln \Pi(\boldsymbol{X})$, it is not of relevance to demonstrating our method, so investigating its origin is beyond the scope of this work.

Table 2 shows the log-evidences for the three models, and the Bayes factors of ΛCDM compared to either of the two extended models. The numbers in the first line were computed via one-pass Gaussianization with ABC transformations, preceded by an

[3] See http://www2.iap.fr/users/kilbinge/CosmoPMC/.

unboxing transformation. To estimate the scatter of the COSMOPMC and SDDR values for $\ln E$ and $\ln B_{12}$ in the second, fourth, and fifth lines, we rerun COSMOPMC 10 times for each model, and determine the mean and average for the COSMOPMC and SDDR estimators. Like for the log-normal sample, 24 independent Gaussianization runs were started for each sample, and the one with the highest log-likelihood value chosen to transform the sample, which is then subjected to the analytic evidence computation procedure. The values in the first row of Table 2 are the weighted averages and standard deviations of all 10 values, where the weights are determined from the analytic error bars as $w_i = \sigma_i^{-2}$ (cf. Appendix A). The values show that the combined data favour ΛCDM over any of the extended models, although the evidence is not strong against either of the two. Our values agree with the numbers of SDDR and PMC within the spread between the latter two estimators, but still small deviations remain, which are larger than the error bars quoted. These may be due to residual non-Gaussianity in the transformed samples, to which the evidence is a sensitive measure.

6 CONCLUSIONS AND OUTLOOK

We have discussed how to transform a posterior probability density approximately into a multivariate Gaussian, and various applications thereof.

(1) From the parameters of the Gaussianized distribution and those of the transformation, we can reconstruct an analytic expression for the original posterior probability density, given a point sample drawn from it. This facilitates the combination of different data sets to obtain the joint posterior density.

(2) Further, this analytic posterior can be used to display contours of the density in question or its marginals, without the need for density estimates or smoothing procedures. Also, in reproducing the contours of the probability density reliably, it outperforms kernel density estimates.

(3) We suggest that, instead of distributing lengthy point samples in the form of a Markov chain, to use a Gaussianizing transformation to disseminate a posterior density. Only the TPs and the first and second moments of the resulting Gaussian are needed to reproduce the posterior density in its functional form; hence we can achieve substantial data compression.

We have demonstrated this algorithm with our implementation in C (code on request), which employs MCMC samples from *Planck* data. We used Box–Cox and ABC transformations to Gaussianize various marginal distributions with distinctive non-Gaussian features, and showed the resulting contours.

Table 2. Values for evidence and Bayes factor for CFHTLenS+BOSS data set in three cosmological models, as computed with Box–Cox Gaussianization (G) of weighted samples with 10 000 points each. For comparison: evidence value $\ln E$ and Bayes factor $\ln B_{12} = \ln E_{\rm base} - \ln E_{\rm extension}$ from PMC, and Bayes factor from SDDR.

	Flat ΛCDM	Curved ΛCDM	Flat wCDM
Dimension	4	5	5
$\ln E$ (G)	486.96 ± 0.01	485.79 ± 0.03	486.09 ± 0.05
$\ln E$ (PMC)	487.02 ± 0.03	485.84 ± 0.01	486.00 ± 0.04
$\ln B_{12}$ (G)	na	1.17 ± 0.04	0.87 ± 0.05
$\ln B_{12}$ (SDDR)	na	1.23 ± 0.04	0.93 ± 0.06
$\ln B_{12}$ (PMC)	na	1.19 ± 0.04	1.03 ± 0.05

One distinctive application of Gaussianizing transformations, which we discuss and demonstrate here, is a novel method to compute the model evidence of a posterior distribution, given a point sample from it. We have tested this method on cosmological data from lensing and BAO, for different cosmological models, and find slight preference for ΛCDM. Compared to the numerical results from PMC and the SDDR, our new method of computing the evidence agrees well within the spread of the other two.

We have introduced the CC plot as a tool to decide whether one probability density reproduces the contours of another, or if they do not, to detect where they deviate.

There are several possible extensions of our method, and directions to advance its scope.

To optimize the Gaussianization algorithm for speed and/or accuracy, it is possible to replace the Nelder–Mead minimum finder with other, more sophisticated algorithms, such as simulated annealing (Černý 1985), or BOBYQA (Powell 2007).

It is possible to engage new families of transformations, designed to cure a wider spectrum of non-Gaussian features that a multivariate probability density may possess – in our implementation, new families can easily be included.

Gaussianization may be employed for fast sampling from a non-Gaussian probability density, in case that the Gaussianizing parameters are either known exactly or to sufficient accuracy. Afterwards, it is possible to quickly draw a point sample from a multivariate Gaussian distribution, and transform this sample with the inverse map.

To improve the accuracy of the evidence computation, it is possible to replace the log-likelihood of equation (5) with another loss function, which penalizes deviations from Gaussianity in a sharper manner.

So far, we have been working with unimodal probability densities. We require the transformations to be bijective, hence we cannot map a multimodal distribution into a unimodal Gaussian. However, we may be able to transform such a density into a mixture of (possibly overlapping) Gaussians, where we now have to estimate the weight factor for each constituent from the transformed sample, in addition to each $\tilde{\boldsymbol{\mu}}$ and $\tilde{\boldsymbol{\Sigma}}$. The requisite number of components can possibly be determined with standard clustering algorithms.

ACKNOWLEDGEMENTS

RLS is grateful to Martin Kilbinger, Karim Benabed, Catherine Heymans, and Stephen Feeney for helpful discussions, and has been supported by the Perren Fund in Astronomy, and by IMPACT (UCL MAPS faculty); BJ acknowledges support by an STFC Ernest Rutherford Fellowship, grant reference ST/J004421/1; HVP was supported by the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013)/ERC grant agreement no.306478-CosmicDawn. This work is based on observations obtained with *Planck* (http://www.esa.int/Planck), an ESA science mission with instruments and contributions directly funded by ESA Member States, NASA, and Canada. In part, this work was supported by National Science Foundation Grant No. PHYS-1066293 and the hospitality of the Aspen Center for Physics.

REFERENCES

Anderson L. et al., 2012, MNRAS, 427, 3435
Benjamin J. et al., 2013, MNRAS, 431, 1547

Box G. E. P., Cox D. R., 1964, J. R. Stat. Soc. B, 26, 211
Černý V., 1985, J. Optim. Theory Appl., 45, 41
Chu M., Kaplinghat M., Knox L., 2003, ApJ, 596, 725
Dickey J., 1971, Ann. Math. Stat., 42, 204
Erben T. et al., 2013, MNRAS, 433, 2545
Fendt W. A., Wandelt B. D., 2007, ApJ, 654, 2
Heymans C. et al., 2012, MNRAS, 427, 146
Hildebrandt H. et al., 2012, MNRAS, 421, 2355
Jaynes E. T., 2003, Probability Theory: The Logic of Science. Cambridge Univ. Press, Cambridge
Jimenez R., Verde L., Peiris H. V., 2004, Phys. Rev. D, 70, 023005
Joachimi B., Taylor A., 2011, MNRAS, 416, 1010
Kass R., Raftery A. E., 1995, J. Am. Stat. Assoc., 90, 779
Kilbinger M. et al., 2010, MNRAS, 405, 2381
Kilbinger M. et al., 2013, MNRAS, 430, 2200
Kitching T., Taylor A., 2010, MNRAS, 408, 865
Kosowsky A., Milosavljevic M., Jimenez R., 2002, Phys. Rev. D, 66, 063007
Lewis A., Bridle S., 2002, Phys. Rev. D, 66, 103511
MacKay D. J. C., 2003, Information Theory, Inference, and Learning Algorithms. Cambridge Univ. Press, Cambridge
Miller L. et al., 2013, MNRAS, 429, 2858
Nelder J. A., Mead R., 1965, Comput. J., 7, 308
Percival W. J., Cole S., Eisenstein D. J., Nichol R. C., Peacock J. A., Pope A. C., Szalay A. S., 2007, MNRAS, 381, 1053
Planck Collaboration XVI 2014, A&A, 571, A16)
Planck Collaboration XIII 2015, A&A, in press
Powell M. J. D., 2007, IMA J. Numer. Anal., 28, 649
Sandvik H., Tegmark M., Wang X., Zaldarriaga M., 2004, Phys. Rev. D, 69, 063005
Silverman B. W., 1998, Density Estimation for Statistics and Data Analysis. Chapman & Hall/CRC, London
Skilling J., 2006, Bayesian Anal., 1, 833
Velilla S., 1993, Stat. Probab. Lett., 17, 259
Verde L., Feeney S. M., Mortlock D. J., Peiris H. V., 2013, J. Cosmol. Astropart. Phys., 09, 0013

APPENDIX A: ANALYTIC EVIDENCE COMPUTATION

We outline how the computation of the log-evidence (see equation 13) can be performed analytically, i.e. without numerical optimization. Our data consist of a Gaussianized weighted sample of $\mathcal{N}$ points in $\mathbf{R}^d$, $\{(\boldsymbol{Y}^a, w^a)\}_{a=1}^{\mathcal{N}}$ and the values of the transformed log-posterior on each of these points, $\{\ell^a\}_{a=1}^{\mathcal{N}}$. To fit a multivariate unnormalized Gaussian

$$\widetilde{\Pi}(\boldsymbol{Y}) = \widehat{\Pi} \exp\left[-\frac{1}{2}(\boldsymbol{Y}-\tilde{\boldsymbol{\mu}})^T \tilde{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{Y}-\tilde{\boldsymbol{\mu}})\right] \tag{A1}$$

through the values of $\{\exp(\ell^a)\}_{a=1}^{\mathcal{N}}$, we use the regression model

$$\ell^{\text{model}}_{\boldsymbol{A},\boldsymbol{B},C}(\boldsymbol{Y}) = \boldsymbol{Y}^T \boldsymbol{A}\, \boldsymbol{Y} + \boldsymbol{B}^T \boldsymbol{Y} + C, \tag{A2}$$

which is linear in each of the $d(d+3)/2+1$ regression parameters: the upper-diagonal components of the symmetric matrix $\mathbf{A}$, the components of vector $\boldsymbol{B}$, and the scalar C. Assuming independence and homoscedasticity, we arrive at our quantity to minimize,

$$\chi^2(\mathbf{A}, \boldsymbol{B}, C) = \sum_{a=1}^{\mathcal{N}} w^a \left[\ell^{\text{model}}_{\boldsymbol{A},\boldsymbol{B},C}(\boldsymbol{Y}^a) - \ell^a\right]^2, \tag{A3}$$

which is quadratic in every regression parameter. Thus, we can write all normal equations of the regression problem,

$$\frac{\mathrm{d}\chi^2}{\mathrm{d}\vartheta} \stackrel{!}{=} 0, \quad \vartheta \in \{A_{ij}, B_k, C\} \tag{A4}$$

as a $[d(d+3)/2+1]$-dimensional linear inhomogeneous vector equation, and solve via singular value decomposition. From the resulting values of $(\mathbf{A}, \boldsymbol{B}, C)$, the parameters of the multivariate Gaussian (equation A1) can readily be computed as

$$\tilde{\boldsymbol{\Sigma}} = -\frac{1}{2}\mathbf{A}^{-1}; \tag{A5}$$

$$\tilde{\boldsymbol{\mu}} = -\frac{1}{2}\mathbf{A}^{-1}\boldsymbol{B}; \tag{A6}$$

$$\ln\widehat{\Pi} = C - \frac{1}{4}\boldsymbol{B}^T\mathbf{A}^{-1}\boldsymbol{B}. \tag{A7}$$

Furthermore, we can use the analytic regression procedure to find error bars on these estimators, and thus on $\ln E$. To this end, we can analytically find the covariance matrix $\mathbf{Cov}$ of all parameters $\{A_{ij}, B_i, C\}$ from the form of χ^2 and the transformed data set $\{(\boldsymbol{Y}^a, w^a, \ell^a)\}_{a=1}^{\mathcal{N}}$, and then approximate the variances of $\ln\widehat{\Pi}$ and of $\ln\det\tilde{\boldsymbol{\Sigma}}$ by standard Gaussian error propagation. In particular

$$\mathrm{Var}\left[\ln E\right] = \mathrm{Var}\left[\ln\widehat{\Pi}\right] + \frac{1}{4}\mathrm{Var}\left[\ln\det\tilde{\boldsymbol{\Sigma}}\right] \tag{A8}$$

$$\simeq \boldsymbol{\Xi}^T\mathbf{Cov}\,\boldsymbol{\Xi} + \frac{1}{4}\boldsymbol{\Upsilon}^T\mathbf{Cov}\boldsymbol{\Upsilon} \tag{A9}$$

with

$$\boldsymbol{\Xi} = \left(\frac{\partial\ln\widehat{\Pi}}{\partial\vartheta_i}\right) = \left(\begin{array}{cc} \vdots & \\ \frac{1}{4}\boldsymbol{B}'_m\boldsymbol{B}'_n(2-\delta_{mn}) & (m=1\ldots d, \\ \vdots & n=m\ldots d) \\ \hline \vdots & \\ -\frac{1}{2}\boldsymbol{B}'_k & (k=1\ldots d) \\ \vdots & \\ \hline 1 & \end{array}\right) \tag{A10}$$

and

$$\boldsymbol{\Upsilon} = \left(\frac{\partial\ln\det\tilde{\boldsymbol{\Sigma}}}{\partial\vartheta_i}\right) = \left(\begin{array}{cc} \vdots & \\ \frac{1}{4}(\mathbf{A}^{-1})_{mn}(2-\delta_{mn}) & (m=1\ldots d, \\ \vdots & n=m\ldots d) \\ \hline \vdots & \\ 0 & \\ \vdots & \\ \hline 0 & \end{array}\right), \tag{A11}$$

where $\boldsymbol{B}' = \mathbf{A}^{-1}\boldsymbol{B}$.

APPENDIX B: UNBOXING TRANSFORMATIONS

A single MP Z, which is assumed to be constrained to an open interval (a, b), is redefined via the unboxing transformation $U_{(a,b)} : (a, b) \to \mathbf{R}$

$$X = U_{(a,b)}(Z) = \frac{a+b}{2} + \frac{b-a}{\sqrt{2\pi}} \Phi^{-1}\left(\frac{Z-a}{b-a}\right), \tag{B1}$$

where Φ^{-1} denotes the inverse of the cumulative distribution function of the Normal distribution:

$$\Phi(x) = \int_{-\infty}^{x} \mathrm{d}y \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right). \tag{B2}$$

$U_{(a,b)}$, thus designed, has the following properties: it is bijective and smooth; the limits are $\lim_{Z\to a} U_{(a,b)}(Z) = -\infty$; $\lim_{Z\to b} U_{(a,b)}(Z) = +\infty$. Further, the mid-point of the interval $m = \frac{1}{2}(a+b)$ is fixed: $U_{(a,b)}(m) = m$, $U'_{(a,b)}(m) = 1$. Sending the interval boundaries to infinity simultaneously will result in the identity transformation

$$\lim_{\substack{a\to-\infty \\ b\to+\infty}} U_{(a,b)}(Z) = Z. \tag{B3}$$

This is a generalization of the widely used probit transformation, which maps the unit interval on to the real numbers as $p \mapsto \Phi^{-1}(p)$. Our modified probit has one huge advantage for the subsequent search for a Gaussianizing transformation: If Z, as a random variable, is uniformly distributed on (a, b), then X is normally distributed with mean m and spread $(b-a)/\sqrt{2\pi}$.

In statistics, a frequently used alternative to probit is the logit map $p \mapsto \log(p/1-p)$. For our purposes, however, the probit is preferable, since a similarly rescaled version of this logit-transformation would yield a distribution with excess kurtosis, instead of a Gaussian.

For a d-dimensional vector of MPs $\boldsymbol{Z} = (Z_1, \ldots, Z_d)$, constrained to intervals $(a_i, b_i) \ni Z_i$, we unbox each dimension separately, with the appropriate boundaries:

$$\boldsymbol{Z} \mapsto \boldsymbol{X} = \left[U_{(a_1,b_1)}(Z_1), \ldots, U_{(a_d,b_d)}(Z_d)\right]. \tag{B4}$$

Before starting the search for the Gaussianization parameters, every point in the original sample is mapped through this transformation.

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