Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

Mon. Not. R. Astron. Soc. **400,** 219–227 (2009) doi:10.1111/j.1365-2966.2009.15202.x

TEASING: a fast and accurate approximation for the low multipole likelihood of the cosmic microwave background temperature

K. Benabed,[1⋆] J.-F. Cardoso,[1,2] S. Prunet[1] and E. Hivon[1]

[1]*Institut d'Astrophysique de Paris, 98bis Bd Arago, 75014 Paris, France*
[2]*Laboratoire de Traitement et Communication de l'Information, LTCI/CNRS 46, rue Barrault, 75013 Paris, France*

Accepted 2009 June 8. Received 2009 June 5; in original form 2009 March 11

ABSTRACT

We propose a novel approximation to the low-ℓ joint likelihood of the angular spectrum C_ℓ of masked cosmic microwave background temperature maps which is both very accurate and very fast to evaluate. We show that, for a flat prior, the posterior distribution of each C_ℓ closely follows an inverse gamma distribution even with partial sky coverage and that the posterior correlation is weak enough that a copula approximation to the joint likelihood is quite accurate. In this paper, the quantities needed to build such a copula approximation (inverse gamma parameters at each angular frequency and a correlation matrix) are computed from an exploration of the posterior using adaptive importance sampling. The accuracy of the proposed approximation is assessed using statistical criteria as well as a mock cosmological parameter fit. When applied to the *Wilkinson Microwave Anisotropy Probe* 5 data set, the copula approximation yields cosmological parameter estimates at the same level of accuracy as the best current techniques.

Key words: methods: data analysis – methods: statistical – cosmic microwave background.

1 INTRODUCTION

The cosmic microwave background (CMB) angular spectrum $\boldsymbol{C} = \{C_\ell\}$ is a central quantity for conducting statistical inference based on CMB observations (Bond & Efstathiou 1987). The high resolution of available (Hinshaw et al. 2009) and forthcoming CMB observations (Efstathiou, Lawrence & Tauber 2005) makes it necessary (at least in the case of partial sky coverage) to adopt a processing scheme in which the low- and high-ℓ parts of the data are processed independently (Efstathiou 2006). This paper addresses the large-scale part of the problem: inference regarding low multipoles based on a partial low-resolution CMB map.

After defining the problem of low-ℓ pixel-based likelihood and introducing some notations (Section 2), we first show how to build a (large) set of *N importance samples* of the angular spectrum such that all integrals of interest for statistical inference can be approximated by Monte Carlo estimates (Section 3). Based on those results, we propose in Section 4 a new approximation to the likelihood for partially observed low-resolution CMB maps. This approximation was initially built as part of the importance sampler, but it turns out to be so accurate that it is of independent interest. This paper and the recent reference (Rudjord et al. 2009) are similar in spirit but differ in the sampling method and in the proposed likelihood approximation.

⋆E-mail: benabed@iap.fr

2 LIKELIHOOD

We recall some well-known facts about the likelihood of the angular spectrum of a CMB temperature map.

In the ideal case of noise-free, beam-free, full-sky map (represented by the vector $\boldsymbol{x}$ of pixels), one has direct access to the harmonic coefficients $a_{\ell m}$ of the sky. Assuming an isotropic Gaussian field, the empirical angular spectrum $\widehat{C}_\ell = \frac{1}{2\ell+1}\sum_m |a_{\ell m}|^2$ is a sufficient statistic for the data and their probability distribution takes the factorized form (Bond, Jaffe & Knox 2000):

$$p(\boldsymbol{x}|\boldsymbol{C}) \propto \prod_{\ell\geq 0} \exp -\frac{2\ell+1}{2}\left(\frac{\hat{C}_\ell}{C_\ell} + \log C_\ell\right). \quad (1)$$

In the case of a flat prior $p(C)$, expression (1) combined with the Bayes rule $p(C|\boldsymbol{x}) = p(\boldsymbol{x}\,|C)p(C)/p(\boldsymbol{x})$ reveals that, given $\boldsymbol{x}$, the angular spectrum C is distributed as a product of inverse gamma densities:

$$p(\boldsymbol{C}|\boldsymbol{x}) = \prod_\ell i\Gamma(C_\ell; \alpha_\ell, \beta_\ell) \quad (2)$$

$$i\Gamma(x;\alpha,\beta) \equiv \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1}\mathrm{e}^{-\beta/x}, \quad (3)$$

with parameters $\alpha_\ell = (2\ell-1)/2$ and $\beta_\ell = (2\ell+1)\hat{C}_\ell/2$.

Such a factorization does not hold when only a fraction of the sky is observed (or has to be ignored because of excessive contamination by foregrounds), or when the stationary CMB is contaminated by non-stationary noise (Gorski 1994; Tegmark 1997).

However, for small sky masks and/or small deviations from stationarity, deviations from the factorized form (1) are expected to be small, suggesting the new likelihood approximation developed in Section 4.

Pixel-based likelihood. We turn to the actual case of interest: partial sky coverage, presence of independent additive Gaussian noise, low-pass effect of a beam. The data set, represented by an $N_{\rm pix} \times 1$ vector $\boldsymbol{x}$ of pixel values, can no longer be losslessly compressed into a sufficient spectral statistic $\hat{C}_\ell$. Rather, one must use the plain Gaussian density:

$$p(\boldsymbol{x}|\mathbf{R}) = |2\pi\mathbf{R}|^{-1/2} \mathrm{e}^{-\frac{1}{2}\boldsymbol{x}^T \mathbf{R}^{-1} \boldsymbol{x}}, \tag{4}$$

where the covariance matrix $\mathbf{R}$ of $\boldsymbol{x}$ has contributions from the CMB signal and from noise: $\mathbf{R} = \mathbf{S}+\mathbf{N}$. For two pixels i and j with angular separation θ_{ij}, the CMB part of the covariance matrix has an (i, j) entry given by (Bond et al. 2000)

$$\boldsymbol{S}_{ij} = \sum_\ell \frac{2\ell+1}{4\pi} W_\ell C_\ell P_\ell(\cos\theta_{ij}), \tag{5}$$

where P_ℓ is the Legendre polynomial of the order of ℓ and where the window function W_ℓ can represent, for example, the spectral response of an azimuthally symmetric beam, or more generally the convolution of the signal with any azimuthally symmetric kernel. Hence, we ignore the complications due to an anisotropic beam as well as the presence of residual foreground contaminants.

The noise part of the covariance matrix could take any form but, in this work, it is taken to correspond to an isotropic noise with angular spectrum N_ℓ. We can thus define a total angular spectrum D_ℓ

$$D_\ell = W_\ell C_\ell + N_\ell, \tag{6}$$

which is unambiguously related to C_ℓ since the beam B_ℓ and the noise spectrum N_ℓ are assumed to be known.

Free parameters. In practice, we consider a more restricted model for the covariance matrix of the observed pixels. First, the adjustable multipoles are restricted to a range $\ell_{\rm min} \leq \ell \leq \ell_{\rm max}$, while other multipoles are kept at constant values. Secondly, we only consider uncorrelated noise with zero mean and variance σ^2 per pixel. It contributes a term $\sigma^2\delta_{ij}$ to $\mathbf{R}$ and corresponds to a flat angular spectrum $N_\ell = \sigma^2/\Omega_{\rm pix}$ if all pixels have the same area $\Omega_{\rm pix}$. Then, the covariance matrix of $\boldsymbol{x}$ as a function of $\boldsymbol{D} = \{D_\ell\}_{\ell=\ell_{\rm min}}^{\ell=\ell_{\rm max}}$ is spelled out as $\mathbf{R}(\boldsymbol{D}) = \mathbf{R}^{\rm var}(\boldsymbol{D}) + \mathbf{R}^{\rm cst}$ with

$$\mathbf{R}^{\rm var}_{ij}(\boldsymbol{D}) = \sum_{\ell=\ell_{\rm min}}^{\ell=\ell_{\rm max}} \frac{2\ell+1}{4\pi}(D_\ell - N_\ell) P_\ell(\cos\theta_{ij}), \tag{7}$$

$$\mathbf{R}^{\rm cst}_{ij} = \sum_{\ell \rm fixed} \frac{2\ell+1}{4\pi} W_\ell C_\ell P_\ell(\cos\theta_{ij}) + \sigma^2\delta_{ij}. \tag{8}$$

Priors and posterior distributions. In all the following, the prior distribution on $\boldsymbol{D}$ is taken to be flat for $D_\ell \geq N_\ell$. At all angular frequencies such that $W_\ell C_\ell \gg N_\ell$ (Fig. 1 illustrates the values used in this paper), this is almost identical to a flat prior on the positive values of C_ℓ. The posterior distribution of $\boldsymbol{D}$ given the data $\boldsymbol{x}$ is

$$\pi(\boldsymbol{D}) = p(\boldsymbol{D}|\boldsymbol{x}) \propto p[\boldsymbol{x}|\mathbf{R}(\boldsymbol{D})] \prod_{\ell=\ell_{\rm min}}^{\ell=\ell_{\rm max}} \mathbf{1}(D_\ell \geq N_\ell),$$

where $p[\boldsymbol{x}|\mathbf{R}(\boldsymbol{D})]$ is evaluated using equations (4), (7) and (8).

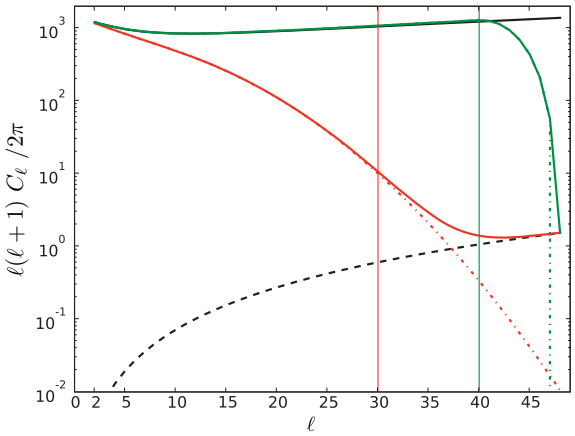

Figure 1. Angular spectra, rescaled by $\ell(\ell+1)/2\pi$: *WMAP* best-fitting spectrum C_ℓ (black solid line); noise spectrum N_ℓ for a variance of $\sigma^2 = 1\,\mu\mathrm{K}^2\,\mathrm{pixel}^{-1}$ (black dashed line); angular spectra $W_\ell C_\ell$ (dot–dashed) and $W_\ell C_\ell + N_\ell$ (solid) for the *WMAP* Gaussian beam (red) and for the window function of equation (11) (green).

About noise and regularization. On a cut sky, the CMB part of the covariance matrix may be poorly conditioned with a trough in its eigenvalue spectrum corresponding to those modes which are mostly localized in the cut. In this case, it is customary (Eriksen et al. 2007; Hinshaw et al. 2007) to add a very small amount of noise to the data and to add the corresponding contribution to the covariance matrix as in equation (8). Another reason for adding uncorrelated noise is to cover spurious noise correlation possibly introduced when the observed sky map is downgraded and to simplify the noise structure (Dunkley et al. 2009). See Fig. 1 for the values used in our experiments. Another possibility is regularization by projection on to the most significant eigenvectors of the covariance matrix (Bond et al. 2000), but this possibility is not considered here.

3 BUILDING A SAMPLE OF THE LOW-ℓ POSTERIOR WITH IMPORTANCE SAMPLING

This section reports on the construction of *importance samples* of the C_ℓ under their joint posterior for two data sets. The principle of importance sampling is first briefly recalled in Section 3.1; our specific technique (an adaptive variant) is described in Section 3.2 and applied to a synthetic CMB cut sky map (Section 3.3) and to the official *Wilkinson Microwave Anisotropy Probe* 5 (*WMAP*5) low-resolution map (Section 3.4).

3.1 Importance sampling

Importance sampling is a well-established technique to explore a probability distribution when no method for directly sampling from it is available [the well-known VEGAS algorithm (Lepage 1978), e.g., is based on importance sampling]. Consider estimating the expectation $Ef(x) = \int f(x)\pi(x)\,\mathrm{d}x$ of some function f of x when the random variable x is distributed under π. If x_i, $i = 1, N$ are N samples of x, then $Ef(x)$ can be estimated by the sample average $\frac{1}{N}\sum_i f(x_i)$. In contrast, importance sampling relies on samples x_i distributed under a *proposal distribution* g not necessarily equal to π. If the support of g includes the support of π, then

$$Ef = \int f(x)\pi(x)\mathrm{d}x = \int f(x)\frac{\pi(x)}{g(x)} g(x)\,\mathrm{d}x,$$

so that if the samples x_i are distributed under g, then Ef is estimated without bias by

$$\frac{1}{N}\sum_{i=1}^{N} w_i f(x_i), \quad \text{where} \quad w_i = w(x_i) \equiv \frac{\pi(x_i)}{g(x_i)}.$$

The factors w_i are called *importance weights*.

The Monte Carlo integration reaches its maximum efficiency when the samples are drawn independently under a proposal distribution g which is identical to the target distribution π. While Markov Chain Monte Carlo (MCMC) methods try to draw from the target distribution π, they do not build independent samples; in contrast, importance sampling (usually) relies on independent draws from an approximate distribution g and corrects the discrepancy using importance weights w_i. Therefore, importance sampling achieves higher relative accuracy (lower variance of posterior integrals) per sample compared to MCMC methods whenever independent samples can be drawn from a proposal distribution which is 'close enough' to the target (see e.g. Wraith et al. 2009).

The agreement between target and proposal distributions can be measured by the Kullback–Leibler divergence

$$K(\pi|g) \equiv \int \log \frac{\pi(x)}{g(x)}\, \pi(x)\, \mathrm{d}x, \tag{9}$$

which is often remapped as the so-called *perplexity criterion*: $\mathcal{P}(\pi|g) \equiv \exp -K(\pi|g)$ so that perfect agreement is reached when $\mathcal{P} = 1$. Another criterion is the *effective sample size* (ESS) of an importance sample:

$$\mathrm{ESS} = \frac{\left(\sum_i w_i\right)^2}{\sum_i w_i^2}. \tag{10}$$

If the proposal matches the target perfectly, then ESS $= N$, otherwise it is smaller than the number of importance samples. The ESS is directly related to the variance of the Monte Carlo estimates.

Importance sampling is well fitted to the problem at hand for at least two reasons: ease of parallelization and availability of a good proposal distribution.

Parallelization is a strong requirement due to the high computational cost of CMB studies. We are planning to sample a 30- to 40-dimensional space, and the computation of the likelihood for a given angular spectrum costs about 5 s for $\ell_{\max} = 48$ and $N_{\mathrm{pix}} = 3072$ on a typical 2 GHz CPU. Since importance sampling can be trivially parallelized, it makes it straightforward to take full advantage of CPU clusters. For instance, computing 10^5 samples would take about 4 days on a single CPU but is reduced to mere hours on a cluster. The MCMC algorithm cannot be parallelized as easily. Indeed, to be able to mix different parallel chains, one has to ensure that they have correctly converged (Rosenthal 2000), which can be a difficult task in 30 to 40 dimensions.

Regarding the proposal distribution, one can draw inspiration from the noise-free, full-sky case (2) since a mask hiding less than 20 per cent of the sky and a high signal-to-noise ratio situation are expected to modify it only slightly.[1] Indeed, as demonstrated below, a product of independent inverse gamma distributions turns out to be a very efficient proposal distribution, provided it is correctly tuned. Such a tuning is achieved via an *adaptive importance sampling*, as explained next.

Other techniques have been proposed to draw samples under the posterior distribution of the power spectrum. In particular, a clever rewriting of the problem allows for the use of some flavours of the MCMC (Gibbs Eriksen et al. 2004; Jewell, Levin & Anderson 2004; Wandelt, Larson & Lakshminarayanan 2004; or Hybrid MC-based, Taylor, Ashdown & Hobson 2008). Those techniques are much faster, i.e. have better scaling properties, than our importance sampler. We decided to stick to an importance sampler for its ease of implementation. In our case, poorer scaling properties are compensated by massive parallelization.

3.2 An adaptive importance sampling algorithm

Importance sampling is efficient only if the proposal distribution is close enough to the target, an objective which may be difficult to reach in large dimensions (sampling angular spectra in the range $0 \leq \ell \leq 40$ qualifies as large problem). To tackle this complexity, we resort to *adaptive importance sampling* which consists in running a sequence of importance runs in which the proposal distribution is improved at each run based on the results of previous runs. A more detailed description of adaptive importance sampling (based upon the Population Monte Carlo algorithm from Cappé et al. 2008) in the context of cosmology can be found in Wraith et al. (2009).

General scheme. The general scheme, based on a parametric family of proposal distributions $g(\boldsymbol{y}; \theta)$, is as follows.

(i) Start with the best available guess of θ for the parameters of the proposal distribution.

(ii) Sample under $g(\boldsymbol{y}; \theta)$. Compute and store the importance weights.

(iii) Re-estimate θ so that $g(\boldsymbol{y}; \theta)$ best matches the current sample set.

(iv) If the (estimated) perplexity $\mathcal{P}[\pi(\boldsymbol{y})|g(\boldsymbol{y}; \theta)]$ is high enough (e.g. above 0.5) or if it has not changed significantly during the last iterations, exit to (v). Otherwise, go to (ii) for another importance run with the re-estimated parameters.

(v) Use the last value of θ for a large final importance sampling run.

Sampling angular spectra. In our experiments, we sample the total angular spectrum, i.e. $\boldsymbol{y} = \boldsymbol{D} = \{D_\ell\}_{\ell=\ell_{\min}}^{\ell=\ell_{\max}}$ and use independent inverse gamma distributions for the proposal:

$$g(\boldsymbol{D}; \theta) = \prod_{\ell=\ell_{\min}}^{\ell=\ell_{\max}} i\Gamma(D_\ell; \alpha_\ell, \beta_\ell).$$

Hence, we must adapt a vector $\theta = \{\alpha_\ell, \beta_\ell\}_{\ell=\ell_{\min}}^{\ell=\ell_{\max}}$ of $2(\ell_{\max} - \ell_{\min} + 1)$ parameters. As a starting point at Step (i), we use

$$\alpha_\ell = \frac{(2\ell+1)}{2} f_{\mathrm{sky}} - 1, \qquad \beta_\ell = \frac{(2\ell+1)}{2} f_{\mathrm{sky}} D_\ell^{\mathrm{ML}},$$

where D_ℓ^{ML} is the maximum-likelihood (ML) estimate of the angular spectrum. At Step (iii), parameters α_ℓ and β_ℓ are re-estimated at their ML values (see the Appendix).

The target density $\pi(\boldsymbol{D})$ is the posterior distribution of $\boldsymbol{D}$ when the prior distribution of $\boldsymbol{D}$ is flat. Hence, it is proportional to the likelihood.

In the two examples presented below, this iterative algorithm reached a perplexity above 0.6 after the first step of 50 k samples, and a 500 k sample set was produced during the final sampling phase.

3.3 Synthetic map

We first describe the results of adaptive importance sampling runs on a synthetic CMB map. The map is prepared at resolution

[1] This situation is representative of CMB data sets from satellites such as *WMAP* and *Planck*.

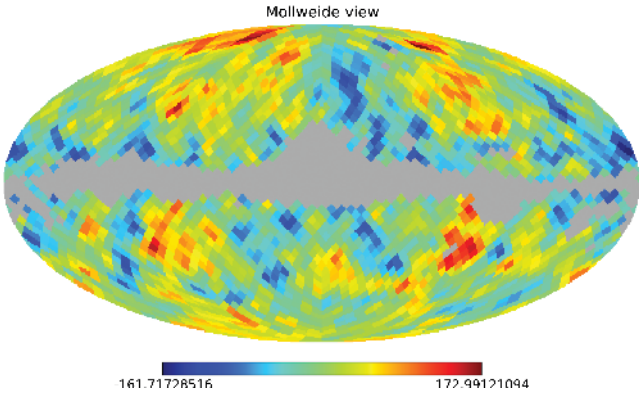

Figure 2. The synthetic CMB map used at Section 3.3.

$N_{\rm side} = 16$ from the *WMAP*5 best-fitting power spectrum (Dunkley et al. 2009) using HEALPIX (Górski et al. 2005). To avoid aliasing small-scale power into large-scale modes, the map is smoothed prior to down-sampling using a synthetic window function w_ℓ:

$$W_\ell = \begin{cases} 1 & 0 \leq \ell \leq 40 \\ \dfrac{1+\cos[(\ell-40)\pi/8]}{2} & 40 \leq \ell \leq 48\,, \\ 0 & 48 \leq \ell \end{cases} \tag{11}$$

which is used to explore the posterior of C_ℓ up to $\ell = 40$.

The posterior of the power spectrum is given by the likelihood described in equation (4), with a flat prior. The Galactic region is excluded using the *WMAP*5 mask, hiding 18 per cent of the sky. The map is shown in Fig. 2. A 1 μK pixel^{-1} noise is taken into account in the likelihood, but no noise is actually added to the map. This level should not affect our results as it is much lower than $\Omega_{\rm pix} C_{40}$ (see Fig. 1). We build a sample of the posterior of the masked map using the adaptive importance sampling algorithm described above. We only explore $\ell = 2$ to 40, the other modes ($\ell = 0$, 1 and $41 \leq \ell \leq 48$) being held constant to the ML estimate.

The initial proposal is given by the product of independent inverse gamma distributions, as described in Section 3.2, centred at $D_\ell^{\rm ML}$ with a width given by an effective sky coverage equal to $f_{\rm sky} \times 0.98$ to ensure that the initial proposal is wide enough.

Only one adaptation step was needed. It took about 58 min on 80 2 GHz CPUs to produce the first 50 k samples (about 6 s for each likelihood evaluation, taking into account all overheads). The final 500 k sample run took 6 h and 21 min on 120 2 GHz CPU (about 5.5 s for each likelihood evaluation, taking into account all overheads). The adaptive algorithm behaved very well: the first step reached $\mathcal{P} = 0.68$, while the second run hit $\mathcal{P} = 0.93$. This last run had an effective sample size ESS = 437 029, i.e. a ratio ESS$/N$ = 0.874.

Figs 3–5 give an overview of the results. First, looking at the 1D marginal distributions, Fig. 3 shows a few marginals (π_ℓ) and their best inverse gamma fits. The inverse gamma model is seen to account very well for both the tails and the mode of the distribution, in line with the high perplexity reached in the last iteration. This agreement validates a posteriori the adaptive approach. On this synthetic map, at least, the marginals closely follow an inverse gamma distribution.

The peaks of the marginals and an effective sky coverage at multipole ℓ, denoted as f_ℓ, are obtained by inverting

$$\alpha_\ell = \frac{(2\ell+1)}{2} f_\ell - 1, \tag{12}$$

$$\beta_\ell = \frac{(2\ell+1)}{2} f_\ell \left(W_\ell C_\ell^{\rm peak} + N_\ell \right). \tag{13}$$

Both quantities are shown in Fig. 4. The $C_\ell^{\rm peak}$ and $C_\ell^{\rm ML}$ discrepancy is small; it is below the per cent order, albeit with a few modes disagreeing by at most 3 per cent. The effective sky coverage, however, is quite different from $f_{\rm sky}$. Its behaviour indicates a transition between scales that are not affected significantly by the cut, and scales that are smaller than the cut, so that their deficit of modes is given by $f_{\rm sky}$. Our resolution is probably not good enough to reach this regime.

One would expect some discrepancy between the $C_\ell^{\rm peak}$ and the ML estimate. Indeed, since the cut induces correlation between scales, there is no reason for the peak of the posterior to be identical to the peak of the marginals in each direction. The small discrepancy can only be explained by a low level of correlation between the C_ℓs, so that the peak of the marginals is close to the joint peak. As a first estimate of the correlation, Fig. 5 shows the correlation matrix measured on our sample

$$[V]_{\ell,\ell'} \equiv {\rm Corr}\,(C_\ell, C_{\ell'}). \tag{14}$$

In this figure, the diagonal of the matrix is removed so as not to dominate the off-diagonal terms. Those exhibit a pattern below the 6 per cent level. Most of the correlation is located around $\ell = 12$, and the correlation seems to extend significantly for about six modes off the diagonal.

Several checks can be performed to assess the accuracy of this matrix. First, the ESS allows us to estimate the error on the matrix measurement to be of the order of 0.15 per cent, which is well below the observed correlation pattern. One can also measure the correlation matrix on the results of the first iteration of the adaptive algorithm, which provides us with an independent exploration of the posterior. The noise was much higher (with a level, according to the ESS of this run of about 0.6 per cent), but the pattern observed in Fig. 5 is easily recovered. Finally, we checked on a *full-sky* run that no correlation pattern is visible.

3.4 *WMAP*5 map

We perform a similar experiment using the *WMAP* map distributed along with the 5-year *WMAP*-likelihood code found on the Lambda web site.[2] The setting is slightly different, since the window function is a 9.°18 Gaussian beam, cutting much more high-frequency power than the window function (11) (see Fig. 1). Therefore, only the range $2 \leq \ell \leq 30$ is explored here, with the other multipole powers held constant at their ML values. As done in the *WMAP*-likelihood code, a 1 μK pixel^{-1} noise is added to the data and to the model. We take care of adding the specific noise realization used in the likelihood code. Indeed, with the beam used, the signal-to-noise ratio at $\ell = 30$ is only ∼14, and our tests have shown a small dependency of the value of the higher C_ℓs on the noise realization.

As in the previous run, only one adaptation step turns out to be needed. It took 32 min on 120 CPUS for 50 k samples, while the second and final run produced 500 k samples in 5 h and 19 min. The first iteration reached $\mathcal{P} = 0.48$, the second one $\mathcal{P} = 0.96$ and an effective sample size ESS = 457 600 (ESS$/N = 0.92$).

The results are generally similar to those reported in Section 3.3. We do not show more 1D marginal plots, but present the recovered $C_\ell^{\rm peak}$ and f_ℓ (Fig. 6), as well as the correlation matrix (Fig. 7). The $C_\ell^{\rm peak}$ and the ML estimates are somewhat similar to the *WMAP*5 power spectrum, with a small discrepancy also observed by Eriksen et al. (2007) using the Gibbs sampling and in Rudjord et al. (2009)

[2] http://lambda.gsfc.nasa.gov/

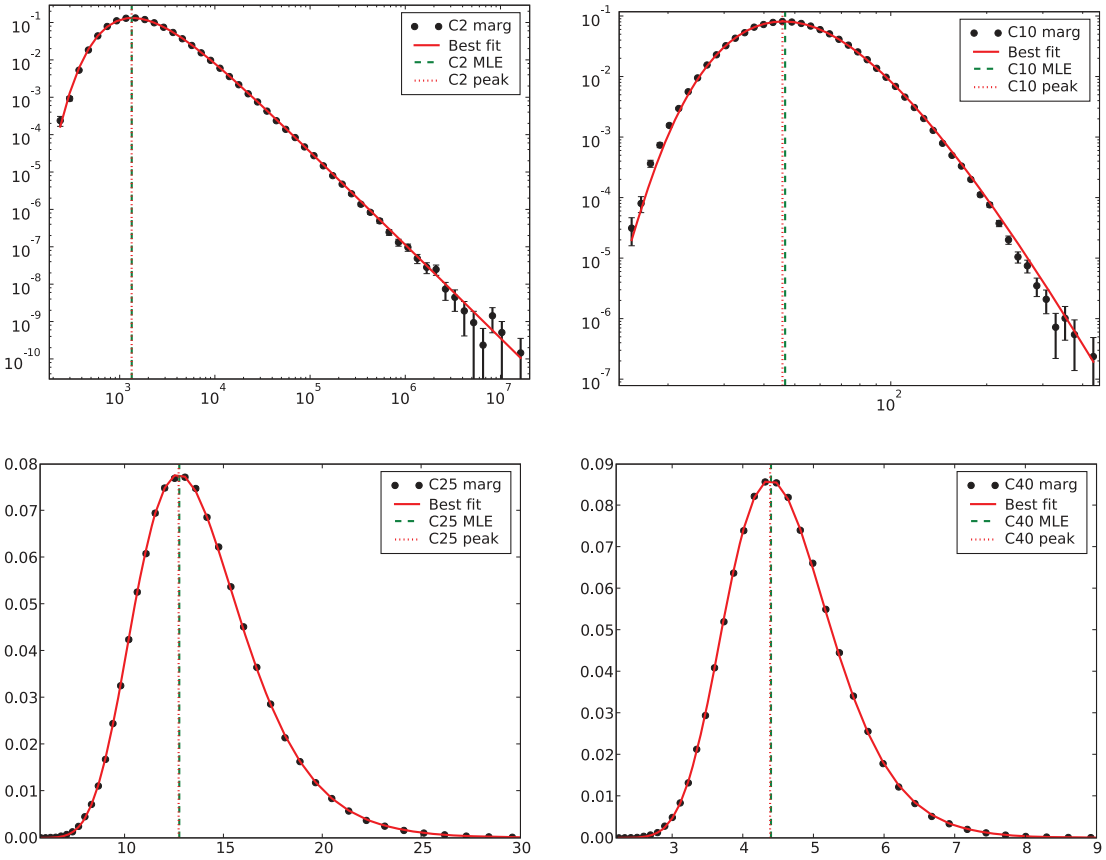

Figure 3. A few marginalized binned posteriors of the C_ℓ. The red line is the best inverse gamma approximation (including binning) obtained using the ML estimates, while the black dots are the binned marginal obtained on the 500 k sample. The red short dotted vertical line gives the location of the peak according to the approximation, and the green long dotted vertical line shows the $C_\ell^{\rm ML}$. Top plots are in log–log scale, while bottom plots are in linear scale to show the behaviour in the tail and at the peak of the marginals.

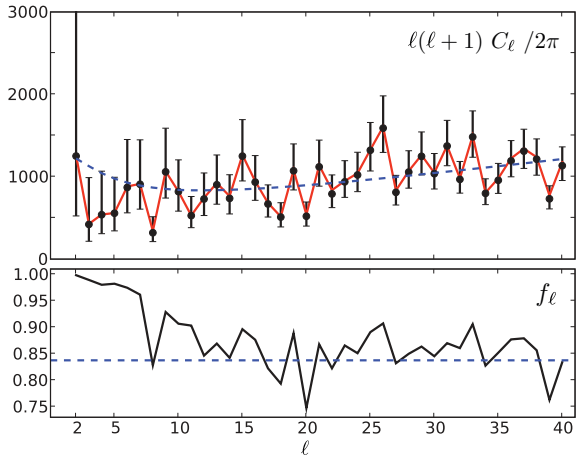

Figure 4. Top panel: angular spectra. Blue dashed line: the power spectrum used to synthesize the map; red: the ML estimate $C_\ell^{\rm ML}$; black dots: $C_\ell^{\rm peak}$. The error bars are 68 per cent limits obtained from the inverse gamma fits for the marginals. Bottom panel: sky coverage f_ℓ. Black line: effective coverage f_ℓ; the blue dashed line shows $f_{\rm sky} = N_{\rm mask}/N_{\rm pix}$.

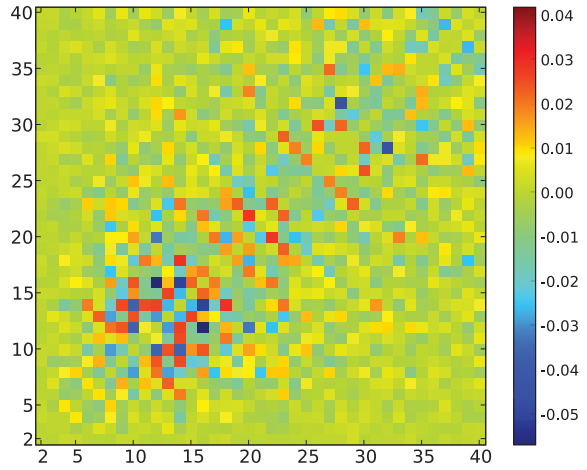

Figure 5. The correlation matrix **V** for C_ℓ (see equation 14) with the diagonal removed. Most of the correlation is located around $\ell = 12$ and extends only to a few neighbouring modes. The correlation is always below the 6 per cent level.

(zooming in their fig. 5). At any rate, the discrepancy is always within the C_ℓ error bars.

The effective coverage f_ℓ is similar to the one reported in Section 3.3, with a transition from 1 to $f_{\rm sky}$ but differs in some details, indicating that it is not only a function of the mask, but also of the actual data set.

Finally, Fig. 7 shows the correlation matrix. It exhibits structures similar to those in Fig. 5. As for the f_ℓ, the differences between Figs 7 and 5 indicate that the correlation matrix does not depend only on the mask.

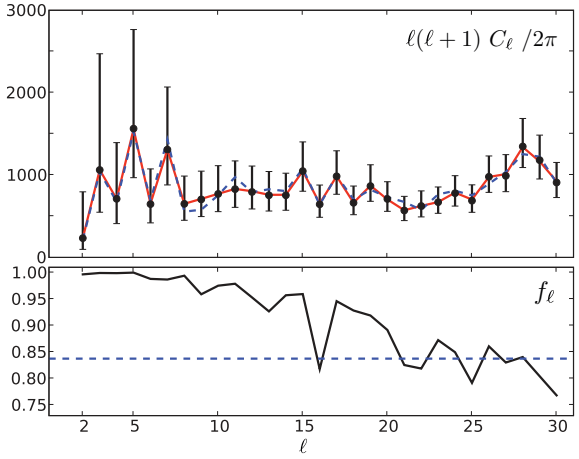

Figure 6. Same as Fig. 4 for the *WMAP*5 data set. The dashed blue top panel line now is the *WMAP* published spectrum.

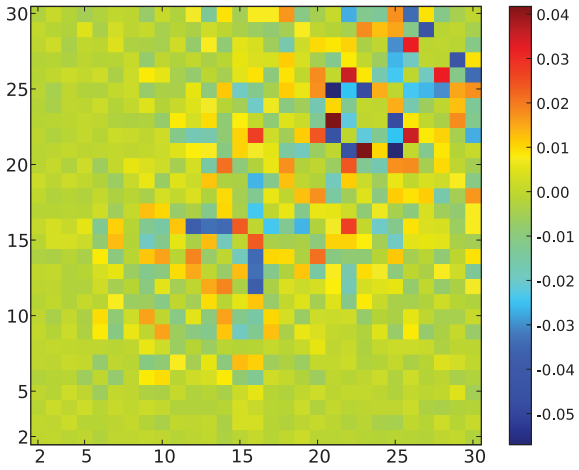

Figure 7. Same as Fig. 5 for the *WMAP*5 data set.

4 APPROXIMATING THE LOW-ℓ LIKELIHOOD

For both data sets considered in the previous section, the posterior distribution of the total angular spectrum D_ℓ revealed similar and striking features: the marginals are very well approximated by inverse gamma distributions, and there is a weak correlation between multipoles (below the 10 per cent level). Since we used a flat prior, these findings suggest that a copula approximation to the likelihood should be quite accurate (in addition to being fast, by design). This approach is somewhat similar to what has been proposed by Bond et al. (2000) and implemented at low ℓ in Rudjord et al. (2009) and at high ℓ in Hamimeche & Lewis (2008). It differs in that instead of offset lognormal (as in Bond et al. 2000) or spline approximation (Rudjord et al. 2009) we use inverse gamma cumulative functions for Gaussianization. For the high-ℓ approximation, Hamimeche & Lewis (2008) use an analytic reparametrization with an approximate covariance matrix, computed using a fiducial C_ℓ.

4.1 Copula approximation

A good approximation formula must at least reproduce the inverse gamma marginals, and the observed level of correlation. A generic way of building multivariate distributions with specified marginals and some correlation is provided by *copula models* (Sklar 1959).

The copula model. Denote $\mathcal{N}^{(d)}(\cdot;\mu,\mathbf{M})$ as the d-variate Gaussian density with mean μ and covariance matrix $\mathbf{M}$. Consider a set of zero-mean unit-variance Gaussian variables G_ℓ with density $\mathcal{N}^{(d)}(G_\ell;0,\mathbf{M}_G)$, where $\mathbf{M}_\mathrm{G}$ has only 1s on the diagonal and possibly non-zero off-diagonal terms. Consider those transformed variables $D_\ell = D_\ell(G_\ell)$ which have an inverse gamma distribution with parameters α_ℓ and β_ℓ, i.e. G_ℓ and D_ℓ are related by

$$\mathcal{N}(G_\ell;0,1)dG_\ell = i\Gamma(D_\ell;\alpha_\ell,\beta_\ell)dD_\ell. \tag{15}$$

The distribution of D_ℓ is then easily seen to be

$$\tilde{\pi}(D_\ell) \equiv \prod_k i\Gamma(D_k;\alpha_k,\beta_k)\,\frac{\mathcal{N}^{(d)}(G_\ell;0,\mathbf{M}_\mathrm{G})}{\prod_k \mathcal{N}^{(1)}(G_k;0,1)}. \tag{16}$$

Distribution (16) is called the *copula approximation*. It belongs to a parametric model with $2d + d(d-1)/2$ parameters: each of the d multipoles requires a pair (α_ℓ,β_ℓ) for the marginal distribution and the correlation matrix $\mathbf{M}_\mathrm{G}$ depends on $d(d-1)/2$ free parameters.

Two properties. Probability distributions of the form (16) enjoy two nice properties which readily follow from their construction. First, the marginal distribution of each D_ℓ remains an inverse gamma regardless of the correlation level (which is independently controlled by the matrix $\mathbf{M}_\mathrm{G}$). Secondly, marginalization over any subset of D_ℓ is readily achieved by removing the corresponding rows and columns of matrix $\mathbf{M}_\mathrm{G}$.

Gaussianization. Evaluating the copula density (16) requires explicit Gaussianization, i.e. mapping D_ℓ to G_ℓ. This is easy, since relation (15) implies that

$$G_\ell \equiv \mathrm{c}N^{-1}[\mathrm{c}i\Gamma(D_\ell;\alpha_\ell,\beta_\ell)], \tag{17}$$

where $\mathrm{c}i\Gamma(\cdot;\alpha,\beta)$ denotes the cumulative distribution function (CDF) of the inverse gamma distribution and $\mathrm{c}N^{-1}$ is the inverse CDF (or quantile function) of the standard normal distribution, sometimes called the *probit* function. The former is

$$\mathrm{c}i\Gamma(x;\alpha,\beta) \equiv \int_0^x i\Gamma(t;\alpha,\beta)\,\mathrm{d}t = \Gamma(\alpha,\beta/x)\,/\,\Gamma(\alpha),$$

while the latter, if missing from a statistical library, can be computed as $\mathrm{c}N(x)^{-1} = \sqrt{2}\mathrm{erf}^{-1}(2x-1)$ with $\mathrm{erf}(y) = \frac{2}{\sqrt{\pi}}\int_0^y \exp(-t^2)\mathrm{d}t$.

Speed. Copula evaluation is very fast. Using a custom code to compute the inverse error function, and the free GSL library[3] for the gamma and cumulative gamma distribution, we can compute about 18 000 samples per second, while the pixel-based likelihood needs about 5.5 s per sample on the same computer within the same setting (i.e. same overheads). Moreover, one can also sample directly from the copula by first drawing G_ℓ under to their multivariate Gaussian distribution and then invert equation (17) to get the D_ℓ values.

Learning the copula model. Learning the $2d + d(d-1)/2$ parameters of a copula model from (importance) samples of D_ℓ is straightforward. In a first step, one estimates for each ℓ, the inverse gamma parameters (α_ℓ,β_ℓ) by ML (see Appendix A). In a second step, the samples are Gaussianized via equation (17) using the estimated values of (α_ℓ,β_ℓ). Finally, matrix $\mathbf{M}_\mathrm{G}$ is plainly estimated as the sample correlation matrix of the Gaussianized samples.

Significance of correlation. Given a copula model $\tilde{\pi}$ with correlation matrix $\mathbf{M}_\mathrm{G}$, there is a simpler copula model with the same marginals

[3] http://www.gnu.org/software/gsl/

but without correlation, i.e. with $\mathbf{M}_G = \mathbf{I}$. This model is denoted $\tilde{\pi}_0$ and called the *uncorrelated model* which, of course, is not as accurate as $\tilde{\pi}$. Since $\tilde{\pi}_0$ and $\tilde{\pi}$ are Gaussian distributions, the loss can be quantified exactly thanks to a Pythagorean property of the Kullback-Leibler divergence which yields

$$K(\pi|\tilde{\pi}_0) = K(\pi|\tilde{\pi}) + K(\tilde{\pi}|\tilde{\pi}_0). \tag{18}$$

It shows that the mismatch $K(\pi|\tilde{\pi}_0)$ of the uncorrelated approximation to the posterior is larger than the mismatch $K(\pi|\tilde{\pi})$ of the regular copula by a positive term $K(\tilde{\pi}|\tilde{\pi}_0)$. This term can be computed in a closed form:

$$K(\tilde{\pi}|\tilde{\pi}_0) = -\frac{1}{2}\log\det\mathbf{M}_G, \tag{19}$$

which is positive unless $\mathbf{M}_G = \mathbf{I}$ and readily gives a measure of the price to pay for ignoring correlation.

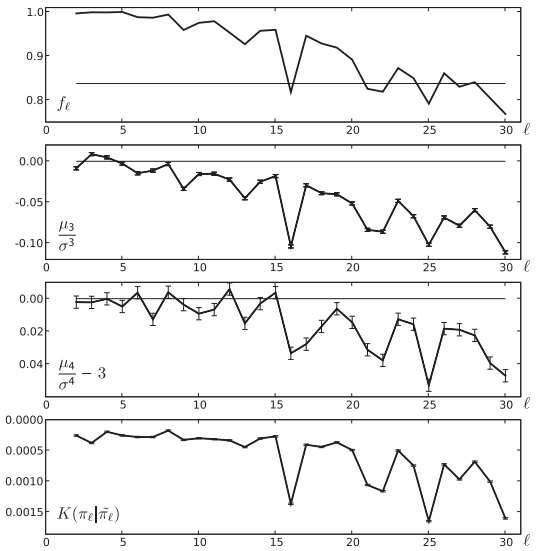

Figure 8. f_ℓ, cumulants and the Kullback divergence of G_ℓ exhibit some correlation. From top to bottom panels, f_ℓ (and $f_{\rm sky}$), skewness, kurtosis and the Kullback divergence between the marginals and standard normal. Note that the Kullback divergence is estimated from a histogram. The last two panels have their ordinates downwards to better show the correlation. Error bars are measured on 500 Gaussian simulations of size ESS (= 457 600).

4.2 Validation: first results

We first look at some self-consistency results when learning a copula model from the importance samples obtained from the *WMAP* data set discussed in Section 3.4.

High perplexity. The first important thing to report is that, on the perplexity scale, the copula approximation is remarkably good: we reach $\mathcal{P}(\pi|\tilde{\pi}) = 0.99$ on a 500 k simulation sample using estimates of the $C_\ell^{\rm peak}$, f_ℓ and $\mathbf{M}_G$ obtained on the same sample. As a simple cross-validation test, we split the 500 k sample into two subsets of equal size, re-estimate the copula parameters on the first subset and compute the perplexity using the second subset. We find a negligible decrease in perplexity of about 5×10^{-4}.

Thus, the copula approximation appears to work extremely well on this data set. Still, one should look further than a single number. This section looks into more details of the approximation.

Gaussianization. Even though the marginals were found to be well approximated by inverse gamma distributions, the Gaussianized importance samples may show some small hints of non-Gaussianity. Indeed, for each G_ℓ, we computed the skewness, the kurtosis and the Kullback divergence to a standard Gaussian. See Fig. 8 for the *WMAP* data set (a similar plot can be obtained on the other data set). The plot shows a small deviation from Gaussianity showing that the target densities are not *exactly* inverse gamma distributed. In addition, those non-Gaussian indicators degrade with ℓ and are correlated with f_ℓ. Since the latter measures deviation from the full-sky case, this is not unexpected.

Correlation matrices. By design, the copula correctly predicts the correlation matrix of the *Gaussianized* variables but it is not necessarily accurate as a predictor of the correlation matrix $\mathbf{V}$ of C_ℓ. Here, we check that $\mathbf{V}$ is well predicted by the covariance matrix of the copula model, denoted as $\tilde{\mathbf{V}}$. Matrix $\mathbf{V}$ is estimated as described before (based on an importance sample); matrix $\tilde{\mathbf{V}}$ is obtained from the same importance samples, reweighed by $\tilde{\pi}/\pi$. The results are displayed in Fig. 9 and show an excellent agreement, with small and evenly distributed errors.

4.3 Perplexities

We briefly report on the relative perplexity and the Kullback divergence between the posterior and its approximations on the *WMAP*5 data set. Some results are reported in Table 1. Since the Gaussianized variables were found to be weakly correlated, it may be tempting simply to ignore this correlation and to resort to the *uncorrelated* approximation $\tilde{\pi}_0$ defined in Section 4.1. In this case, the fit is slightly degraded: we measure $\mathcal{P}(\pi|\tilde{\pi}_0) = 0.97$, in line with the perplexity obtained after the last step of the adaptive importance run

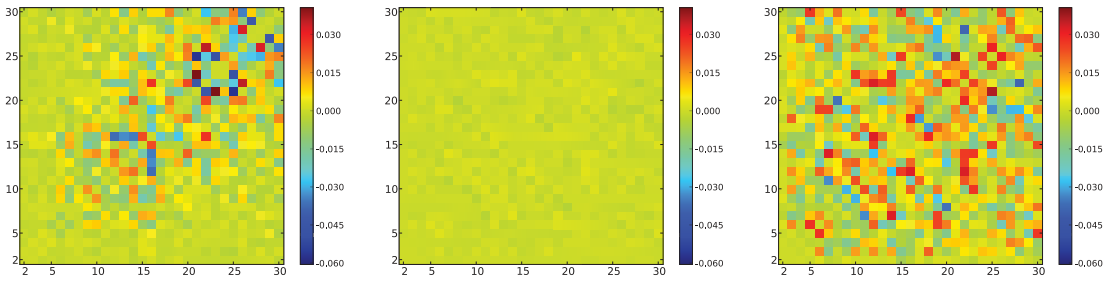

Figure 9. Left-hand panel: $\mathbf{V}$; centre panel: $\mathbf{V} - \tilde{\mathbf{V}}$; right-hand panel: $10 \times (\mathbf{V} - \tilde{\mathbf{V}})$. All panels share the same colour scale. Matrices $\mathbf{V}$ and $\tilde{\mathbf{V}}$ have been obtained on the same importance sample with appropriate weights.

Table 1. Perplexities (see the text).

Approximation	Perplexity	Kullback ($\times 10^{-3}$)
Copula $\tilde{\pi}$	0.991	8.6
Uncorrelated copula $\tilde{\pi}_0$	0.965	35.2
Uncorrelated last run	0.956	45.0
Naive $\tilde{\pi}_{\rm naive}$	0.779	249.6
Lognormal	0.191	1655.3

($\mathcal{P} = 0.96$; Section 3.4) showing that the determination of $C_\ell^{\rm peak}$ and f_ℓ is only marginally improved by the 500 k simulation. The contribution of correlation to the quality of the fit is given on the Kullback scale by the Pythagorean decomposition (18). Numerical evaluation by Monte Carlo integration gives, term to term,

$$K(\pi|\tilde{\pi}_0) = 35.18 \times 10^{-3} \approx 8.61 \times 10^{-3} + 27.3 \times 10^{-3}. \quad (20)$$

This is only an approximate equality because of MC errors. The last term was also evaluated using equation (19), yielding 27.5×10^{-3}. These values show that correlation accounts for most part of the mismatch in the sense that $K(\pi|\tilde{\pi}) \approx \frac{1}{3} K(\tilde{\pi}|\tilde{\pi}_0)$.

Those results can be compared to the naive approximation used as the initial proposal in our adaptive importance sampling runs, i.e. the copula approximation $\pi_{\rm naive}$ with $C_\ell^{\rm ML}$, $f_{\rm sky}$ and ignoring the correlation. It gives a perplexity of $\mathcal{P}(\pi|\tilde{\pi}_{\rm naive}) = 0.76$ corresponding to a huge increase in the Kullback divergence.

Finally, we compute, as a comparison baseline, the perplexity of the classical offset lognormal approximation (Bond et al. 2000). The estimation of the curvature at the peak is easily derived from f_ℓ. The perplexity goes down to $\mathcal{P} = 0.2$ for that approximation.

4.4 Validation: pseudo-cosmological parameters

We now compare several likelihood functions via their impact on estimation of (pseudo-) cosmological parameters from *WMAP* data. Since only the low-ℓ part of the spectrum is considered, only a few cosmological parameters can be fitted. We choose to perform our comparisons using a simple model with only two parameters, amplitude and spectral index, i.e. we consider

$$\tilde{C}_\ell \equiv C_\ell^{\rm ref} \times A \left(\frac{\ell}{\ell_0}\right)^n, \quad (21)$$

where $C_\ell^{\rm ref}$ is a reference angular spectrum (here the *WMAP*1 best-fitting spectrum), and the relative amplitude A and the relative spectral index n are our pseudo-cosmological parameters. The reference power spectrum being a fit on a broader range of multipoles, the posterior of (A, n) is not centred at $(1, 0)$.

Fig. 10 shows the 1σ, 2σ and 3σ contours and the peak position for different likelihood approximations. The top panel presents a comparison between the *WMAP*5 likelihood code, used in both pixel-based and Gibbs modes (Dunkley et al. 2009), and copula approximations with or without correlations (i.e. $\tilde{\pi}$ and $\tilde{\pi}_0$). They all appear to be in remarkably good agreement. The small discrepancies in the contour curves (which are smaller than the grid step size) are much smaller than the width of the mode. The peaks of the copula approximations and of the Gibbs approximation are very slightly displaced compared to the official *WMAP*5 results, at a distance of the order of the step size of the grid on which likelihoods are evaluated. The bottom panel presents a comparison with the lognormal approximation described in the previous chapter. As expected, the quality of that last approximation is poor, with a deviation of the best fit (A, n) of the order of $\sigma/4$. None the less, the areas of the 1σ, 2σ and 3σ regions are similar, probably because these areas are mostly controlled by the values of f_ℓ.

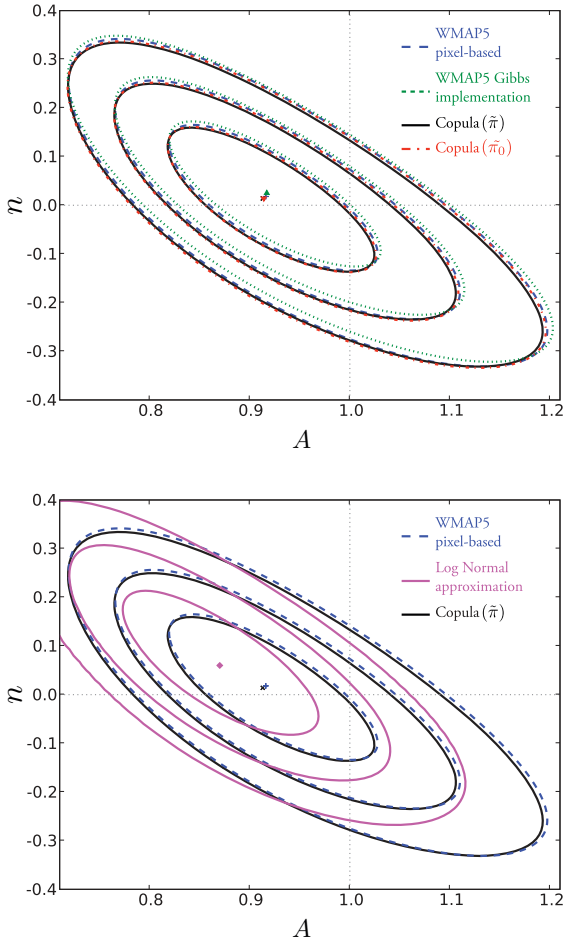

Figure 10. Posterior distribution for (A, n) using different likelihood approximations. Both panels: the dashed blue line shows official *WMAP*5 likelihood code and the black solid line shows the copula approximation $\tilde{\pi}$. Top panel: green dotted line is the Gibbs implementation included in the official code, the red dash–dotted line is the copula approximation ignoring correlations, $\tilde{\pi}_0$. Bottom panel: solid magenta line is the lognormal approximation. The coloured symbols mark the peak of each posterior.

5 CONCLUSION

Using an adaptive importance sampling algorithm, we explored the low-ℓ posterior of partially observed CMB maps, both synthetic and real. From this exploration, we built a copula-based approximation for that posterior distribution. Numerical evaluation of that approximation is much faster than the pixel-based computation. We showed that the approximation is very close to the actual posterior with an accuracy which is probably sufficient for most cosmological applications. For example, on a simple two-parameter pseudo-cosmological model, we found a discrepancy which is negligible with respect to the width of the posterior mode (Fig. 10).

The copula approximation uses two ingredients: a model of marginal distributions and a correlation matrix. The marginals are mostly distributed as inverse gammas, as in the full-sky case, but with different parameters. Maybe surprisingly, the correlations between (Gaussianized) multipoles are found to be quite low (<10 per cent). Ignoring them in the toy cosmological model

illustrated by Fig. 10 does not significantly change the posterior. However, when considering the full joint distribution of the multipoles (as opposed to its *projection* on to the two-parameter toy model), the correlation is significant: the Kullback divergence from the true posterior to its copula approximation quadruples if the correlation is left out. In both cases, however, the Kullback divergence remains small.

The main limitation of the proposed approximation is that it requires an exploration of the posterior to measure the parameters of the approximation. We used an adaptive importance sampling algorithm, but an MCMC algorithm, Gibbs-based (Eriksen et al. 2004; Jewell et al. 2004; Wandelt et al. 2004) or Hybrid MC-based (Taylor et al. 2008) can also be used. Both methods exhibit good scaling properties thanks to a smart rewriting of the posterior and could, if convergence is well controlled, provide estimates at higher ℓ. Indeed, a very recent work, published at the time we were finishing this paper, follows a similar path and demonstrates a Gaussianization technique based on splines rather than on inverse gamma models (Rudjord et al. 2009).

Another approach would be to determine the parameters of the marginals directly from the likelihood, without resorting to a sampling-based exploration. We are currently working on an analytical derivation of the approximation which would make it possible to build an approximation valid for higher ℓ at low computational cost. Being able to reach smaller scales is also important to explore the transition between low-ℓ estimates and high-ℓ ones. Indeed, at very small scales, the problem becomes intractable and requires the use of asymptotic approximations to the likelihood (Percival & Brown 2006; Smith, Challinor & Rocha 2006).

Finally, it is not clear yet whether the same kind of approximation can be built for polarized fields. In the temperature case addressed here, we took advantage of a low correlation situation, thanks to a high signal-to-noise ratio and relatively small masked area. Polarized observations will be noisier, and it remains to be seen if copula approximations are up to the task. This is the subject of current investigations.

ACKNOWLEDGMENTS

We thank J. Dunkley for her detailed description of the large-scale map used in the *WMAP*5 likelihood. The authors were greatly helped by the comments and remarks from F. Bouchet, H. K. Eriksen, members of the ECOSSTAT ANR project and the Planck CTP working group. The ANR grant ECOSSTAT (ANR-05-BLAN-0283-04) provided financial support for part of this work. We acknowledge the use of the HEALPIX package.[4] We thank the Planck HFI Data Processing Center for the use of its computing resources (Magique).

REFERENCES

Bond J. R., Efstathiou G., 1987, MNRAS, 226, 655
Bond J. R., Jaffe A. H., Knox L., 2000, ApJ, 533, 19
Cappé O., Douc R., Guillin A., Marin J.-M., Robert C. P., 2008, Statistics and Computing, 18, 447
Dunkley J. et al., 2009, ApJS, 180, 306
Efstathiou G., 2006, MNRAS, 370, 343
Efstathiou G., Lawrence C. R., Tauber J., 2005, European Space Agency, ESA-SCI(2005)-1
Eriksen H. K. et al., 2007, ApJ, 656, 641
Eriksen H. K. et al., 2004, ApJS, 155, 227
Gorski K. M., 1994, ApJ, 430, L85
Górski K. M., Hivon E., Banday A. J., Wandelt B. D., Hansen F. K., Reinecke M., Bartelmann M., 2005, ApJ, 622, 759
Hamimeche S., Lewis A., 2008, Phys. Rev. D, 77, 103013
Hinshaw G. et al., 2007, ApJS, 170, 288
Hinshaw G. et al., 2009, ApJS, 180, 225
Jewell J., Levin S., Anderson C. H., 2004, ApJ, 609, 1
Lepage G., 1978, J. Comput. Phys., 27, 192
Percival W. J., Brown M. L., 2006, MNRAS, 372, 1104
Rosenthal J. S., 2000, Far East J. Theor. Stat., 4, 207
Rudjord Ø., Groeneboom N. E., Eriksen H. K., Huey G., Górski K. M., Jewell J. B., 2009, ApJ, 692, 1669
Sklar A., 1959, Publ. Inst. Stat. Univ. Paris, 8, 229
Smith S., Challinor A., Rocha G., 2006, Phys. Rev. D, 73, 023517
Taylor J. F., Ashdown M. A. J., Hobson M. P., 2008, MNRAS, 389, 1284
Tegmark M., 1997, Phys. Rev. D, 55, 5895
Wandelt B. D., Larson D. L., Lakshminarayanan A., 2004, Phys. Rev. D, 70, 083511
Wraith D., Kilbinger M., Benabed K., Cappé O., Cardoso J.-F., Fort G., Prunet S., Robert C. P., 2009, Phys. Rev. D, 80, 023507

APPENDIX A: ML ESTIMATION OF INVERSE GAMMA PARAMETERS

The log-likelihood $\log \mathcal{L}(\alpha, \beta)$ for a sample of N independent realizations X_i under an inverse gamma density is

$$\log \mathcal{L} = \sum_i^N \left[\alpha \log \beta - \log \Gamma(\alpha) - (\alpha + 1) \log X_i - \frac{\beta}{X_i} \right]$$

as seen from equation (3). The ML estimate for (α, β) is the solution of $\frac{\partial \log \mathcal{L}}{\partial \alpha} = 0$ and $\frac{\partial \log \mathcal{L}}{\partial \beta} = 0$ leading to the two estimating equations:

$$\log \beta - \psi(\alpha) = \frac{1}{N} \sum_i^N \log X_i, \qquad \frac{\alpha}{\beta} - = \frac{1}{N} \sum_i^N \frac{1}{X_i},$$

where $\psi(u)$ is the log-derivative of the gamma function, also known as the digamma function. Using the last equation to express β in terms of α, the ML estimate can be obtained by solving

$$\log \alpha - \psi(\alpha) = \frac{1}{N} \sum_i^N \log X_i - \log \left(\frac{1}{N} \sum_i^N \frac{1}{X_i} \right). \tag{A1}$$

This is quickly done numerically in a few steps of a Newton algorithm; both the digamma function and its derivative being available in the GSL package.

[4] http://healpix.jpl.nasa.gov

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