arXiv:1108.0600v1 [astro-ph.IM] 2 Aug 2011

Improving stochastic estimates with inference methods: calculating matrix diagonals

Marco Selig, Niels Oppermann, and Torsten A. Enßlin
Max-Planck-Institut für Astrophysik, Karl-Schwarzschild-Str. 1, 85741 Garching, Germany

(Dated: August 3, 2011)

Estimating the diagonal entries of a matrix, that is not directly accessible but only available as a linear operator in form of a computer routine, is a common necessity in many computational applications, especially in statistical inference. Here, methods of statistical inference itself are used to improve the accuracy and/or the computational costs of matrix *probing* methods to estimate matrix diagonals. In particular, the generalized Wiener filter methodology, as developed within information field theory, is shown to significantly improve estimates based on only a few sampling probes, in cases in which some form of continuity of the solution can be assumed. The strength, length scale and precise functional form of the exploited autocorrelation function of the matrix diagonal is determined from the probes themselves. The developed algorithm is successfully applied to mock and real world problems. These performance tests show that the method pays off best in situations where a matrix diagonal has to be calculated from only a small number of computationally expensive probes.

PACS numbers: 02.50.Tt,02.30.Zz,11.10.-z,89.70.-a,89.70.Eg
Keywords: Inference methods, Inverse problems, Field theory, Information and communication theory, Computational complexity

I. INTRODUCTION

A. Estimation and inference

Many computational problems are approached by stochastic methods. Numerical integrals over very high dimensional spaces using regular grids are prohibited by the exploding computational costs, and are often approximated by the summation of suitably constructed stochastic samples. For example, the extraction of the diagonal entries of a matrix, that is only available as a linear operator on vectors, would require that this operator is applied once to each basis vector of the vector space. For the high dimensional image spaces one encounters in signal reconstruction problems, this is most often computationally too expensive. Therefore matrix or operator probing methods have been developed, in which the operator is probed by a number of random vectors, from which a stochastic estimate of the matrix diagonal can be obtained. The estimate will have a stochastic error, that can be diminished by enlarging the probing sample size. This of course also increases the computational costs.

Thus, a calculation problem has been replaced by a numerical experiment, or more precisely with a measurement which has many similarities with physical measurements. The outcome of this numerical measurement is therefore subject to uncertainties. The uncertainties can be reduced by repeating independent measurements and averaging the results. And – this is the main point of this paper – the averaging procedure can be improved by inference methods which exploit additional knowledge on the solution. For example, just the knowledge that the matrix diagonal should exhibit some level of smoothness on some (not necessarily known) spatial scales will turn out to be sufficient to improve the estimate precision by a factor of a few for a given budget of computational resources.

The analogy between stochastic estimation and signal inference might become more convincing for estimation problems for which the computational operations required by a probing estimator can be separated into prohibitively expensive ones (i.e. applying the linear operator to a matrix) and relatively cheap ones (i.e calculating some average of the probes). The expensive operations are then the measurement, and the data they deliver carry information on the signal we are interested in (the matrix diagonal) as well as noise contamination due to aspects of the matrix that we are not interested in (the impact of the non diagonal terms on the matrix-vector-product). Although in principle these two contributions to the data could be identified and separated, the computational costs for this are very high. Thus we should admit that due to our limited computational resources we do not know how a given datum (the result of the matrix-vector operation) is to be separated into signal and noise, because this would involve breaking down the matrix explicitly. However, for a sufficiently large set of suitably constructed measurements (i.e. using appropriate random vectors as operator probes) the signal part in the data will always be the same where the noise part will be independent and of zero mean. Thus, a suitably constructed averaging scheme will reveal the signal with larger and larger accuracy with increasing data size.

Only two questions remain. What is the optimal averaging scheme, that exploits, in addition to the data, any prior knowledge we have on the problem? Additionally what are its computational costs? If a more sophisticated and better scheme exceeds the costs for the additional probing required to gain the same accuracy, it would not be worth implementing. Furthermore, one might worry that one gets into an infinite recursion of inference schemes within inference schemes. For example, an image reconstruction problem requires the calculation of a matrix diagonal. For this an inference scheme is used, which itself could again require the calculation of another matrix diagonal, and so on.[1]

Here, a pragmatic point of view is adopted, and higher order problems are ignored. It will be shown that more sophisticated inference schemes can significantly outperform simple probing and averaging schemes in terms of total computational costs to reach a given accuracy level. This is especially true in case the computational costs for the matrix operation are large.

B. Previous work

The search for estimators of properties of matrices led to the stochastic method of probing. For this purpose the matrices are multiplied with test vectors in a way which statistically projects out the property of interest, here, the matrix diagonal. A first proposal for such a probing method can be found in the work by Hutchinson (1989) [1, and references therein]. There the functionality and efficiency of probing for obtaining trace estimates has been proven.

Bekas et al. [2] extended the probing to Hadamard vectors (rows of the Hadamard matrix) to improve the estimation of diagonals of banded matrices. Those methods have been oriented to applications in density functional theory. Similar problems were approached by Tang and Saad [3] and references therein, with a focus on nonstochastic estimators.

The recent paper by Aune and Simpson [4] transfers the probing technique to the field of information theory, in particular to the calculation of log-likelihoods.

Finally in the extensive work of Rohde and Tsybakov [5] the noise corrupted observation of unknown matrix entries is investigated from a more mathematical point of view.

C. Structure of this work

First, in Sect. II, we highlight the importance of obtaining estimates for matrix diagonals, in particular of uncertainty covariance matrices in the framework of information theory. We discuss the problem of their calculation in high-dimensional cases.

We will review in Sect. III the frequentist approach called probing, which is completely general and can be applied to all kinds of matrices. In Sect. IV we use information field theory (IFT) to present a Bayesian estimate for matrix diagonals. Our proposal focuses on covariance matrices, which are positive and symmetric by definition, and in practice often have sparse off-diagonal entries or at least off-diagonal entries that are decaying with distance from the diagonal.

Subsequently, Sect. V is devoted to the verification and application of both methods, where we investigate simple mock examples as well as a real example: the uncertainty covariance matrix of the all sky Faraday depth derived from the NVSS catalog.

We conclude in Sect. VI.

II. PROBLEM OF MATRIX DIAGONALS

Linear operators are fundamental in any area of computation and thereby often expressed in their matrix representation.

In the field of information theory the covariance matrix of a quantity (which equals the inverse of the precision matrix) holds a key role. To stress this, let us consider a multi dimensional zero-mean Gaussian

$$\mathcal{G}(\varphi, X) = \frac{1}{\sqrt{\det\left[2\pi X\right]}} \exp\left(-\frac{1}{2}\varphi^{\intercal} X^{-1}\varphi\right) \tag{1}$$

with the covariance matrix $X = \langle \varphi\varphi^{\intercal}\rangle_{\mathcal{G}}$ where φ is a random field defined over some pixelized vector space and $\langle\,\cdot\,\rangle_{\mathcal{G}}$ denotes the expectation value weighted by this Gaussian. (In this matter 'pixel' is to be understood as a discretized coordinate which elsewhere may be referred to as 'grid point', 'bin' or 'voxel'.)

A diagonal entry of the covariance matrix is the squared standard deviation σ_i assigned to pixel i expressing the pixelwise uncertainty in φ,

$$\sigma_i^2 = \left\langle \varphi_i^2 \right\rangle_{\mathcal{G}} = X_{ii}. \tag{2}$$

A sophisticated and effective reconstruction tool is the generic filter [6] that we will review in the following. We provide this review in order to further emphasize the importance and problem of obtaining matrix diagonals for stochastic inference. Furthermore because this filter will form the basis of our proposed algorithm discussed in Sect. IV B.

[1] However, in practice, such an infinite recursion can easily be truncated. First, one can use the pure frequentist averaging method at some level of this loop, thereby truncating it. Second, although the dimension of the matrix diagonal appearing in an image reconstruction problem has as many entries as the image, due to potentially existing spatial symmetries (e.g. statistical translational and rotational invariance) and smoothness properties (the uncertainty map of an image usually has less structure than the image itself) the covariance structure of the diagonal entries has often a lower number of degrees of freedom. Thus, the computational complexity of the series of nested inference problems gets simpler and simpler to the point where a direct matrix operation is affordable and truncates the need for further recursion.

A. Generic filter

The generalized Wiener filter, as derived e.g. in [7] in a Bayesian framework, is for one thing based on a linear forward data model

$$d = R\,s + n, \tag{3}$$

where the data d is a sum of signal response $R\,s$ and noise n. In this scenario, the response is a linear operator that inherits all aspects of the signal detection, i.e. the detector's input-output relation (e.g. the detector's point spread function, survey coverage, etc.).

The generalized Wiener filter arises in case one can in addition assume a Gaussian distribution for the signal's prior and the signal-independent noise,

$$P(s) = \mathcal{G}(s, S), \tag{4}$$
$$P(n|s) = \mathcal{G}(n, N), \tag{5}$$

where S and N stand for the signal and noise covariance matrix respectively. Ergo the likelihood of the data given the signal becomes

$$P(d|s) = P(n|s) = \mathcal{G}(d - Rs, N). \tag{6}$$

The resulting filter formula, whose derivation is detailed in [6–8] and will therefore not be repeated in this work, is

$$m = \underbrace{\left(S^{-1} + R^{\intercal} N^{-1} R\right)^{-1}}_{D} \underbrace{\left(R^{\intercal} N^{-1} d\right)}_{j}, \tag{7}$$

where the map m is the Bayesian estimator for the signal, i.e. its posterior mean, D is referred to as *information propagator* and j as *information source*. The inverse problem of estimating the signals given the data leads to a Gaussian posterior,

$$P(s|d) = \mathcal{G}(s - m, D), \tag{8}$$

with with the mean m and covariance D which encodes the a posteriori signal uncertainty. Both, the signal and noise covariance needed for this, are here assumed to be known. A convenient description of these covariances is in terms of their power spectra, the spectra of the eigen values of these matrices.

In the following we decompose the signal covariance by $S = \sum_l C_l S_l$, where the S_l are the projections onto a suitable eigen basis (in our examples in Sect. V this will be spherical harmonics). An analogous decomposition exists for the inverse $S^{-1} = \sum_l C_l^{-1} S_l^{-1}$.

The signal's power spectrum might be unknown a priori, whereas the eigen basis can often be guessed from statistical symmetries (e.g. the spherical harmonics basis in case of a statistically isotropic distribution on the sphere). Thus, the spectral coefficients C_l allow for a parameterization of the covariance. In such applications without spectral knowledge, the generalized Wiener filter can be extended to a generic filter derived in [6]. The generic filter formulas are Eq. (7) complemented by a reconstruction rule for the power spectrum, i.e. for each spectral coefficient one calculates

$$C_l = \frac{1}{\varrho_l + 2\epsilon_l} \operatorname{tr}\left[\left(m m^{\intercal} + \delta_l D\right) S_l^{-1}\right] \tag{9}$$

where $\varrho_l = \operatorname{tr}\left[S_l^{-1}\right]$ are the degrees of freedom for each spectral band. A prior that is flat on a logarithmic scale has been assumed for the spectral coefficients in the derivation of this formula. The parameters (δ_l, ϵ_l) characterize the different filter options: Two specific forms are the *classical filter* for which one chooses $(\delta_l, \epsilon_l) = (0, 0)$ and the *critical filter* for which $(\delta_l, \epsilon_l) = (1, 0)$. The former can be derived from a 'classical' maximum a posteriori (MAP) approximation of the spectral uncertainty marginalized problem. The latter is called 'critical' because it exhibits (in contrast to the classical filter) only a marginal perception threshold. (For a filter with a perception threshold the signal to noise ratio of a spectral mode has to exceed a certain threshold in the data before the filter recognizes it at all.) There exists a critical line in the δ-ϵ-plane separating filters that fully suppress bands with insufficient spectral power from filters that do not. The critical filter resides exactly on this line while the classical filter is in the region with such a perception threshold.

All in all Eq. (7) and (9) provide an iterative scheme for the full inverse problem of signal reconstruction with unknown power spectrum, i.e. unknown correlation structure. The signal reconstruction benefits from this additional spectral information since it encodes internal structure of the signal. Furthermore, it can be shown, see [6, 8], that the posterior can usually be approximated by a Gaussian,

$$P(s|d) \approx \mathcal{G}(s - m, D), \tag{10}$$

where the map m is the expected signal given the data and $\sqrt{\operatorname{diag}[D]}$ the associated uncertainty map (to which the square root is to be applied pixelwise). Therefore, the resulting map is an approximative solution of the inference problem. This map is constructed to be close to the minimum mean square error (MMSE) estimate (averaged over the remaining uncertainties given the data).

In order to apply the critical or other generic filters we may need to calculate the trace of DS_l^{-1} in (9) in each iteration, and we have to evaluate the diagonal of D in order to interpret the reliability of our results. This motivates our ambition to develop faster and more accurate matrix probing schemes.

Generic filters are applied e.g. in [6–10].

B. Exact matrix diagonal

The diagonal of the uncertainty covariance D is a quantity of interest, but unfortunately not directly accessible in most cases. Its calculation involves complex matrix operations such as matrix inversion, see Eq. (7). Often the complete matrix is not known explicitly, only the matrix-vector-multiplication is available in form of a computer routine which reads in and returns a vector.

Calculating the diagonal of the matrix X of dimension r seems still possible using normalized unit vectors $e^{(k)}$ (with $e_i^{(k)} = \delta_{ik} \, \forall \, i,k \in \{1,\ldots,r\}$).

$$\mathrm{diag}\,[X] = \sum_k e^{(k)} * X \, e^{(k)}, \tag{11}$$

where $*$ denotes a componentwise product in the way that $(a*b)_i = a_i b_i \, \forall \, i \in \{1,\ldots,r\}$.

It is obvious that this 'true' diagonal is too expensive computationally because one needs to evaluate the matrix-vector-multiplication exactly r times looping through all unit vectors where the dimension r of the problem can be very high ($r \gg 1$). In addition each of those products alone can be expensive because it may invoke numerical inversion techniques, e.g. conjugate gradient [11], which is the case in most of the examples in Sect. V.

III. PROBING ESTIMATE

The question arises if one can choose another set of vectors instead of the full set of normalized unit vectors to speed up the computation. Independent and identically distributed (i.i.d.) random variables stored in a set of vectors $\{\xi\}$ (with sample size $|\{\xi\}| = A$) will work if they fulfill the property

$$\langle \xi_i \xi_j \rangle_{\{\xi\}} \xrightarrow{A\to\infty} \delta_{ij}. \tag{12}$$

Here the average $\langle \, \cdot \, \rangle_{\{\xi\}}$ stands for the arithmetic mean over a set $\{\xi\}$ and δ_{ij} for the Kronecker delta.

Two of many possible options are (i) equally probable values of ±1 for the components of ξ [1][2] or (ii) zero-mean Gaussian random numbers with unit variance. Both were originally developed for trace estimation. We will use (ii) in the following examples.

Regardless of the choice of the random vectors, the sample average

$$\langle \xi * X \, \xi \rangle_{\{\xi\}} \xrightarrow{A\to\infty} \mathrm{diag}\,[X] \tag{13}$$

over an infinite set will result in the 'true' diagonal, see App. A.

The average over a finite but sufficiently large set ($A < r < \infty$) will therefore give the probing estimator f of the matrix diagonal,

$$\mathrm{diag}\,[X] \approx \langle \xi * X \, \xi \rangle_{\{\xi\}} = f. \tag{14}$$

Given this estimator we obtain one for the trace by summing up all elements of f, as

$$\mathrm{tr}\,[X] \approx \langle \xi^{\mathsf{T}} X \, \xi \rangle_{\{\xi\}} = \sum_i f_i. \tag{15}$$

Since one wants to obtain an estimator in a finite period of time, one has to find an acceptable trade-off between the sample size A and the residual error, where the latter scales with $1/\sqrt{A}$ according to the law of large numbers. Aiming for a certain precision therefore requires a particular amount of computation time.

The estimator given by Eq. (14) is absolutely generic and applicable to a variety of matrices. Recent applications of it can be found in [2, 4, 9].

IV. BAYESIAN ESTIMATE

A. Forward model

Instead of doing a blind probing we now want to develop a Bayesian estimate which exploits additional knowledge of the problem to infer the matrix diagonal from a smaller set of samples. For this purpose we consider the sampling described by Eq. (14) as a linear forward model of a measurement process for the signal $\tilde{s} = \mathrm{diag}\,[X]$ we are interested in. (In order to avoid confusion we will mark with a tilde already introduced synonymous quantities that appear now in another context.)

For one sample, $a \in \{1,\ldots,A\}$, the measurement equation takes the form

$$\begin{aligned} \tilde{d}^{(a)} &= \xi^{(a)} * X \, \xi^{(a)} \\ &= \underbrace{\mathrm{diag}\left[\left(\xi_1^{(a)}\right)^2, \ldots, \left(\xi_r^{(a)}\right)^2\right]}_{\tilde{R}^{(a)}} \, \tilde{s} + \tilde{n}^{(a)}. \end{aligned} \tag{16}$$

For all samples it is

$$\begin{aligned} \tilde{d} &= \left(\tilde{d}^{(1)}, \ldots, \tilde{d}^{(A)}\right)^{\mathsf{T}} \\ &= \underbrace{\left(\tilde{R}^{(1)}, \cdots, \tilde{R}^{(A)}\right)^{\mathsf{T}}}_{\tilde{R}} \, \tilde{s} + \underbrace{\left(\tilde{n}^{(1)}, \ldots, \tilde{n}^{(A)}\right)^{\mathsf{T}}}_{\tilde{n}} \\ &= \tilde{R} \, \tilde{s} + \tilde{n}, \end{aligned} \tag{17}$$

where $\tilde{d}$ represents the 'measured' data, $\tilde{R}$ the signal response and $\tilde{n}$ the noise. The contributions from all off-diagonal matrix elements are considered to be noise, i.e.

$$\tilde{n}^{(a)} = \xi^{(a)} * \left(X - \mathrm{diag}\,[X_{11}, \ldots, X_{rr}]\right) \xi^{(a)}, \tag{18}$$

and they can be estimated using (17), once we have an estimator for the signal.

Note that if one chooses the random variables ξ to be ±1, firstly one does not have to evaluate normal variables

[2] In [2] a much more sophisticated choice, based on [1], is presented.

as originally pointed out by [1] and secondly all the response martices $R^{(a)}$ equal $\mathbb{1}$ and hence do not need to be treated separately for the different samples. This speeds up the algorithm and reduces the memory requirements.

B. Proposed algorithm

Our goal is to find an estimator for the matrix diagonal which is close to the MMSE, but still computationally affordable. This estimator has to account for our missing knowledge about the underlying correlation structure. Given these requirements the generic filter formulas are potentially an appropriate choice. Therefore, our proposed algorithm is based on this filter.

We start by probing the matrix as described in Sect. III and as a result obtain a first estimator f for our signal, i.e. the matrix diagonal. This additional information changes our state of knowledge about the matrix diagonal in the way that the assumed prior in (4) is not adequate anymore. However, after only a few samples, f will not yet be a good approximation for the diagonal entries of the matrix, since several entries may be considerably over- or underestimated. By contrast f provides already a good enough estimator for the trace, see (15). For that reason we can a priori expect the matrix diagonal $\tilde{s}$ to be distributed around some $\tilde{t}$ rather than around zero, where for all $i \in \{1, \dots, r\}$ we set $\tilde{t}_i = \sum_i f_i/r \approx \langle \mathrm{tr}\,[X] \rangle_{\{\xi\}} / \dim\,[X]$. Therefore the prior of the matrix diagonal is chosen to have a non-zero mean $\tilde{t}$,

$$P(\tilde{s}) = \mathcal{G}(\tilde{s} - \tilde{t}, \tilde{S}). \tag{19}$$

As a consequence the filter formulas (7) and (9) undergo a shift,

$$\tilde{m} = \tilde{D}\left(\tilde{R}^{\intercal}\tilde{N}^{-1}\tilde{d} + \tilde{S}^{-1}\tilde{t}\right), \tag{20}$$

$$\tilde{D} = \left(\tilde{S}^{-1} + \tilde{R}^{\intercal}\tilde{N}^{-1}\tilde{R}\right)^{-1}, \tag{21}$$

$$\tilde{C}_l = \frac{1}{\tilde{\varrho}_l + 2\tilde{\epsilon}_l}\,\mathrm{tr}\left[\left((\tilde{m} - \tilde{t})\,(\tilde{m} - \tilde{t})^{\intercal} + \tilde{\delta}_l\tilde{D}\right)\tilde{S}_l^{-1}\right]. \tag{22}$$

Furthermore the noise covariance, i.e. its required inverse, is unknown a priori and needs to be estimated for our algorithm. If we use the data model described in Sect. IV A, $\tilde{N}^{-1}$ can be approximated by the noise given the data and an estimator for the signal,

$$\tilde{n} = \tilde{d} - \tilde{R}\,\tilde{m}. \tag{23}$$

We simplify our calculation be using

$$\tilde{N}^{-1} = (\tilde{n}\tilde{n}^{\intercal})^{-1} \approx (\mathrm{diag}\,[\tilde{n} * \tilde{n}])^{-1}. \tag{24}$$

This is done in order to limit the computational effort. (A more correct treatment of unknown noise covariance matrices is addressed in [10].)

Equations (20) to (22) are solved iteratively in the following scheme:

1. Start with $\tilde{m}^{(\nu=0)} = f$.
2. Compute $\tilde{n}^{(\nu+1)}$ according to (23).
3. Compute $\tilde{C}_l^{(\nu+1)}$ according to (22), while ignoring $\tilde{t}$ and $\tilde{D}$ for $\nu = 0$.
4. Compute $\tilde{m}^{(\nu+1)}$ according to (20) using (21) and (24).
5. Repeat steps 2 to 4 until convergence.

As an initial guess for the power spectrum in step 3, we use an overestimation. This accelerates the convergence process as can be seen in the extreme limits: $\tilde{C}_l \to \infty : \tilde{m} \sim \tilde{R}^{-1}\tilde{d}$, whereas $\tilde{C}_l \to 0^+ : \tilde{m} \sim \tilde{t}$. I.e. a strong overestimate still gives a non trivial result for $\tilde{m}$, whereas an underestimate gives a trivial one.

Following Sect. II A we generally recommend the critical filter, since it does not exhibit a significant perception threshold. Nevertheless in the presented examples the correction term $\mathrm{tr}[\tilde{D}\tilde{S}_l^{-1}]$ does only marginally contribute to the accuracy and therefore the classical filter which does not require the calculation of this term has been applied in the following.

V. VERIFICATION & APPLICATION

A. Numerical experiments

To verify the proposed algorithm, we perform some numerical experiments that are posed on signals living on the sphere. The examples in this Sect. are represented by all sky HEALPix[3] maps with $N_{\mathrm{side}} = 8$, resulting in $r = 768$ pixels and in a maximal spectral index $l_{\max} = 23$ of the spherical harmonics basis in which we a priori assume our signal covariance to be diagonal due to a statistical isotropy of the signal.

1. Trivial case

At first we consider a trivial case where the matrix in question is given explicitly. To ensure that this matrix is covariant, i.e. it is positive and symmetric, we constructed it to be

$$X = \begin{pmatrix} X_{11} & -1 & & 0 \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ 0 & & -1 & X_{rr} \end{pmatrix}, \tag{25}$$

[3] See HEALPix homepage http://healpix.jpl.nasa.gov/

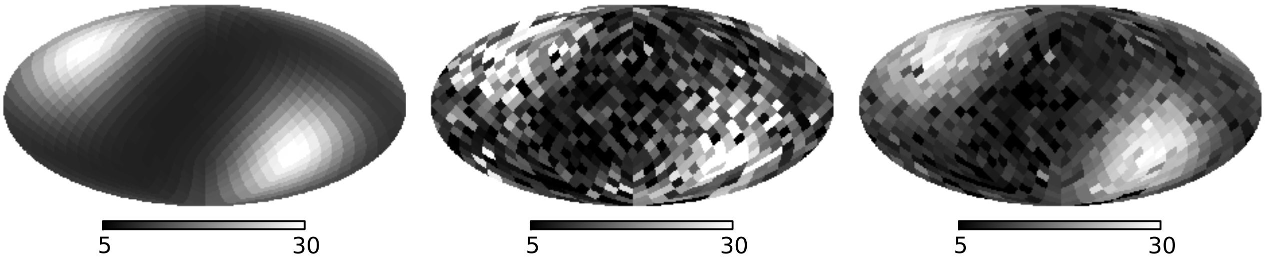

Figure 1: Result from the trivial case: (left) the exact matrix diagonal, (middle) the probing estimate and (right) the Baysian estimate started with four probes, both after around 0.3 seconds.

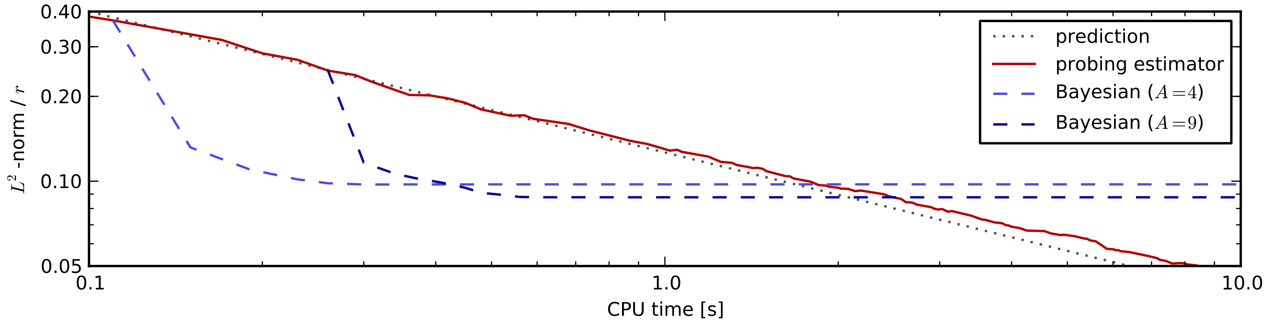

Figure 2: The L^2-norm of the error (divided by the number of pixels r) as a function of CPU time for the trivial case: The evolution of the probing estimator (solid), its theoretical prediction $\propto 1/\sqrt{A}$ (dotted) and the Bayesian estimators starting with four (dashed light) and nine samples (dashed dark) respectively are shown.

where the diagonal entries need to fulfill $X_{ii} \geq 2 \, \forall \, i \in \{1, \ldots, r\}$ for positive definiteness and are drawn with a simple structure on the sphere, see Fig. 1.

The normalized L^2-norms of the residual error[4] serve as an accuracy measure and are shown as a function of CPU time in Fig. 2.

Although X as an operator could be implemented very efficiently, we use the much more expensive full matrix multiplication to have realistic computational costs like those of applying a more complex matrix. But this trivial case should only hold as a proof of concept, more sensible examples are discussed in the following.

As one can clearly see in Fig. 2 the probing estimator improves continuously with an increasing number of probes and shows an overall proportionality to $1/\sqrt{A}$ as argued in Sect. III. However, the Bayesian estimator given a set of samples converges in only a couple of iterations to a result with an accuracy the pure probing will first reach after investing a factor of a few more CPU time. For a fixed amount of computation time the Bayesian estimator excels the probing estimator, as can be seen in Fig. 1. It is also evident that the Bayesian estimator must reach a lower limit in its progress because only a limited data set is provided containing finitely accurate information.

2. *Realistic case*

For a more realistic mock example we consider a covariance matrix $D = (S^{-1} + N^{-1})^{-1}$ similar to the one described by (7) where the signal covariance S is completely defined by a power spectrum

$$C_l \propto (\max\{1, l\})^{-2} \,. \tag{26}$$

The noise covariance N is characterized by two effects, first high noise in one of the twelve HEALPix basis pixels representing a defect in the detector, and second smoothly increased noise towards the poles imitating an observational effect[5]. The described noise covariance and the resulting propagator D are illustrated in Fig. 3, where one can see the conservation of the noise structure and the smoothing effect of the power spectrum.

[4] Meaning mathematically $||\mathrm{diag}\,[X] - f||_2/r$ or $||\mathrm{diag}\,[X] - m||_2/r$ respectively to be exact.

[5] The noise variance is assigned to each pixel i according to $N_{ii}^{-1} = (0.005 + 8\,(h_i(h_i - h_{\max}))^2 \,/ h_{\max}^4)$, where h_i is the HEALPix ring number associated to the pixel i.

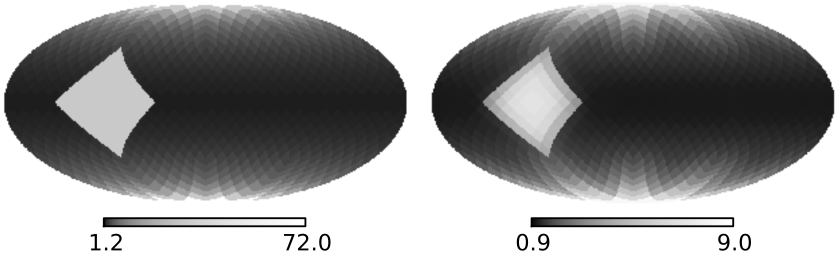

Figure 3: The realistic case: (left) the matrix diagonal of the mock noise covariance N and (right) the 'true' diagonal of the propagator $D = (S^{-1} + N^{-1})^{-1}$.

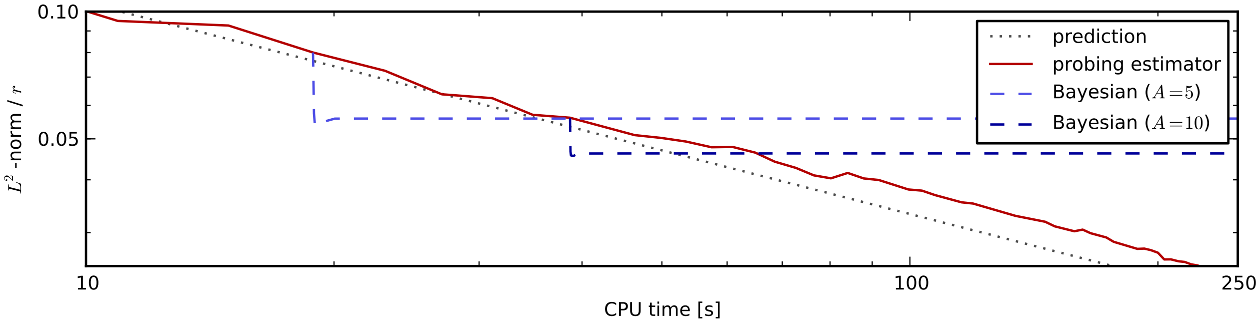

Figure 4: Same as Fig. 2, only for the propagator for the realistic case.

The performance of both algorithms is shown in Fig. 4. Our algorithm performs qualitatively in the same way as in the trivial case but the overall gain in accuracy or time is quantitatively lower. It is also noticeable that the relative advantage of the proposed method decreases with the number of used random vectors. Consequently the matrix diagonal inference method pays off best in cases where a rough estimate using only a few probes is sufficient.

B. Faraday sky uncertainty

Next, we attempt to use our algorithm in a real physical application. We consider the inference problem discussed in [9]. In that work, an all sky map of the galactic Faraday depth was reconstructed from a catalog of measurements of the Faraday depths of 37 543 point sources [12]. The data were modeled according to a linear measurement procedure, Eq. (3), with a response matrix R that encodes both the probing of the all sky field in the directions of the point sources and a multiplication with a scalar function of galactic latitude, $p(\vartheta)$. This galactic profile function was introduced to partially account for the large scale anisotropy introduced by the presence of the galactic disk on the sky. It was calculated as the root mean square rotation measure value per latitude bin from the same data set as a first step of the reconstruction.

With this response, the signal field becomes a down scaled version of the galactic Faraday depth, i.e. the dimensionless ratio of the Faraday depth to the profile function. To reconstruct this field the critical filter algorithm that was discussed in Sect. II A was used, yielding an estimate for the posterior mean m of the signal field s, as well as an estimate for the components of the angular power spectrum of this field, C_l. In addition, a map showing the uncertainty of the signal estimate m, given by diag $[D]$, is provided in [9]. This was calculated from the information propagator D, which takes on the form (21), by applying the probing estimator discussed in Sect. III.

Here, we show how the application of our Bayesian algorithm to this problem can improve the accuracy and speed up the calculation of this matrix diagonal. In order to be able to compare the results of the probing and Bayesian estimators to the correct matrix diagonal, we reduce the dimensionality of the problem to facilitate the exact calculation of the diagonal via Eq. (7). We do this by reducing the resolution of the all sky map with respect to the one presented in [9] to HEALPix parameter $N_{\text{side}} = 16$, leading to 3 072 pixels, and truncating the reconstructed power spectrum at $l_{\max} = 47$. Furthermore, we use a coarser version of the galactic profile function. In this way, a coarse-grained version of the propagator D is defined and we can calculate its diagonal exactly, as well as in the frequentist and Bayesian way.

Fig. 5 shows the results of these calculations, where the matrix-vector-multiplication was conducted for ten different random vectors in the case of the probing and Bayesian estimators. While still exhibiting a large

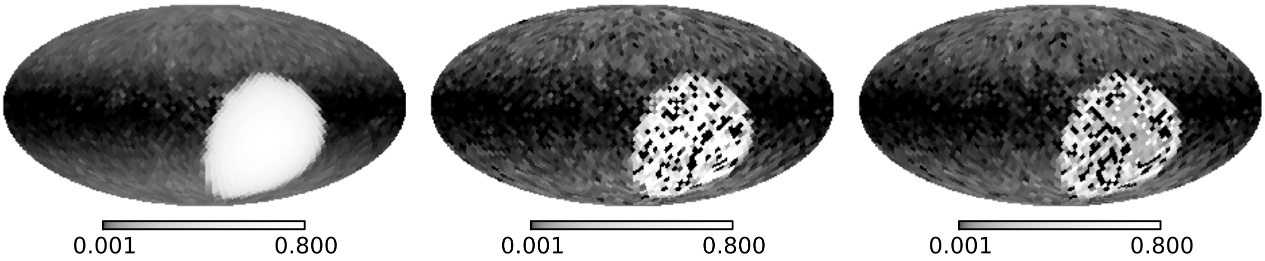

Figure 5: Diagonal of the propagator for the reconstruction of the galactic Faraday depth. The left panel shows the result of the exact calculation according to Eq. (11), the middle panel the probing result after ten iterations, and the right panel the result of the Bayesian estimator, using ten random vectors as well.

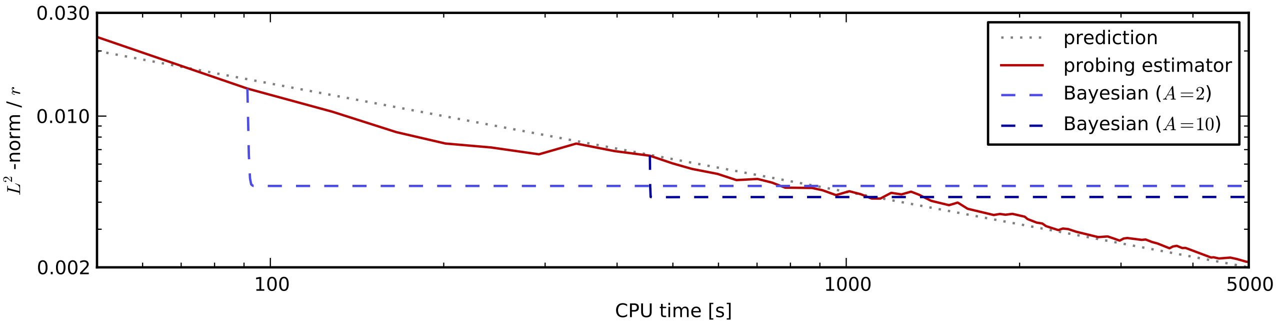

Figure 6: Same as Fig. 2, only for the propagator for the reconstruction of the galactic Faraday depth.

amount of noise, both the probing and Bayesian results show roughly the right structure after only a few iterations. This structure is determined mainly by an oval region of high uncertainty, i.e. large diagonal entries, where no data were taken and a dependence on galactic latitude due to the profile function. Since the signal response includes a multiplication with the galactic profile, it is larger near the galactic plane than near the galactic poles, leading to an overall lower uncertainty within the galactic plane.

From Fig. 5 alone, it is hard to judge whether the Bayesian estimator leads to an improvement over the probing one. We therefore plot again the L^2-norm of the difference between the estimated matrix diagonal and the 'true' one as a function of CPU-time in Fig. 6. Shown is the curve for the pure probing estimator as well as two examples for Bayesian improvements, using two and ten random vectors, respectively. It is evident that for both cases the Bayesian method gives a boost in accuracy with only marginal time consumption. The absolute and relative improvement is larger if one uses fewer random vectors. This shows again that the main strength of the Bayesian method does not lie in the absolute accuracy that can be reached, but rather in the speed-up it provides for obtaining an estimate for the matrix diagonal with intermediate accuracy.

VI. CONCLUSIONS

We reviewed the reliability and robustness of the probing techniques for diagonals of matrices and applied them to several examples where they performed as expected.

A new inference algorithm has been proposed that interprets the probing of the matrix diagonal as a numerical experiment. The outcome of the experiment exhibits all features of a measurement like signal response and noise. Exploiting additional knowledge on the existence of an underlying continuous structure of the matrix diagonal which exhibits (a priori maybe unknown) correlations allowed for an inference method which improves the estimates acquired from probing.

Applying this new inference algorithm on a sample of matrix probes, we retrieved estimators for the matrix diagonal which exhibit a higher accuracy for a small investment of additional computation time. As fewer samples are needed by this new method to achieve the same accuracy, we reduce the number of computational expensive calculations and this way save CPU (and real) time. The new algorithm is especially effective when matrix diagonals need to be calculated only roughly, since the relative gain in accuracy is larger in cases where only a few probes are available.

This has been shown in numerical examples as well as for the uncertainty map appearing in the reconstruction of the galactic Faraday depth.

VII. ACKNOWLEDGMENTS

We thank Mike Bell, Henrik Junklewitz and Georg Robbers for the insightful discussions and productive comments on the manuscript.

Some of the results in this paper have been derived using the HEALPix [13] package. For all computations the free open-source mathematics software system SAGE[6] has been used.

[6] See SAGE homepage `http://www.sagemath.org/`

Appendix A: Proof of the probing estimator

Here we prove that Eq. (13) is indeed implied by Eq. (12). Given a sufficiently large but not necessarily finite set $\{\xi\}$ (with $|\{\xi\}| = A$), the condition given by Eq. (12) becomes

$$\lim_{A\to\infty} \langle \xi_i \xi_j \rangle_{\{\xi\}} = \lim_{A\to\infty} \frac{1}{A} \sum_{a=1}^{A} \xi_i^{(a)} \xi_j^{(a)} = \delta_{ij}. \tag{A1}$$

Inserting this equality in Eq. (13), w.l.o.g. restricted to the average's component $i \in \{1, \ldots, r\}$,

$$\begin{aligned}\Big(\langle \xi * X\, \xi \rangle_{\{\xi\}} \Big)_i &= \frac{1}{A} \sum_{a=1}^{A} \sum_{j=1}^{r} \xi_i^{(a)} X_{ij} \xi_j^{(a)} \\ &= \sum_{j=1}^{r} X_{ij} \underbrace{\frac{1}{A} \sum_{a=1}^{A} \xi_i^{(a)} \xi_j^{(a)}}_{\to \delta_{ij}} \to X_{ii}, \end{aligned} \tag{A2}$$

proves the statement.

[1] M. F. Hutchinson, Communications in Statistics - Simulation and Computation **18**, 1059 (1989).

[2] C. Bekas, E. Kokiopoulou, and Y. Saad, Applied Numerical Mathematics archive **57** (2007).

[3] J. M. Tang and Y. Saad, Numerical Linear Algebra with Applications (2011).

[4] E. Aune and D. P. Simpson, ArXiv e-prints (2011), 1105.5256.

[5] A. Rohde and A. B. Tsybakov, ArXiv e-prints (2009), 0912.5338.

[6] T. A. Enßlin and M. Frommert, Phys. Rev. D **83**, 105014 (2011), 1002.2928.

[7] T. A. Enßlin, M. Frommert, and F. S. Kitaura, Phys. Rev. D **80**, 105005 (2009), 0806.3474.

[8] T. A. Enßlin and C. Weig, Phys. Rev. E **82**, 051112 (2010), 1004.2868.

[9] N. Oppermann, H. Junklewitz, G. Robbers, and T. A. Enßlin, A&A **530**, A89+ (2011), 1008.1246.

[10] N. Oppermann, G. Robbers, and T. A. Ensslin, ArXiv e-prints (2011), 1107.2384.

[11] J. R. Shewchuk, Technical report, Carnegie Mellon University, Pittsburgh, PA (1994).

[12] A. R. Taylor, J. M. Stil, and C. Sunstrum, Astrophys. J. **702**, 1230 (2009).

[13] K. M. Górski, E. Hivon, A. J. Banday, B. D. Wandelt, F. K. Hansen, M. Reinecke, and M. Bartelmann, Astrophys. J. **622**, 759 (2005), arXiv:astro-ph/0409513.