Bayesian Inference of Log Determinants

Jack Fitzsimons[1] **Kurt Cutajar**[2] **Michael Osborne**[1] **Stephen Roberts**[1] **Maurizio Filippone**[2]

[1] Information Engineering, University of Oxford, UK
[2] Department of Data Science, EURECOM, France

Abstract

The log determinant of a kernel matrix appears in a variety of machine learning problems, ranging from determinantal point processes and generalized Markov random fields, through to the training of Gaussian processes. Exact calculation of this term is often intractable when the size of the kernel matrix exceeds a few thousands. In the spirit of probabilistic numerics, we reinterpret the problem of computing the log determinant as a Bayesian inference problem. In particular, we combine prior knowledge in the form of bounds from matrix theory and evidence derived from stochastic trace estimation to obtain probabilistic estimates for the log determinant and its associated uncertainty within a given computational budget. Beyond its novelty and theoretic appeal, the performance of our proposal is competitive with state-of-the-art approaches to approximating the log determinant, while also quantifying the uncertainty due to budget-constrained evidence.

1 INTRODUCTION

Developing scalable learning models without compromising performance is at the forefront of machine learning research. The scalability of several learning models is predominantly hindered by linear algebraic operations having large computational complexity, among which is the computation of the log determinant of a matrix (Golub & Van Loan, 1996). The latter term features heavily in the machine learning literature, with applications including spatial models (Aune et al., 2014; Rue & Held, 2005), kernel-based models (Davis et al., 2007; Rasmussen & Williams, 2006), and Bayesian learning (Mackay, 2003).

The standard approach for evaluating the log determinant of a positive definite matrix involves the use of Cholesky decomposition (Golub & Van Loan, 1996), which is employed in various applications of statistical models such as kernel machines. However, the use of Cholesky decomposition for general dense matrices requires $\mathcal{O}(n^3)$ operations, whilst also entailing memory requirements of $\mathcal{O}(n^2)$. In view of this computational bottleneck, various models requiring the log determinant for inference bypass the need to compute it altogether (Anitescu et al., 2012; Stein et al., 2013; Cutajar et al., 2016; Filippone & Engler, 2015).

Alternatively, several methods exploit sparsity and structure within the matrix itself to accelerate computations. For example, sparsity in Gaussian Markov random fields (GMRFs) arises from encoding conditional independence assumptions that are readily available when considering low-dimensional problems. For such matrices, the Cholesky decompositions can be computed in fewer than $\mathcal{O}(n^3)$ operations (Rue & Held, 2005; Rue et al., 2009). Similarly, Kronecker-based linear algebra techniques may be employed for kernel matrices computed on regularly spaced inputs (Saatçi, 2011). While these ideas have proven successful for a variety of specific applications, they cannot be extended to the case of general dense matrices without assuming special forms or structures for the available data.

To this end, general approximations to the log determinant frequently build upon stochastic trace estimation techniques using iterative methods (Avron & Toledo, 2011). Two of the most widely-used polynomial approximations for large-scale matrices are the Taylor and Chebyshev expansions (Aune et al., 2014; Han et al., 2015). A more recent approach draws from the possibility of estimating the trace of functions using stochastic Lanczos quadrature (Ubaru et al., 2016), which has been shown to outperform polynomial approximations from both a theoretic and empirical perspective.

Inspired by recent developments in the field of probabilistic numerics (Hennig et al., 2015), in this work we propose an alternative approach for calculating the log determinant of a matrix by expressing this computation as a Bayesian quadrature problem. In doing so, we reformulate the problem of *computing* an intractable quantity as an *estimation* problem, where the goal is to infer the correct result using tractable computations that can be carried out within a given time budget. In particular, we model the eigenvalues of a matrix A from noisy observations of $\mathrm{Tr}(A^k)$ obtained through stochastic trace estimation using the Taylor approximation method (Zhang & Leithead, 2007). Such a model can then be used to make predictions on the infinite series of the Taylor expansion, yielding the estimated value of the log determinant. Aside from permitting a probabilistic approach for predicting the log determinant, this approach inherently yields uncertainty estimates for the predicted value, which in turn serves as an indicator of the quality of our approximation.

Our contributions are as follows.

1. We propose a probabilistic approach for computing the log determinant of a matrix which blends different elements from the literature on estimating log determinants under a Bayesian framework.
2. We demonstrate how bounds on the expected value of the log determinant improve our estimates by constraining the probability distribution to lie between designated lower and upper bounds.
3. Through rigorous numerical experiments on synthetic and real data, we demonstrate how our method can yield superior approximations to competing approaches, while also having the additional benefit of uncertainty quantification.
4. Finally, in order to demonstrate how this technique may be useful within a practical scenario, we employ our method to carry out parameter selection for a large-scale determinantal point process.

To the best of our knowledge, this is the first time that the approximation of log determinants is viewed as a Bayesian inference problem, with the resulting quantification of uncertainty being hitherto unexplored thus far.

1.1 RELATED WORK

The most widely-used approaches for estimating log determinants involve extensions of iterative algorithms, such as the conjugate gradient and Lanczos methods, to obtain estimates of functions of matrices (Chen et al., 2011; Han et al., 2015) or their trace (Ubaru et al., 2016). The idea is to rewrite log determinants as the trace of the logarithm of the matrix, and employ trace estimation techniques (Hutchinson, 1990) to obtain unbiased estimates of these. Chen et al. (2011) propose an iterative algorithm to efficiently compute the product of the logarithm of a matrix with a vector, which is achieved by computing a spline approximation to the logarithm function. A similar idea using Chebyshev polynomials has been developed by Han et al. (2015). Most recently, the Lanczos method has been extended to handle stochastic estimates of the trace and obtain probabilistic error bounds for the approximation (Ubaru et al., 2016). Blocking techniques, such as in Ipsen & Lee (2011) and Ambikasaran et al. (2016), have also been proposed.

In our work, we similarly strive to use a small number of matrix-vector products for approximating log determinants. However, we show that by taking a Bayesian approach we can combine priors with the evidence gathered from the intermediate results of matrix-vector products involved in the aforementioned methods to obtain more accurate results. Most importantly, our proposal has the considerable advantage that it provides a full distribution on the approximated value.

Our proposal allows for the inclusion of explicit bounds on log determinants to constrain the posterior distribution over the estimated log determinant (Bai & Golub, 1997). Nyström approximations can also be used to bound the log determinant, as shown in Bardenet & Titsias (2015). Similarly, Gaussian processes (Rasmussen & Williams, 2006) have been formulated directly using the eigendecomposition of its spectrum, where eigenvectors are approximated using the Nyström method (Peng & Qi, 2015). There has also been work on estimating the distribution of kernel eigenvalues by analyzing the spectrum of linear operators (Braun, 2006; Wathen & Zhu, 2015), as well as bounds on the spectra of matrices with particular emphasis on deriving the largest eigenvalue (Wolkowicz & Styan, 1980; Braun, 2006). In this work, we directly consider bounds on the log determinants of matrices (Bai & Golub, 1997).

2 BACKGROUND

As highlighted in the introduction, several approaches for approximating the log determinant of a matrix rely on stochastic trace estimation for accelerating computations. This comes about as a result of the relationship between the log determinant of a matrix, and the corresponding trace of the log-matrix, whereby

$$\log\big(\mathrm{Det}\,(A)\big) = \mathrm{Tr}\big(\log\,(A)\big). \tag{1}$$

Provided the matrix $\log(A)$ can be efficiently sampled, this simple identity enables the use of stochastic trace es-

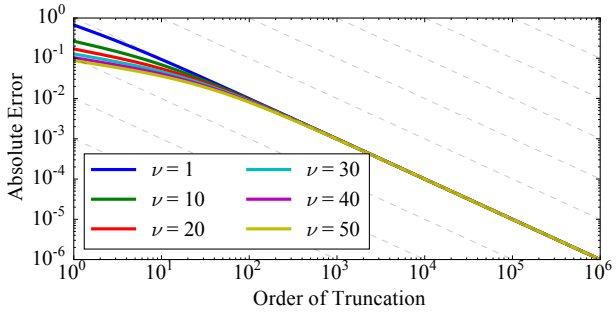

Figure 1: Expected absolute error of truncated Taylor series for stationary ν-continuous kernel matrices. The dashed grey lines indicate $\mathcal{O}(n^{-1})$.

timation techniques (Avron & Toledo, 2011; Fitzsimons et al., 2016). We elaborate further on this concept below.

2.1 STOCHASTIC TRACE ESTIMATION

It is possible to obtain a stochastic estimate of the trace term such that the expectation of the estimate matches the term being approximated (Avron & Toledo, 2011). In this work, we shall consider the Gaussian estimator, whereby we introduce $N_{\mathbf{r}}$ vectors $\mathbf{r}^{(i)}$ sampled from an independently and identically distributed zero-mean and unit variance Gaussian distribution. This yields the unbiased estimate

$$\mathrm{Tr}(A) = \frac{1}{N_r} \sum_{i=1}^{N_r} {\mathbf{r}^{(i)}}^\top A \, \mathbf{r}^{(i)}. \tag{2}$$

Note that more sophisticated trace estimators (see Fitzsimons et al., 2016) may also be considered; without loss of generality, we opt for a more straightforward approach in order to preserve clarity.

2.2 TAYLOR APPROXIMATION

Against the backdrop of machine learning applications, in this work we predominantly consider covariance matrices taking the form of a Gram matrix $K = \{\kappa(\mathbf{x}_i, \mathbf{x}_j)\}_{i,j=1,\ldots,n}$, where the kernel function κ implicitly induces a feature space representation of data points $\mathbf{x}$. Assume K has been normalized such that the maximum eigenvalue is less than or equal to one, $\lambda_0 \leq 1$, where the largest eigenvalue can be efficiently found using Gershgorin intervals (Gershgorin, 1931). Given that covariance matrices are positive semidefinite, we also know that the smallest eigenvalue is bounded by zero, $\lambda_n \geq 0$. Motivated by the identity presented in (1), the Taylor series expansion (Barry & Pace, 1999; Zhang & Leithead, 2007) may be employed for evaluating the log determinant of matrices having eigenvalues bounded between zero and one. In particular, this approach relies on the following logarithm identity,

$$\log\left(I - A\right) = -\sum_{k=1}^{\infty} \frac{A^k}{k}. \tag{3}$$

While the infinite summation is not explicitly computable in finite time, this may be approximated by computing a truncated series instead. Furthermore, given that the trace of matrices is additive, we find

$$\mathrm{Tr}\big(\log\left(I - A\right)\big) \approx -\sum_{k=1}^{m} \frac{\mathrm{Tr}\left(A^k\right)}{k}. \tag{4}$$

The $\mathrm{Tr}(A^k)$ term can be computed efficiently and recursively by propagating $\mathcal{O}(n^2)$ vector-matrix multiplications in a stochastic trace estimation scheme. To compute $\mathrm{Tr}(\log(K))$ we simply set $A = I - K$.

There are two sources of error associated with this approach; the first due to stochastic trace estimation, and the second due to truncation of the Taylor series. In the case of covariance matrices, the smallest eigenvalue tends to be very small, which can be verified by Weyl (1912) and Silverstein (1986)'s observations on the eigenspectra of covariance matrices. This leads to A^k decaying slowly as $k \to \infty$.

In light of the above, standard Taylor approximations to the log determinant of covariance matrices are typically unreliable, even when the exact traces of matrix powers are available. This can be verified analytically based on results from kernel theory, which state that the approximate rate of decay for the eigenvalues of positive definite kernels which are ν-continuous is $\mathcal{O}(n^{-\nu-0.5})$ (Weyl, 1912; Wathen & Zhu, 2015). Combining this result with the absolute error, $E(\lambda)$, of the truncated Taylor approximation we find

$$\begin{aligned}
\mathbb{E}\left[E\left(\lambda\right)\right] &= \mathcal{O}\left(\int_0^1 \lambda^{\nu+0.5}\left(\log\left(\lambda\right) - \sum_{j=1}^{m} \frac{\lambda^j}{j}\right)\mathrm{d}\lambda\right) \\
&= \mathcal{O}\left(\int_0^1 \lambda^{\nu+0.5} \sum_{j=m}^{\infty} \frac{\lambda^j}{j}\mathrm{d}\lambda\right) \\
&= \mathcal{O}\left(\frac{\Psi^{(0)}\left(m+\nu+1.5\right) - \Psi^{(0)}\left(m\right)}{\nu+1.5}\right),
\end{aligned}$$

where $\Psi^{(0)}(\cdot)$ is the Digamma function. In Figure 1, we plot the relationship between the order of the Taylor approximation and the expected absolute error. It can be observed that irrespective of the continuity of the kernel, the error converges at a rate of $\mathcal{O}(n^{-1})$.

3 THE PROBABILISTIC NUMERICS APPROACH

We now propose a probabilistic numerics (Hennig et al., 2015) approach: we reframe a numerical computation (in this case, trace estimation) as probabilistic inference. Probabilistic numerics usually requires distinguishing: an appropriate latent function; data and; the ultimate object of interest. Given the data, a posterior distribution is calculated for the object of interest. For instance, in numerical integration, the latent function is the integrand, f, the data are evaluations of the integrand, $f(x)$, and the object of interest is the value of the integral, $\int f(x)p(x)\mathrm{d}x$ (see § 3.1.1 for more details). In this work, our latent function is the distribution of eigenvalues of A, the data are noisy observations of $\mathrm{Tr}(A^k)$, and the object of interest is $\log(\mathrm{Det}(K))$. For this object of interest, we are able to provide both expected value and variance. That is, although the Taylor approximation to the log determinant may be considered unsatisfactory, the intermediate trace terms obtained when raising the matrix to higher powers may prove to be informative if considered as observations within a probabilistic model.

3.1 RAW MOMENT OBSERVATIONS

We wish to model the eigenvalues of A from noisy observations of $\mathrm{Tr}\left(A^k\right)$ obtained through stochastic trace estimation, with the ultimate goal of making predictions on the infinite series of the Taylor expansion. Let us assume that the eigenvalues are i.i.d. random variables drawn from $P(\lambda_i = x)$, a probability distribution over $x \in [0, 1]$. In this setting $\mathrm{Tr}(A) = n\mathbb{E}_x[P(\lambda_i = x)]$, and more generally $\mathrm{Tr}\left(A^k\right) = n\mathbf{R}_x^{(k)}[P(\lambda_i = x)]$, where $\mathbf{R}_x^{(k)}$ is the k^{th} raw moment over the x domain. The raw moments can thus be computed as,

$$\mathbf{R}_x^{(k)}\left[P\left(\lambda_i = x\right)\right] = \int_0^1 x^k P\left(\lambda_i = x\right)\mathrm{d}x. \quad (5)$$

Such a formulation is appealing because if $P\left(\lambda_i = x\right)$ is modeled as a Gaussian process, the required integrals may be solved analytically using Bayesian Quadrature.

3.1.1 Bayesian Quadrature

Gaussian processes (GPs; Rasmussen & Williams, 2006) are a powerful Bayesian inference method defined over functions $X \to \mathbb{R}$, such that the distribution of functions over any finite subset of the input points $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ is a multivariate Gaussian distribution. Under this framework, the moments of the conditional Gaussian distribution for a set of predictive points, given a set of labels $\mathbf{y} = (y_1, \ldots, y_n)^\top$, may be computed as

$$\mu = \mu_0 + K_*^\top K^{-1}(\mathbf{y} - \mu_0), \quad (6)$$

$$\Sigma = K_{*,*} - K_*^\top K^{-1} K_*, \quad (7)$$

with μ and Σ denoting the posterior mean and variance, and K being the $n \times n$ covariance matrix for the observed variables $\{\mathbf{x}_i, y_i; i \in (1, 2, \ldots n)\}$. The latter is computed as $\kappa(\mathbf{x}, \mathbf{x}')$ for any pair of points $\mathbf{x}, \mathbf{x}' \in X$. Meanwhile, K_* and $K_{*,*}$ respectively denote the covariance between the observable and the predictive points, and the prior over the predicted points. Note that μ_0, the prior mean, may be set to zero without loss of generality.

Bayesian Quadrature (BQ; O'Hagan, 1991) is primarily concerned with performing integration of potentially intractable functions. In this work, we limit our discussion to the setting where the integrand is modeled as a GP,

$$\int p(x)\, f(x)\, \mathrm{d}x, \quad f \sim \mathrm{GP}(\mu, \Sigma),$$

where $p(x)$ is some measure with respect to which we are integrating. A full discussion of BQ may be found in O'Hagan (1991) and Rasmussen & Ghahramani (2002); for the sake of conciseness, we only state the result that the integrals may be computed by integrating the covariance function with respect to $p(x)$ for both K_*,

$$\kappa\left(\int \cdot\, \mathrm{d}x, x'\right) = \int p\left(x\right)\kappa\left(x, x'\right)\mathrm{d}x,$$

and $K_{*,*}$,

$$\kappa\left(\int \cdot\, \mathrm{d}x, \int \cdot\, \mathrm{d}x'\right) = \iint p\left(x\right)\kappa\left(x, x'\right)p\left(x'\right)\mathrm{d}x\mathrm{d}x'.$$

3.2 KERNELS FOR RAW MOMENTS AND INFERENCE ON THE LOG DETERMINANT

Recalling (5), if $P(\lambda_i = x)$ is modeled using a GP, in order to include observations of $\mathbf{R}_x^{(k)}[P(\lambda_i = x)]$, denoted as $\mathbf{R}_x^{(k)}$, we must be able to integrate the kernel with respect to the polynomial in x,

$$\kappa\left(\mathbf{R}_x^{(k)}, x'\right) = \int_0^1 x^k \kappa\left(x, x'\right)\mathrm{d}x, \quad (8)$$

$$\kappa\left(\mathbf{R}_x^{(k)}, \mathbf{R}_{x'}^{(k')}\right) = \int_0^1\int_0^1 x^k \kappa\left(x, x'\right)x'^{k'}\mathrm{d}x\mathrm{d}x'. \quad (9)$$

Although the integrals described above are typically analytically intractable, certain kernels have an elegant analytic form which allows for efficient computation. In this section, we derive the raw moment observations for a histogram kernel, and demonstrate how estimates of the log determinant can be obtained. An alternate polynomial kernel is described in Appendix A.

3.2.1 Histogram Kernel

The entries of the histogram kernel, also known as the piecewise constant kernel, are given by $\kappa(x, x') = \sum_{j=0}^{1-m} \mathcal{H}(\frac{j}{m}, \frac{j+1}{m}, x, x')$, where

$$\mathcal{H}\left(\frac{j}{m}, \frac{j+1}{m}, x, x'\right) = \begin{cases} 1 & x, x' \in \left[\frac{j}{m}, \frac{j+1}{m}\right] \\ 0 & \text{otherwise} \end{cases}.$$

Covariances between raw moments may be computed as follows:

$$\begin{aligned} \kappa\left(\mathbf{R}_x^{(k)}, x'\right) &= \int_0^1 x^k \kappa\left(x, x'\right) \mathrm{d}x \\ &= \frac{1}{k+1}\left(\left(\frac{j+1}{m}\right)^{k+1} - \left(\frac{j}{m}\right)^{k+1}\right), \end{aligned} \tag{10}$$

where in the above x lies in the interval $\left[\frac{j}{m}, \frac{j+1}{m}\right]$. Extending this to the covariance function between raw moments we have,

$$\begin{aligned} \kappa\left(\mathbf{R}_x^{(k)}, \mathbf{R}_{x'}^{(k')}\right) &= \int_0^1 \int_0^1 x^k x'^{k'} \kappa\left(x, x'\right) \mathrm{d}x \mathrm{d}x' \\ &= \sum_{j=0}^{m-1} \prod_{\bar{k} \in (k, k')} \frac{1}{(\bar{k}+1)}\left(\left(\frac{j+1}{m}\right)^{\bar{k}+1} - \left(\frac{j}{m}\right)^{\bar{k}+1}\right). \end{aligned} \tag{11}$$

This simple kernel formulation between observations of the raw moments compactly allows us to perform inference over $P(\lambda_i = x)$. However, the ultimate goal is to predict $\log(\mathrm{Det}(K))$, and hence $\sum_{i=1}^{\infty} \frac{\mathrm{Tr}\left(A^k\right)}{k}$. This requires a seemingly more complex set of kernel expressions; nevertheless, by propagating the implied infinite summations into the kernel function, we can also obtain the closed form solutions for these terms,

$$\begin{aligned} \kappa\left(\sum_{k=1}^{\infty} \frac{\mathbf{R}_x^{(k)}}{k}, \mathbf{R}_{x'}^{(k')}\right) = \sum_{j=0}^{m-1} \frac{1}{k'+1}\Bigg(\left(\frac{j+1}{m}\right)^{k'+1} - \\ \left(\frac{j}{m}\right)^{k'+1}\Bigg)\left(S\left(\frac{j+1}{m}\right) - S\left(\frac{j}{m}\right)\right) \end{aligned} \tag{12}$$

$$\kappa\left(\sum_{k=1}^{\infty} \frac{\mathbf{R}_x^{(k)}}{k}, \sum_{k'=1}^{\infty} \frac{\mathbf{R}_{x'}^{(k')}}{k'}\right) = \sum_{j=0}^{m-1}\left(S\left(\frac{j+1}{m}\right) - S\left(\frac{j}{m}\right)\right)^2 \tag{13}$$

where $S(\alpha) = \sum_{k=1}^{\infty} \frac{\alpha^{k+1}}{k(k+1)}$, which has the convenient identity for $0 < \alpha < 1$,

$$S(\alpha) = \alpha + (1-\alpha)\log(1-\alpha).$$

Following the derivations presented above, we can finally go about computing the prediction for the log determinant, and its corresponding variance, using the GP posterior equations given in (6) and (7). This can be achieved by replacing the terms K_* and $K_{*,*}$ with the constructions presented in (12) and (13), respectively. The entries of K are filled in using (11), whereas $\mathbf{y}$ denotes the noisy observations of $\mathrm{Tr}\left(A^k\right)$.

3.2.2 Prior Mean Function

While GPs, and in this case BQ, can be applied with a zero mean prior without loss of generality, it is often beneficial to have a mean function as an initial starting point. If $P(\lambda_i = x)$ is composed of a constant mean function $g(\lambda_i = x)$, and a GP is used to model the residual, we have that

$$P\left(\lambda_i = x\right) = g\left(\lambda_i = x\right) + f\left(\lambda_i = x\right).$$

The previously derived moment observations may then be decomposed into,

$$\begin{aligned} \int x^k P\left(\lambda_i = x\right) \mathrm{d}x = \int x^k g\left(\lambda_i = x\right) \mathrm{d}x \\ + \int x^k f\left(\lambda_i = x\right) \mathrm{d}x. \end{aligned} \tag{14}$$

Due to the domain of $P\left(\lambda_i = x\right)$ lying between zero and one, we set a beta distribution as the prior mean, which has some convenient properties. First, it is fully specified by the mean and variance of the distribution, which can be computed using the trace and Frobenius norm of the matrix. Secondly, the r-th raw moment of a Beta distribution parameterized by α and β is

$$\mathbf{R}_x^{(k)}\left[g\left(\lambda_i = x\right)\right] = \frac{\alpha + r}{\alpha + \beta + r},$$

which is straightforward to compute.

In consequence, the expectation of the logarithm of random variables and, hence, the 'prior' log determinant yielded by $g\left(\lambda_i = x\right)$ can be computed as

$$\mathbb{E}[\log(X); X \sim g(\lambda_i = x)] = \Psi(\alpha) - \Psi(\alpha + \beta). \tag{15}$$

This can then simply be added to the previously derived GP expectation of the log determinant.

3.2.3 Using Bounds on the log determinant

As with most GP specifications, there are hyperparameters associated with the prior and the kernel. The optimal settings for these parameters may be obtained via

optimization of the standard GP log marginal likelihood, defined as

$$\text{LML}_{GP} = -\frac{1}{2}\mathbf{y}^\top K^{-1}\mathbf{y} - \frac{1}{2}\log(\text{Det}(K)) + \text{const.}$$

Borrowing from the literature on bounds for the log determinant of a matrix, as described in Appendix B, we can also exploit such upper and lower bounds to truncate the resulting GP distribution to the relevant domain, which is expected to greatly improve the predicted log determinant. These additional constraints can then be propagated to the hyperparameter optimization procedure by incorporating them into the likelihood function via the product rule, as follows:

$$\text{LML} = \text{LML}_{GP} + \log\left(\Phi\left(\frac{a-\hat{\mu}}{\hat{\sigma}}\right) - \Phi\left(\frac{b-\hat{\mu}}{\hat{\sigma}}\right)\right),$$

with a and b representing the upper and lower log determinant bounds respectively, $\hat{\mu}$ and $\hat{\sigma}$ representing the posterior mean and standard deviation, and $\Phi(\cdot)$ representing the Gaussian cumulative density function. Priors on the hyperparameters may be accounted for in a similar way.

3.2.4 Algorithm Complexity and Recap

Due to its cubic complexity, GP inference is typically considered detrimental to the scalability of a model. However, in our formulation, the GP is only being applied to the noisy observations of $\text{Tr}\left(A^k\right)$, which rarely exceed the order of tens of points. As a result, given that we assume this to be orders of magnitude smaller than the dimensionality n of the matrix K, the computational complexity is dominated by the matrix-vector operations involved in stochastic trace estimation, i.e. $\mathcal{O}(n^2)$ for dense matrices and $\mathcal{O}(ns)$ for s-sparse matrices.

The steps involved in the procedure described within this section are summarized as pseudo-code in Algorithm 1. The input matrix A is first normalized by using Gershgorin intervals to find the largest eigenvalue (line 1), and the expected bounds on the log determinant (line 2) are calculated using matrix theory (Appendix B). The noisy Taylor observations up to an expansion order M (lines 3-4), denoted here as $\mathbf{y}$, are then obtained through stochastic trace estimation, as described in § 2.2. These can be modeled using a GP, where the entries of the kernel matrix K (lines 5-7) are computed using (11). The kernel parameters are then tuned as per § 3.2.3 (line 8). Recall that we seek to make a prediction for the infinite Taylor expansion, and hence the exact log determinant. To this end, we must compute K_* (lines 9-10) and $k_{*,*}$ (line 11) using (12) and (13), respectively. The posterior mean and variance (line 12) may then be evaluated by filling in (6) and (7). As outlined in the previous section, the resulting posterior distribution can be truncated using the derived bounds to obtain the final estimates for the log determinant and its uncertainty (line 13).

Algorithm 1 Computing log determinant and uncertainty using probabilistic numerics

Input: PSD matrix $A \in \mathbb{R}^{n\times n}$, raw moments kernel κ, expansion order M, and random vectors Z
Output: Posterior mean MTRN, and uncertainty VTRN

1: $A \leftarrow$ NORMALIZE(A)
2: BOUNDS $\leftarrow$ GETBOUNDS(A)
3: **for** $i \leftarrow 1$ to M **do**
4: $\quad \mathbf{y}_i \leftarrow$ STOCHASTICTAYLOROBS(A, i, Z)
5: **for** $i \leftarrow 1$ to M **do**
6: $\quad$ **for** $j \leftarrow 1$ to M **do**
7: $\quad\quad K_{ij} \leftarrow \kappa(i, j)$
8: $\kappa, K \leftarrow$ TUNEKERNEL$(K, \mathbf{y}, \text{BOUNDS})$
9: **for** $i \leftarrow 1$ to M **do**
10: $\quad K_{*,i} \leftarrow \kappa(*, i)$
11: $k_{*,*} \leftarrow \kappa(*, *)$
12: MEXP, VEXP $\leftarrow$ GPPRED$(\mathbf{y}, K, K_*, k_{*,*})$
13: MTRN, VTRN $\leftarrow$ TRUNC(MEXP, VEXP, BOUNDS)

4 EXPERIMENTS

In this section, we show how the appeal of this formulation extends beyond its intrinsic novelty, whereby we also consistently obtain performance improvements over competing techniques. We set up a variety of experiments for assessing the model performance, including both synthetically constructed and real matrices. Given the model's probabilistic formulation, we also assess the quality of the uncertainty estimates yielded by the model. We conclude by demonstrating how this approach may be fitted within a practical learning scenario.

We compare our approach against several other estimations to the log determinant, namely approximations based on Taylor expansions, Chebyshev expansions and Stochastic Lanczos quadrature. The Taylor approximation has already been introduced in § 2.2, and we briefly describe the others below.

Chebyshev Expansions: This approach utilizes the m-degree Chebyshev polynomial approximation to the function $\log(I - A)$ (Han et al., 2015; Boutsidis et al., 2015; Peng & Wang, 2015),

$$\text{Tr}\left(\log\left(I - A\right)\right) \approx \sum_{k=0}^{m} c_k \text{Tr}\left(T_k\left(A\right)\right), \qquad (16)$$

where $T_k(x) = AT_{k-1}(A) - T_{k-2}(A)$ starting with

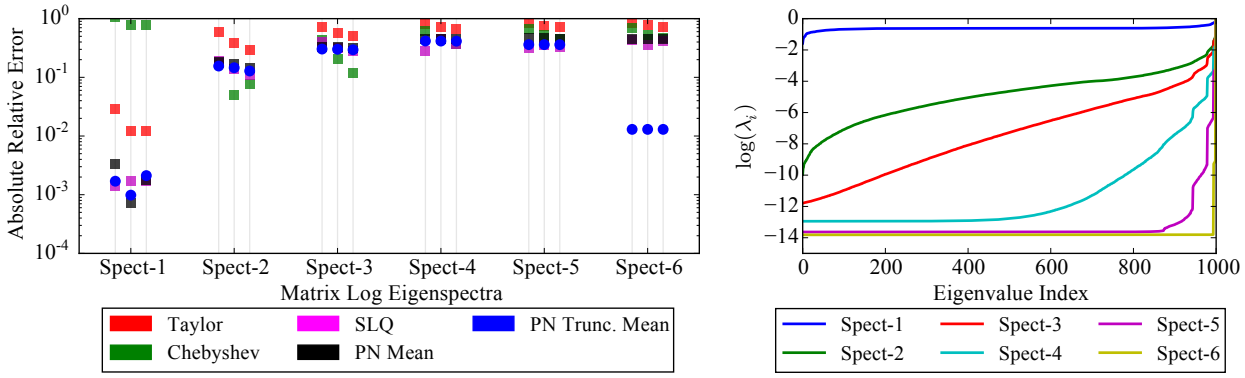

Figure 2: Empirical performance over 6 covariance matrices described in § 4.1. The right figure displays the log eigenspectrum of the matrices and their respective indices. The left figure displays the relative performance of the algorithms for the stochastic trace estimation order set to 5, 25 and 50 (from left to right respectively).

$T_0(A) = 1$ and $T_1(A) = A$, and c_k is defined as

$$c_k = \frac{2}{n+1} \sum_{i=0}^{n} \log(I - x_i) T_k(x_i), \qquad x_i = \cos\left(\frac{\left(i + \frac{1}{2}\right)\pi}{n+1}\right). \tag{17}$$

The Chebyshev approximation is appealing as it gives the best m-degree polynomial approximation of $\log(I - x)$ under the L_∞-norm. The error induced by general Chebyshev polynomial approximations has also been thoroughly investigated (Han et al., 2015).

Stochastic Lanczos Quadrature: This approach (Ubaru et al., 2016) relies on stochastic trace estimation to approximate the trace using the identity presented in (1). If we consider the eigendecomposition of matrix A into $Q\Lambda Q^\top$, the quadratic form in the equation becomes

$$\begin{aligned} \mathbf{r}^{(i)\top} \log(A) \mathbf{r}^{(i)} &= \mathbf{r}^{(i)\top} Q \log(\Lambda) Q^\top \mathbf{r}^{(i)} \\ &= \sum_{k=1}^{n} \log(\lambda_k) \mu_k^2 \end{aligned},$$

where μ_k denotes the individual components of $Q^\top \mathbf{r}^{(i)}$. By transforming this term into a Riemann-Stieltjes integral $\int_a^b \log(t) \mathrm{d}\mu(t)$, where $\mu(t)$ is a piecewise constant function (Ubaru et al., 2016), we can approximate it as

$$\int_a^b \log(t) \mathrm{d}\mu(t) \approx \sum_{j=0}^{m} \omega_j \log(\theta_j),$$

where m is the degree of the approximation, while the sets of ω and θ are the parameters to be inferred using Gauss quadrature. It turns out that these parameters may be computed analytically using the eigendecomposition of the low-rank tridiagonal transformation of A obtained using the Lanczos algorithm (Paige, 1972). Denoting the resulting eigenvalues and eigenvectors by θ and y respectively, the quadratic form may finally be evaluated as,

$$\mathbf{r}^{(i)\top} \log(A) \mathbf{r}^{(i)} \approx \sum_{j=0}^{m} \tau_j^2 \log(\theta_j), \tag{18}$$

with $\tau_j = \left[e_1^T y_j\right]$.

4.1 SYNTHETICALLY CONSTRUCTED MATRICES

Previous work on estimating log determinants have implied that the performance of any given method is closely tied to the shape of the eigenspectrum for the matrix under review. As such, we set up an experiment for assessing the performance of each technique when applied to synthetically constructed matrices whose eigenvalues decay at different rates. Given that the computational complexity of each method is dominated by the number of matrix-vector products (MVPs) incurred, we also illustrate the progression of each technique for an increasing allowance of MVPs. All matrices are constructed using a Gaussian kernel evaluated over 1000 input points.

As illustrated in Figure 2, the estimates returned by our approach are consistently on par with (and frequently superior to) those obtained using other methods. For matrices having slowly-decaying eigenvalues, standard Chebyshev and Taylor approximations fare quite poorly, whereas SLQ and our approach both yield comparable results. The results become more homogeneous across methods for faster-decaying eigenspectra, but our method is frequently among the top two performers. For our approach, it is also worth noting that truncating the GP using known bounds on the log determinant indeed

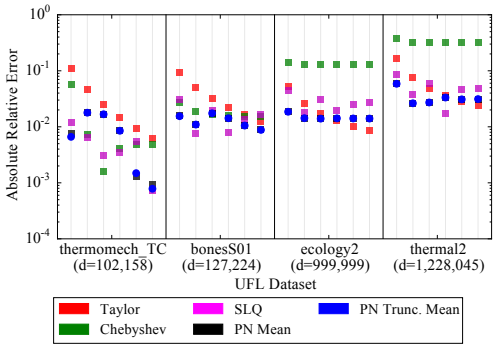

Figure 3: Methods compared on a variety on UFL Sparse Datasets. Each dataset was ran the matrix approximately raised to the power of 5, 10, 15, 20, 25 and 30 (left to right) using stochastic trace estimation.

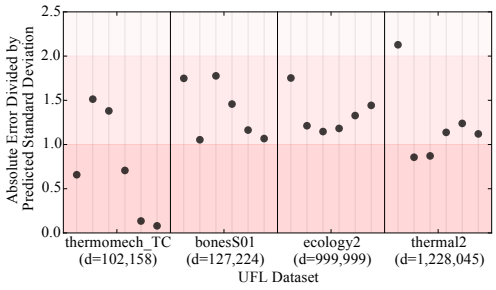

Figure 4: Quality of uncertainty estimates on UFL datasets, measured as the ratio of the absolute error to the predicted variance. As before, results are shown for increasing computational budgets (MVPs). The true value lay outside 2 standard deviations in only one of 24 trials.

results in superior posterior estimates. This is particularly evident when the eigenvalues decay very rapidly. Somewhat surprisingly, the performance does not seem to be greatly affected by the number of budgeted MVPs.

4.2 UFL SPARSE DATASETS

Although we have so far limited our discussion to covariance matrices, our proposed method is amenable to any positive semi-definite matrix. To this end, we extend the previous experimental set-up to a selection of real, sparse matrices obtained from the SuiteSparse Matrix Collection (Davis & Hu, 2011). Following Ubaru et al. (2016), we list the true values of the log determinant reported in Boutsidis et al. (2015), and compare all other approaches to this baseline.

The results for this experiment are shown in Figure 3. Once again, the estimates obtained using our probabilistic approach achieve comparable accuracy to the competing techniques, and several improvements are noted for larger allowances of MVPs. As expected, the SLQ approach generally performs better than Taylor and Chebyshev approximations, especially for smaller computational budgets. Even so, our proposed technique consistently appears to have an edge across all datasets.

4.3 UNCERTAINTY QUANTIFICATION

One of the notable features of our proposal is the ability to quantify the uncertainty of the predicted log determinant, which can be interpreted as an indicator of the quality of the approximation. Given that none of the other techniques offer such insights to compare against, we assess the quality of the model's uncertainty estimates by measuring the ratio of the absolute error to the predicted standard deviation (uncertainty). For the latter to be meaningful, the error should ideally lie within only a few multiples of the standard deviation.

In Figure 4, we report this metric for our approach when using the histogram kernel. We carry out this evaluation over the matrices introduced in the previous experiment, once again showing how the performance varies for different MVP allowances. In all cases, the absolute error of the predicted log determinant is consistently bounded by at most twice the predicted standard deviation, which is very sensible for such a probabilistic model.

4.4 MOTIVATING EXAMPLE

Determinantal point processes (DPPs; Macchi, 1975) are stochastic point processes defined over subsets of data such that an established degree of repulsion is maintained. A DPP, $\mathcal{P}$, over a discrete space $y \in \{1, \ldots, n\}$ is a probability measure over all subsets of y such that

$$\mathcal{P}(A \in y) = \mathrm{Det}(K_A),$$

where K_A is a positive definite matrix having all eigenvalues less than or equal to 1. A popular method for modeling data via K is the *L-ensemble* approach (Borodin, 2009), which transforms kernel matrices, L, into an appropriate K,

$$K = (L + I)^{-1} L.$$

The goal of inference is to correctly parameterize L given observed subsets of y, such that the probability of unseen subsets can be accurately inferred in the future.

Given that the log-likelihood term of a DPP requires the log determinant of L, naïve computations of this term are intractable for large sample sizes. In this experiment, we demonstrate how our proposed approach can be employed to the purpose of parameter optimization for

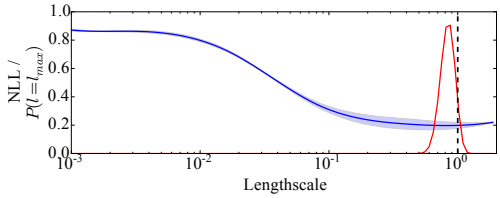

Figure 5: The rescaled negative log-likelihood (NLL) of DPP with varying length scale (blue) and probability of maximum likelihood (red). Cubic interpolation was used between inferred likelihood observations. Ten samples, z, were taken to polynomial order 30.

large-scale DPPs. In particular, we sample points from a DPP defined on a lattice over $[-1, 1]^5$, with one million points at uniform intervals. A Gaussian kernel with lengthscale parameter l is placed over these points, creating the true L. Subsets of the lattice points can be drawn by taking advantage of the tensor structure of L, and we draw five sets of 12,500 samples each. For a given selection of lengthscale options, the goal of this experiment is to confirm that the DPP likelihood of the obtained samples is indeed maximized when L is parameterized by the true lengthscale, l. As shown in Figure 5, the computed uncertainty allows us to derive a distribution over the true lengthscale which, despite using few matrix-vector multiplications, is very close to the optimal.

5 CONCLUSION

In a departure from conventional approaches for estimating the log determinant of a matrix, we propose a novel probabilistic framework which provides a Bayesian perspective on the literature of matrix theory and stochastic trace estimation. In particular, our approach enables the log determinant to be inferred from noisy observations of $\mathrm{Tr}\left(A^k\right)$ obtained from stochastic trace estimation. By modeling these observations using a GP, a posterior estimate for the log determinant may then be computed using Bayesian Quadrature. Our experiments confirm that the results obtained using this model are highly comparable to competing methods, with the additional benefit of measuring uncertainty.

We forecast that the foundations laid out in this work can be extended in various directions, such as exploring more kernels on the raw moments which permit tractable Bayesian Quadrature. The uncertainty quantified in this work is also a step closer towards fully characterizing the uncertainty associated with approximating large-scale kernel-based models.

References

Ambikasaran, S., Foreman-Mackey, D., Greengard, L., Hogg, D. W., and O'Neil, M. Fast Direct Methods for Gaussian Processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2):252–265, 2016.

Anitescu, M., Chen, J., and Wang, L. A Matrix-free Approach for Solving the Parametric Gaussian Process Maximum Likelihood Problem. *SIAM J. Scientific Computing*, 34(1), 2012.

Aune, E., Simpson, D. P., and Eidsvik, J. Parameter Estimation in High Dimensional Gaussian Distributions. *Statistics and Computing*, 24(2):247–263, 2014.

Avron, H. and Toledo, S. Randomized Algorithms for Estimating the Trace of an Implicit Symmetric Positive Semi-definite Matrix. *J. ACM*, 58(2):8:1–8:34, 2011.

Bai, Z. and Golub, G. H. Bounds for the Trace of the Inverse and the Determinant of Symmetric Positive Definite Matrices. *Annals of Numerical Mathematics*, 4:29–38, 1997.

Bardenet, R. and Titsias, M. K. Inference for Determinantal Point Processes Without Spectral Knowledge. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, pp. 3393–3401, 2015.

Barry, R. P. and Pace, R. K. Monte Carlo Estimates of the Log-Determinant of Large Sparse Matrices. *Linear Algebra and its applications*, 289(1):41–54, 1999.

Borodin, A. Determinantal point processes. *arXiv preprint arXiv:0911.1153*, 2009.

Boutsidis, C., Drineas, P., Kambadur, P., and Zouzias, A. A Randomized Algorithm for Approximating the Log Determinant of a Symmetric Positive Definite Matrix. *CoRR*, abs/1503.00374, 2015.

Braun, M. L. Accurate Error Bounds for the Eigenvalues of the Kernel Matrix. *Journal of Machine Learning Research*, 7:2303–2328, December 2006.

Chen, J., Anitescu, M., and Saad, Y. Computing f(A)b via Least Squares Polynomial Approximations. *SIAM Journal on Scientific Computing*, 33(1):195–222, 2011.

Cutajar, K., Osborne, M., Cunningham, J., and Filippone, M. Preconditioning Kernel Matrices. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24*, 2016.

Davis, J. V., Kulis, B., Jain, P., Sra, S., and Dhillon, I. S. Information-theoretic Metric Learning. In *Proceedings of the Twenty-Fourth International Conference (ICML 2007), Corvallis, Oregon, USA, June 20-24, 2007*, pp. 209–216, 2007.

Davis, T. A. and Hu, Y. The University of Florida Sparse Matrix Collection. *ACM Transactions on Mathematical Software (TOMS)*, 38(1):1, 2011.

Filippone, M. and Engler, R. Enabling Scalable Stochastic Gradient-based inference for Gaussian processes by employing the Unbiased LInear System SolvEr (ULISSE). In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, July 6-11*, 2015.

Fitzsimons, J. K., Osborne, M. A., Roberts, S. J., and Fitzsimons, J. F. Improved Stochastic Trace Estimation using Mutually Unbiased Bases. *CoRR*, abs/1608.00117, 2016.

Gershgorin, S. Uber die Abgrenzung der Eigenwerte einer Matrix. *Izvestija Akademii Nauk SSSR, Serija Matematika*, 7(3):749–754, 1931.

Golub, G. H. and Van Loan, C. F. *Matrix computations*. The Johns Hopkins University Press, 3rd edition, October 1996. ISBN 080185413.

Han, I., Malioutov, D., and Shin, J. Large-scale Log-Determinant computation through Stochastic Chebyshev Expansions. In Bach, F. R. and Blei, D. M. (eds.), *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, 2015.

Hennig, P., Osborne, M. A., and Girolami, M. Probabilistic Numerics and Uncertainty in Computations. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 471(2179), 2015.

Hutchinson, M. A Stochastic Estimator of the Trace of the Influence Matrix for Laplacian Smoothing Splines. *Communications in Statistics - Simulation and Computation*, 19(2):433–450, 1990.

Ipsen, I. C. F. and Lee, D. J. Determinant Approximations, May 2011.

Macchi, O. The Coincidence Approach to Stochastic point processes. *Advances in Applied Probability*, 7:83–122, 1975.

Mackay, D. J. C. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, first edition edition, June 2003. ISBN 0521642981.

O'Hagan, A. Bayes-Hermite Quadrature. *Journal of Statistical Planning and Inference*, 29:245–260, 1991.

Paige, C. C. Computational Variants of the Lanczos method for the Eigenproblem. *IMA Journal of Applied Mathematics*, 10(3):373–381, 1972.

Peng, H. and Qi, Y. EigenGP: Gaussian Process Models with Adaptive Eigenfunctions. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, pp. 3763–3769. AAAI Press, 2015.

Peng, W. and Wang, H. Large-scale Log-Determinant Computation via Weighted L_2 Polynomial Approximation with Prior Distribution of Eigenvalues. In *International Conference on High Performance Computing and Applications*, pp. 120–125. Springer, 2015.

Rasmussen, C. E. and Williams, C. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

Rasmussen, C. E. and Ghahramani, Z. Bayesian Monte Carlo. In *Advances in Neural Information Processing Systems 15, NIPS 2002, December 9-14, 2002, Vancouver, British Columbia, Canada*, pp. 489–496, 2002.

Rue, H. and Held, L. *Gaussian Markov Random Fields: Theory and Applications*, volume 104 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London, 2005.

Rue, H., Martino, S., and Chopin, N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):319–392, 2009.

Saatçi, Y. *Scalable Inference for Structured Gaussian Process Models*. PhD thesis, University of Cambridge, 2011.

Silverstein, J. W. Eigenvalues and Eigenvectors of Large Dimensional Sample Covariance Matrices. *Contemporary Mathematics*, 50:153–159, 1986.

Stein, M. L., Chen, J., and Anitescu, M. Stochastic Approximation of Score functions for Gaussian processes. *The Annals of Applied Statistics*, 7(2):1162–1191, 2013. doi: 10.1214/13-AOAS627.

Ubaru, S., Chen, J., and Saad, Y. Fast Estimation of tr (f (a)) via Stochastic Lanczos Quadrature. 2016.

Wathen, A. J. and Zhu, S. On Spectral Distribution of Kernel Matrices related to Radial Basis functions. *Numerical Algorithms*, 70(4):709–726, 2015.

Weyl, H. Das asymptotische Verteilungsgesetz der Eigenwerte linearer partieller Differentialgleichungen (mit einer Anwendung auf die Theorie der Hohlraumstrahlung). *Mathematische Annalen*, 71(4):441–479, 1912.

Wolkowicz, H. and Styan, G. P. Bounds for Eigenvalues using Traces. *Linear algebra and its applications*, 29: 471–506, 1980.

Zhang, Y. and Leithead, W. E. Approximate Implementation of the logarithm of the Matrix Determinant in Gaussian process Regression. *Journal of Statistical Computation and Simulation*, 77(4):329–348, 2007.

A POLYNOMIAL KERNEL

Similar to the derivation of the histogram kernel, we can also derive the polynomial kernel for moment observations. The entries of the polynomial kernel, given by $k(x, x') = (xx' + c)^d$, can be integrated over as,

$$\begin{aligned}\kappa\left(\mathbf{R}_x^{(k)}, x'\right) &= \int_0^1 \sum_{i=1}^{d} \binom{d}{i} x^{k+i} x'^i c^{d-i} dx \\ &= \sum_{i=1}^{d} \binom{d}{i} \frac{x'^i c^{d-i}}{k+i+1},\end{aligned} \tag{19}$$

$$\begin{aligned}\kappa\left(\mathbf{R}_x^{(k)}, \mathbf{R}_{x'}^{(k')}\right) &= \int_0^1 \int_0^1 \sum_{i=1}^{d} \binom{d}{i} x^{k+i} x'^{k'+i} c^{d-i} dx dx' \\ &= \sum_{i=1}^{d} \binom{d}{i} \frac{c^{d-i}}{(k+i+1)(k'+i+1)}.\end{aligned} \tag{20}$$

As with the histogram kernel, the infinite sum of the Taylor expansion can also be combined into the Gaussian process,

$$\begin{aligned}\kappa\left(\sum_{k=1}^{\infty} \frac{\mathbf{R}_x^{(k)}}{k}, \mathbf{R}_{x'}^{(k')}\right) &= \frac{1}{k} \sum_{k=1}^{\infty} \sum_{i=1}^{d} \binom{d}{i} \frac{c^{d-i}}{(k+i+1)(k'+i+1)} \\ &= \sum_{i=1}^{d} \binom{d}{i} \frac{c^{d-i}\left(\Psi^{(0)}(i+2)+\gamma\right)}{(i+1)(k'+i+1)},\end{aligned} \tag{21}$$

$$\begin{aligned}\kappa\left(\sum_{k=1}^{\infty} \frac{\mathbf{R}_x^{(k)}}{k}, \sum_{k'=1}^{\infty} \frac{\mathbf{R}_{x'}^{(k')}}{k'}\right) &= \frac{1}{kk'} \sum_{k=1}^{\infty} \sum_{k'=1}^{\infty} \sum_{i=1}^{d} \binom{d}{i} \frac{c^{d-i}}{(k+i+1)(k'+i+1)} \\ &= \sum_{i=1}^{d} \binom{d}{i} \frac{c^{d-i}\left(\Psi^{(0)}(i+2)+\gamma\right)^2}{(i+1)^2}.\end{aligned} \tag{22}$$

In the above, $\Psi^{(0)}(\cdot)$ is the digamma function and γ is the Euler-Mascheroni constant. We strongly believe that the polynomial and histogram kernels are not the only kernels which can be analytically derived to include moment observations but act as a reasonable initial choice for practitioners.

B BOUNDS ON LOG DETERMINANTS

For the sake of completeness, we restate the bounds on the log determinants used throughout this paper (Bai & Golub, 1997).

Theorem 1 *Let A be an n-by-n symmetric positive definite matrix, $\mu_1 = Tr(A)$, $\mu_2 = \|A\|_F^2$ and $\lambda_i(A) \in [\alpha; \beta]$ with $\alpha > 0$, then*

$$\begin{bmatrix} \log\alpha \\ \log t \end{bmatrix}^T \begin{bmatrix} \alpha & t \\ \alpha^2 & t^2 \end{bmatrix}^{-1} \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \leq \textit{Tr}(\log(A)) \leq \begin{bmatrix} \log\beta \\ \log\bar{t} \end{bmatrix}^T \begin{bmatrix} \beta & \bar{t} \\ \beta^2 & \bar{t}^2 \end{bmatrix}^{-1} \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$$

where,

$$t = \frac{\alpha\mu_1 - \mu_2}{\alpha n - \mu_1}, \quad \bar{t} = \frac{\beta\mu_1 - \mu_2}{\beta n - \mu_1}$$

This bound can be easily computed while loading the matrix as both the trace and Frobenius norm can be readily calculated using summary statistics. However, bounds on the maximum and minimum must also be derived. We chose to use Gershgorin intervals to bound the eigenvalues (Gershgorin, 1931).