Carnegie Mellon University

Constructing approximately sufficient ABC summary statistics

A Dissertation Submitted to the Graduate School in Partial Fulfillment of the Requirements

for the degree

Doctor of Philosophy

in

Statistics

by

Michael Vespe

Department of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213

August 29, 2016

Abstract

Approximate Bayesian Computation (ABC) is a relatively recent method, made possible by recent advances in computation, which allows for inference on parameters in settings where forward simulation models exist but wherein the likelihood need not be specified explicitly. Instead, candidate parameter values are retained or rejected according to the extent to which data simulated using those parameters resemble the observed data. This requires a means of comparing data sets, a step which, for high-dimensional data, typically occurs in the space of lower-dimensional summary statistics. Ideal ABC summary statistics would be minimally sufficient, but sufficient statistics are not readily available in most complex model settings. We propose the *common-space mapping method* (CSMM) for constructing low-dimensional ABC summary statistics based on the relationship between parameters and data in a simulated training set. We also develop an information-theoretic criterion for tuning the CSMM in order to minimize, in expectation, the Kullback-Leibler loss of conditioning on a non-sufficient statistic. In ABC simulation studies for toy problems where the true posterior distribution is known, we find that the CSMM-derived summary statistics perform comparably to the (known) minimal sufficient statistics for those models. We describe weak gravitational lensing as a motivating problem, to which the canonical approach involves potentially restrictive assumptions whose relaxation ABC may permit, and analyze simulated and real weak lensing data using ABC in conjunction with the CSMM.

Acknowledgments

This work is based on observations obtained with MegaPrime/MegaCam, a joint project of CFHT and CEA/DAPNIA, at the Canada-France-Hawaii Telescope (CFHT) which is operated by the National Research Council (NRC) of Canada, the Institut National des Sciences de l'Univers of the Centre National de la Recherche Scientifique (CNRS) of France, and the University of Hawaii. This research used the facilities of the Canadian Astronomy Data Centre operated by the National Research Council of Canada with the support of the Canadian Space Agency. CFHTLenS data processing was made possible thanks to significant computing support from the NSERC Research Tools and Instruments grant program.

Most of the graphics were made using ggplot (Wickham, 2009), which is great. While its initial appeal to me was in the superficial sense of aesthetic refinement and maturity it lent my plots, in using it more extensively I gained an appreciation for the elegant structural coherence of the grammar of graphics.

***

Although my name is written on the first page of this report (and attached to the wholly unnecessary copyright on the second), the work in this dissertation is the fruit of many years of dutiful effort on the part of too many people to enumerate in this space. But I'll try.

I must begin by acknowledging my advisors Chad Schafer and Peter Freeman; my weekly meetings with Chad and Peter (starting with ADA and continuing through this semester) provided both valuable guidance and a sense of structure and continuity to my time here. Their patient direction allowed me to wrestle with ideas at the depth necessary to expand upon them. I am also grateful for the time and generosity of my other committee members: Ann Lee and Cosma Shalizi, from our department, and Rachel Mandelbaum, from the Department of Physics and the McWilliams Center for Cosmology. Rachel deserves particular recognition for managing to answer not only the simplistic questions I did ask her but also the more sophisticated ones I should have been asking.

While Rachel's input was indispensable in helping to fill in — or, failing that, to paper over — the considerable gaps in my knowledge of physics and astronomy, I am similarly indebted to Melanie Simet, Ying Zu, Tim Eifler, Alina Kiessling, Chieh-An Lin, Benjamin Joachimi, Andrea Petri, and Mike Jarvis (among others) for constructive conversations, in person or via email, about cosmology problems and the standard approaches to solving them.

The past years have afforded me a great deal of intellectual and personal enrichment, for which credit goes to the collegial culture our department has succeeded in fostering. This credit extends to the faculty, the graduate (and undergraduate) student community, and the administrative staff. Through membership in our department community, I have enjoyed pleasant company and made friendships that I hope will persist long after the notion of gleaning inference from data has followed geocentrism and phrenology into quaint obsolescence. These fine people, some of whom may bristle at not being acknowledged individually by name, kept me rooted in those moments when the process appeared interminable. Likewise, the connections I made in Pittsburgh beyond the academic realm helped me to enlarge my perspective and make the most of the opportunities this great city presents.

I must also acknowledge, with immense gratitude, all of my teachers, from elementary school all the way to graduate school, for constructing the foundation that allowed me to pursue research. Their hard work allowed me not only to deepen my grasp of specific content but to refine my understanding of how I might learn further; each step in the process was crucial in laying the groundwork for the next, ultimately preparing me for the excellent instruction I received from the coursework component of this doctoral program.

Finally, to my mom and dad and my brother David: none of my success (or failure) would be possible without your unconditional love, support, and patience over almost thirty years.

Thank you.

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

In recent years, technological developments have greatly increased the availability of computational resources, rendering feasible approaches to statistical inference that were computationally prohibitive in the past. These newly feasible computationally intensive techniques are exemplified by Approximate Bayesian Computation (ABC), a so-called "likelihood-free" approach to inference. Such likelihood-free approaches are advantageous in parametric inference problems where the complexity of the process believed to generate the data may prohibit explicit specification or calculation of the likelihood, but where realistic simulation models exist to produce realizations of that process given parameters. Likelihood-free inference thus replaces the burden of evaluating a likelihood with that of repeated simulation from a forward process.

In simple terms, ABC proceeds according to a trial-and-error scheme: parameter values which yield simulated data similar (in some sense) to the observed data are preferred. However, for high-dimensional data, determining whether observed and simulated data are similar requires some manner of dimension reduction, even in the era of abundant computational resources. Ideally, the summary statistics used in ABC would be (minimally) sufficient; indeed, much of the theoretical guarantees of ABC posterior estimation require that any summarizing be via a sufficient statistic. However, in any situation where a model is so complex that likelihood-free inference is preferable to more standard approaches, it will be difficult or impossible to establish that a statistic is sufficient, let alone minimally sufficient.

While some early ABC analyses chose summary statistics on an *ad hoc* basis, the literature contains a variety of methods for selecting or constructing summary statistics with desired properties. Some such methods involve learning the relationship between corresponding parameters and data using a training set produced via simulation from the forward model; it is to this category that we contribute the *common-space mapping method* (CSMM). The CSMM employs diffusion mapping, which identifies a data set's lower-dimensional structure via the spectral decomposition of a pairwise similarity matrix, to find potentially nonlinear structure in the parameter-data relationship. Our method produces a flexible collection of mappings which can be tuned by optimizing a surrogate for approximate sufficiency. This mapping can then be used to construct ABC summary statistics. We argue that conditioning on a summary statistic optimized thus will, in expectation, minimize the discrepancy between the true posterior and the ABC posterior (in terms of Kullback-Leibler loss).

When the true posterior distribution is analytically known, we measure the performance of a summary statistic by the extent to which the posterior distribution conditioning on that summary statistic resembles the known true posterior distribution. In the setting of simple Gaussian toy examples, we conduct ABC simulation studies in which the performance of the CSMM-derived summary statistic is competitive with that of the minimal sufficient statistic. We comment on the relationship between the CSMM tuning parameter, the associated approximate sufficiency criterion, and the performance of the resulting summary statistic.

We present weak gravitational lensing, a phenomenon which allows for cosmological parameter constraint via the patterns of galaxy shape distortion due to dark matter, as a motivating scientific problem. In weak lensing, we find a natural application for ABC in conjunction with the CSMM approach to dimension reduction: complex underlying processes produce high-dimensional observed data, and the canonical approaches to inference require simplifying assumptions to make likelihood evaluation tractable. We argue that weak lensing parameter inference via ABC permits the relaxation of some simplifying assumptions (albeit at some computational cost and introducing different sources of approximation error). We present the results for both simulated data (where the true input cosmology is known) and for real data taken from the CFHTLens survey.

This report is structured as follows. In Chapter 2, we provides some background and literature review on ABC, concentrating especially on summary statistic choice; we also motivate the need for dimension reduction with a toy example where we encounter a curse of dimensionality. In Chapter 3 we formally specify the common-space mapping method and introduce a criterion useful for tuning the method. In the next two chapters we present ABC analyses using the CSMM-derived summary statistics: Chapter 4 explores toy example problems wherein a ground truth is known, while Chapter 5 provides proof-of-concept analyses of weak lensing data (both simulated and real). Chapter 6 offers a brief guide to software that implements the CSMM. Chapter 7 concludes with ideas for future development, as well as a brief discussion of the advantages and limitations of the present work.

Chapter 2

Background and motivation

2.1 Overview of ABC

2.1.1 Likelihood-free inference

Likelihood-free methods allow for parameter inference in settings where it is difficult or impossible to evaluate a likelihood function for a given set of parameter values but in which a process is available to generate data using those parameter values as inputs. Such situations may arise when the underlying processes or sources of uncertainty are too complex to permit a formal mathematical specification of the likelihood; this includes the case described by Blum (2012) where the data-generating process includes a latent high-dimensional variable which would have to be marginalized out of the ultimate likelihood.

In this work, we concentrate on approximate Bayesian computation (ABC), a likelihood-free Bayesian approach that was first applied by Pritchard et al. (1999) in the computational biology literature, although an earlier thought experiment of Rubin (1984) laid out its conceptual underpinning. We give here a brief explanation of ABC, deferring a more comprehensive exposition to the literature (see, e.g., Marin et al., 2012).

Consider the setting where data $x \in \mathbb{R}^d$ is generated from a model parameterized by θ; suppose that a density for x is given by $f(x|\theta)$. Having observed x, the Bayesian approach to inference in this setting is to assume a prior distribution $\pi(\theta)$ for the parameter and explore the posterior distribution $\pi(\theta|x)$ of parameters conditional on observed data:

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{f(x)} = \frac{f(x|\theta)\pi(\theta)}{\int f(x|\theta)\pi(\theta)d\theta} \propto f(x|\theta)\pi(\theta) \tag{2.1}$$

Direct evaluation of the $\pi(\theta|x)$ would require a means of evaluating both the likelihood $f(x|\theta)$ and the marginal distribution of $f(x)$. The broad class of methods referred to as Markov Chain Monte Carlo (MCMC) permit sampling from the posterior distribution, provided that the likelihood $f(x|\theta)$ can be evaluated. By contrast, ABC does not require the evaluation of $f(x|\theta)$; thus, it can be applied in settings where it is feasible to simulate a realization of x given θ, but where evaluating $f(x|\theta)$ is analytically or computationally intractable.

Since its introduction, likelihood-free analyses have been performed in domains as diverse as population genetics (Beaumont et al., 2009; Csilléry et al., 2010), dynamic biological systems (Toni et al., 2009), and astronomy (Cameron & Pettitt, 2012; Weyant et al., 2013). Our application of ABC to weak gravitational lensing finds a recent precedent in the work of Lin & Kilbinger (2015), which uses ABC to obtain parameter constraints from weak lensing peak counts. Cosmology problems, which are generally structured as parametric inference for models with relatively few parameters, provide a very natural application setting for ABC. Bayesian approaches already enjoy wide usage in the field. Additionally, complex forward simulation models incorporate sources of uncertainty for which it would be difficult to write down a likelihood.

We remark that the use of likelihood-free methods is not restricted to the framework of Bayesian inference. In one frequentist approach known as indirect inference (Gourieroux et al., 1993), first developed for estimating the parameters in econometric time series models, a simulation mechanism is used to produce multiple mock data sets from each candidate parameter value θ. The simulated data are then summarized according to the estimated parameters of an auxiliary model, using as a point estimate for θ that value which produced the simulated data set most similar to the observed data, and appealing to asymptotic normality arguments for uncertainty (see also Gourieroux & Monfort, 1997). Other recent work (Wood, 2010) introduces the *synthetic likelihood*, which is the likelihood calculated assuming that a chosen set of summary statistics has a multivariate normal distribution; the resulting synthetic likelihood can be part of a frequentist likelihood calculation or a Bayesian MCMC analysis.

2.1.2 ABC rejection algorithm

Suppose that we that we observe data x_{obs} and that a forward model, M_θ, is available to simulate data given a value for θ. The simplest ABC algorithm is as follows:

Algorithm 2.1 Basic ABC Rejection Algorithm

```
1: for $i = 1$ to $N$ do
2:     Draw $\tau$ from $\pi$
3:     Simulate $y$ from $M_\tau$
4:     if $y = x_{obs}$ then
5:         Retain $\tau$
6:     end if
7: end for
```

The resulting sample consists of N retained τ values drawn from the posterior distribution $\pi(\theta|x_{obs})$. However, in most cases of application, the event that $y = x_{obs}$ exactly is an event of probability zero, so no values of τ will be retained. Instead, comparisons between observed and simulated data are usually made using some summary statistic of the data, $S(\cdot) : \mathbb{R}^d \to \mathbb{R}^{d'}$. This requires modifying lines 4-6 of Algorithm 2.1 as follows:

```
if $S(y) = S(x_{obs})$ then
    Retain $\tau$
end if
```

After this modification, the resulting sample of retained τ values will have distribution $\pi\left(\theta|S(x) = S(x_{obs})\right)$. Note that if $S(\cdot)$ is sufficient for θ, then $\pi\left(\theta|S(x) = S(x_{obs})\right)$ is equal to $\pi(\theta|x_{obs})$. If $S(\cdot)$ is not sufficient for θ, then the distribution of the retained τ values will be an approximation to the desired posterior. (In Section 3.2.1, we consider the implications of conditioning on non-sufficient statistics on estimating the posterior distribution.)

However, for any continuously-distributed $S(x)$, the probability that $S(y) = S(x_{obs})$ is zero. Hence, we modify the algorithm again, retaining τ if the difference between $S(y)$ and $S(x_{obs})$ is small in some metric $\rho : \mathbb{R}^{d'} \times \mathbb{R}^{d'} \to \mathbb{R}$. The algorithm, modified to incorporate these changes, is given in Algorithm 2.2.

The distribution of the resulting sample of retained τ values is now $\pi_\epsilon\left(\theta|\rho(S(x), S(x_{obs})) \leq \epsilon\right)$, where the subscript ϵ refers to the tolerance up to which the summarized observed and simu-

Algorithm 2.2 Basic ABC Rejection Algorithm, Modified

```
for $i = 1$ to $N$ do
    Draw $\tau$ from $\pi$
    Simulate $y$ from $M_\tau$
    if $\rho\left(S(y), S(x_{obs})\right) \leq \epsilon$ then
        Retain $\tau$
    end if
end for
```

lated data may differ. If and only if $\epsilon = 0$ and $S(\cdot)$ is sufficient for θ, then $\pi_\epsilon\left(\theta | S(x) = S(x_{obs})\right)$ is equal to the desired $\pi(\theta|x_{obs})$. By contrast, as ϵ grows arbitrarily large, all τ values are retained, so the resulting sample has distribution equal to the prior $\pi(\theta)$.

In practice, the choice of ϵ involves a tradeoff. Taking ϵ too large will make $\pi_\epsilon\left(\theta | S(x) \approx S(x_{obs})\right)$ a poor approximation to $\pi(\theta|x_{obs})$; on the other hand, taking ϵ too small will either result in a sample too small to be of inferential value, or require N, the number of candidate (τ, y) pairs generated, to be prohibitively large.

Indeed, using any positive value of ϵ in an ABC procedure introduces some approximation error to the resulting estimate of $\pi(\theta|x_{obs})$. Hence, standard practice is to use the smallest computationally feasible value of ϵ in order to neutralize this source of approximation error, though some approaches argue that an ϵ value of zero is not in fact ideal: Wilkinson (2013) demonstrates that for a value of ϵ consistent with the measurement error in the data, the resulting ABC posterior is exact, while the so-called *noisy ABC* approach of Fearnhead & Prangle (2012) advocates adding noise (related to the ABC tolerance) to the summarized data so that the posterior will satisfy a *calibration* property.

Additional error in approximating $\pi(\theta|x_{obs})$ can result from using a summary statistic $S(\cdot)$ that is not sufficient for θ, as well as the so-called *Monte Carlo error* due to estimating a distribution from a finite sample.

2.1.3 ABC variants

Since basic rejection ABC was first introduced, many variants have been developed to improve particular aspects of the ABC procedure. Some authors (e.g., Blum et al., 2013) generalize the formulation of ABC so that instead of simply accepting or rejecting τ, a weight is assigned by a smoothing kernel function $K\left(\rho(S(y), S(x_{obs}))\right)$. (Indeed, taking the rectangular kernel $K_\epsilon\left(\rho(S(y), S(x_{obs}))\right) = \frac{1}{2}\mathbb{I}_{\{\rho(S(y),S(x_{obs}))\leq\epsilon\}}$ recovers Algorithm 2.2.)

Other variants incorporate the ABC idea within an iterative parameter estimation framework. One such approach of Marjoram et al. (2003) modifies the traditional MCMC procedure by replacing a likelihood calculation with a data simulation and comparison step; however, this so-called ABC-MCMC procedure (as with standard MCMC) may require a burn-in period, and the resulting chains must be 'thinned out' to account for the inherent correlation between consecutive draws. The ABC-PRC (Sisson et al., 2007) and ABC SMC (Toni et al., 2009) algorithms are iterative particle methods wherein the tolerance ϵ is decreased at each time step.

We implement a version of ABC SMC (specified in Algorithm 2.3) for the simulation studies of Chapters 4 and 5. At each time step t, new particles are updated via weighted resampling from the old set of particles (with replacement) and perturbation via a Gaussian kernel. (Note that in the specification of Algorithm 2.3, superscripts refer to time steps rather than exponentials.) In our implementation, we simulate data with those perturbed parameter values as inputs, retaining only those closer (in the space of summary statistics) than some ϵ_t, which we allow to decrease as t increases. We also shrink the bandwidth of the perturbation kernel by a multiplicative factor of $\alpha = 0.9$ at each time step.

These iterative variants of ABC improve upon the basic rejection algorithm by exploring the parameter space in ways more efficient than simple repeated sampling from the prior. However, many of the issues (discussed in subsequent sections of this report) that arise in the context of simple rejection ABC persist even in the context of the more efficient variants of ABC.

2.1.4 Choice of summary statistic

As mentioned above, comparing observed and simulated data using a non-sufficient summary statistic contributes somewhat to the approximation error of an ABC-derived posterior estimate. This is unfortunate, as ABC finds its most natural application in the setting of an intractable likelihood, where determining that a statistic is sufficient might be difficult

Algorithm 2.3 ABC SMC Algorithm

```
Initialize tolerance $\epsilon^1$ and perturbation bandwidth $\sigma^1$
for $t = 1$ to $T$ do
    for $i = 1$ to $N$ do
        while $\rho\left(S(y_i^t), S(x_{obs})\right) > \epsilon^t$ do
            if $t = 1$ then
                Draw $\tau_i^t$ from $\pi$
            else
                Select $\tau_i^t$ from $\tau_1^{t-1}, \ldots, \tau_N^{t-1}$ with probabilities given by $w_i^{t-1}$
                Perturb $\tau_i^t$ by adding noise drawn from Gaussian kernel $K_{\sigma^t}(\cdot)$
            end if
            Draw $y_i^t$ from $M_{\tau_i^t}$
        end while
        Retain $\tau_i^t$
        Set distance $d_i^t = \rho\left(S(y_i^t), S(x_{obs})\right)$
        Set weight $w_i^t = \frac{\pi(\tau_i^t)}{\sum_{j=1}^N w_j^{t-1} K_{\sigma^t}(\tau_i^t - \tau_j^{t-1})}$
    end for
    Normalize the weights $w_i^t$
    Set $\epsilon^{t+1} = \text{median}(d_1^t, \ldots, d_N^t)$ and $\sigma^{t+1} = \alpha\sigma^t$
end for
```

or impossible.

Barber et al. (2013) establishes convergence rates for ABC posterior estimation which require $S(\cdot)$ to be sufficient, noting only that “an additional error will be introduced” otherwise. In general, the literature regarding the effect of the choice of summary statistic on estimation is not extensive. Frazier et al. (2015) derive conditions on the summary statistic — including to the one-to-one-ness of the mapping $\theta \rightarrow \mathbf{b}(\theta)$, where $\mathbf{b}(\theta)$ is the limit of the summary statistic based on data generated from θ — under which ABC methods achieve Bayesian consistency. The same paper develops diagnostic criteria to assess whether a summary statistic will meet these conditions and demonstrate that, in some previous studies where ABC is applied, those conditions are not met.

Numerous approaches to choosing an ABC summary statistic exist in the literature; we present a cursory overview of those here, again deferring to reviews (Blum et al., 2013; Prangle, 2015) for a comprehensive treatment. Given a pool of potential summary statistics, Joyce & Marjoram (2008) propose a scheme for scoring a candidate statistic for inclusion based on the variability in the distribution of the candidate statistic conditioned on both the data and those statistics already included. A similar approach (Nunes & Balding, 2010) involves minimizing the empirically estimated entropy over the set of possible summary statistic combinations. However, these techniques have limited value in the case where all potential summary statistics cannot be enumerated simply. Ruli et al. (2015) argues that the score function of the likelihood would be a well-behaved ABC summary statistic and suggests an approach wherein the score function of a composite likelihood is used as a proxy for that of the true likelihood.

By contrast, other approaches involve constructing statistics believed to be approximately sufficient by learning a projection from a simulated training set. One such strategy, the so-called *semi-automatic ABC* approach of Fearnhead & Prangle (2012) attempts to construct summary statistics by simulating a training set of parameters and data and regressing parameter values on data. A similar method (Jiang et al., 2015) involves training a multilevel neural network, using simulated data as the input and corresponding parameter values as the target; the fitted target values that result, which approximate the posterior mean $\mathbb{E}[\theta|x]$, are used as the ABC summary statistic.

Another strategy (Drovandi et al., 2011) demonstrates the link between ABC and the frequentist likelihood-free approach known as indirect inference, in the sense that the estimated parameters of an auxiliary model are used as ABC summary statistics. Related work (Drovandi et al., 2015) suggests that "the summary statistic produced by ABC II should carry most of the information contained in the observed data provided that the auxiliary model provides a good description of the data." However, the authors caution that this intuition is difficult to make formal and suggest that "goodness-of-fit tests and/or residual analysis on the auxiliary model fit to the data will at least provide some guidance on the usefulness of the summary statistic produced by ABC II."

2.2 Toy example: a curse of dimensionality

We consider the issues that arise in a simple toy example in order to motivate the development of methods to construct approximately sufficient statistics of low dimension.

We recall first that in standard Bayesian inference, the posterior distribution $\pi(\theta|x)$ is proportional to $\pi(\theta)f(x|\theta)$, where $f(x|\theta)$ is also the likelihood $\mathcal{L}(\theta;x)$. By contrast, the probability of a parameter value τ appearing in an ABC posterior sample is

$$\pi_{ABC}(\tau|x) = \pi(\tau)\,\mathbb{P}(\text{generate some } y \sim M_\tau \text{ such that } \rho(S(y),S(x)) \leq \epsilon|x)$$

Consequently, in an ABC scheme, it would be desirable for the rightmost term above, i.e., the probability of retaining a candidate parameter value τ, to be proportional to the likelihood $\mathcal{L}(\tau;x)$. In the example that follows, we will assess the performance of summary statistics by quantifying how well the probability of retaining τ in an ABC analysis using those summary statistics (suitably normalized) approximates the likelihood of τ.

Now, suppose that data $X_1,\ldots,X_n \overset{IID}{\sim} \mathcal{N}(\mu,1)$, where μ is unknown. We will use ABC to explore the posterior distribution $\pi(\mu|X_1,\ldots,X_n)$. Recall that performing an ABC analysis requires a choice of distance metric ρ, summary statistic $S(\cdot)$, and tolerance ϵ. For simplicity, and lacking motivation for any other form, we will use the Euclidean norm for ρ in the ABC procedure.

For $S(\cdot)$, we will consider what happens under various extents of summarizing the observed and simulated data sets. One extreme is no summarizing, i.e., $S(X_1,\ldots,X_n) = (X_1,\ldots,X_n)$. Another extreme is summarizing the entire data set by its mean, $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Both extremes can be shown to be sufficient for μ; the former is trivially so, while the latter is a basic result in elementary statistics. In between these extremes there are intermediate extents of summarizing. It is also possible to summarize the data too much, in the sense of using a non-sufficient summary statistic (or, put another way), throwing away information. We will now consider these extents of summarizing formally.

2.2.1 Summarizing too coarsely

Assuming for the sake of argument that $n = 2^\ell$ for some integer ℓ, we will define tiers of summary statistics at each of $\ell + 1$ levels: 0, 1, …, ℓ. We will use the notation

$$\overline{X}_k^j = \frac{1}{2^{\ell-j}} \sum_{i=2^{\ell-j}(k-1)+1}^{2^{\ell-j}k} X_i$$

Our approach is to partition the data set $X_1, \ldots, X_n$ into chunks such that at level j there will be 2^j chunks, each of size $2^{\ell-j}$. Then, the summary statistics $\overline{X}_k^j$ will be the means of the data in each chunk. (Indeed, $\overline{X}_1^0$ will just be our familiar $\overline{X}$.) A graphical representation of these tiers of summary statistic is given in Figure 2.1.

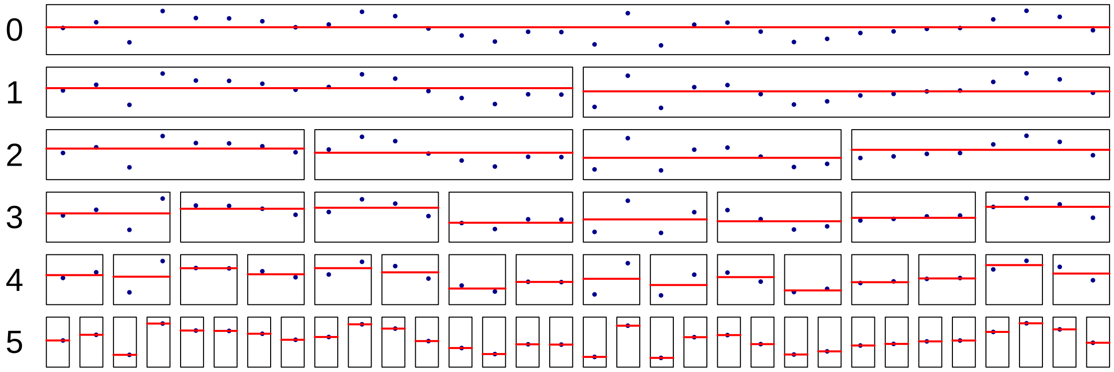

Figure 2.1: For $X_1, \ldots, X_{\ell}$ distributed $\mathcal{N}(\mu, 1)$, each row represents a tier of summary statistics $\overline{X}_1^j, \ldots \overline{X}_{2^j}^j$ (in red) that are jointly sufficient for μ.

At any level j, the collection $\overline{X}_1^j, \ldots \overline{X}_{2^j}^j$ is jointly sufficient for μ and thus is nominally adequate for use in an ABC analysis. Intuition suggests that sufficient statistics of lower dimension should be preferred. We demonstrate first that a lower dimensional summary statistic requires a lower tolerance ϵ to achieve the same probability (e.g., 0.25) of accepting the true μ value. We then attempt to assess the relative usefulness of the tiers of sufficient statistics by calculating, for each tier, what probability of retaining the true μ is needed so that the retention probabilities $p_j(\tau)$ approximate the likelihood of τ with the same accuracy as is obtained when $\overline{X}$, the minimal sufficient statistic, is used with a 0.25 probability of retaining the true μ value.

To this end, suppose we are in the setting where the true value of μ is 0; we have already observed $X_1, \ldots, X_n$, which were distributed $\mathcal{N}(0,1)$, and τ has already been drawn from the prior distribution to also be equal to 0, which is the true μ. Thus, we will not consider the contribution of any uncertainty due to the prior draw because the prior has, so to speak, guessed the true μ correctly. We may first ask: once the true μ has been drawn from the prior, how strict a tolerance will be required (for each tier of summary statistics) to retain that true μ with probability 0.25?

Proposition 2.2.1. *At level j, the ϵ_j^* needed so that the true μ will be retained with probability 0.25 is given by*

$$\epsilon_j^* = \sqrt{\frac{2^j}{n} F^{-1}_{2^j,\lambda}(0.25)} \tag{2.2}$$

where $F^{-1}_{2^j,\lambda}$ is the inverse CDF for the noncentral χ^2 distribution with 2^j degrees of freedom and noncentrality parameter $\lambda = \frac{n}{2^j}\sum_{i=1}^{2^j}(\overline{X}_k^j)^2$.

Proof. Having fixed $\tau = 0$, each $Y_i \sim \mathcal{N}(0,1)$. Thus, each $\overline{Y}_k^j$ is the mean of $2^{\ell-j}$ independent $\mathcal{N}(0,1)$ random variables, so it has a $\mathcal{N}(0, 2^{j-\ell})$ distribution. Since the X_i are fixed (and consequently, so are the $\overline{X}_k^j$), the quantity $\overline{X}_k^j - \overline{Y}_k^j$ has a $\mathcal{N}(\overline{X}_k^j, 2^{j-\ell})$ distribution.

The sum of squares of normally distributed random variables divided by their variances has a noncentral χ^2 distribution. Generally, if $U_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$, then $\sum_{i=1}^p \left(\frac{U_i}{\sigma_i}\right)^2$ has a χ^2 distribution with p degrees of freedom and noncentrality parameter $\lambda = \sum_{i=1}^p \left(\frac{\mu_i}{\sigma_i}\right)^2$. Hence, the quantity

$$\frac{n}{2^j}\sum_{k=1}^{2^j}(\overline{X}_k^j - \overline{Y}_k^j)^2 \tag{2.3}$$

has a χ^2 distribution with 2^j degrees of freedom and noncentrality parameter $\lambda = \frac{n}{2^j}\sum_{i=1}^{2^j}(\overline{X}_k^j)^2$.

Suppose now we want to calculate the probability of retaining the candidate τ which is equal to the true μ. We will retain τ if and only if

$$\|S(X_1,\dots,X_n)-S(Y_1,\dots,Y_n)\|_2\leq\epsilon$$

for some chosen ϵ.

Hence, as a function of ϵ, the probability of retaining τ is

$$p(\epsilon)=\mathbb{P}(\|S(X_1,\dots,X_n)-S(Y_1,\dots,Y_n)\|_2\leq\epsilon)$$

Summarizing at level j, we have that $S(X_1,\dots,X_n)=(\overline{X}_1^j,\dots,\overline{X}_{2^j}^j)$, and analogously for $S(Y_1,\dots,Y_n)$, meaning that

$$\begin{aligned}
p_j(\epsilon)&=\mathbb{P}\left(\sqrt{\sum_{k=1}^{2^j}(\overline{X}_k^j-\overline{Y}_k^j)^2}\leq\epsilon\right)\text{ or equivalently,}\\
p_j(\epsilon)&=\mathbb{P}\left(\sum_{k=1}^{2^j}(\overline{X}_k^j-\overline{Y}_k^j)^2\leq\epsilon^2\right)\text{, or finally,}\\
p_j(\epsilon)&=\mathbb{P}\left(\frac{n}{2^j}\sum_{k=1}^{2^j}(\overline{X}_k^j-\overline{Y}_k^j)^2\leq\frac{n}{2^j}\epsilon^2\right). \qquad (2.4)
\end{aligned}$$

We argued earlier that the quantity on the right-hand side of the inequality has a χ^2 distribution with 2^j degrees of freedom and noncentrality parameter $\lambda=\frac{n}{2^j}\sum_{i=1}^{2^j}(\overline{X}_k^j)^2$. Hence, if $F_{2^j,\lambda}$ is the distribution function for that χ^2 distribution, then

$$p_j(\epsilon)=F_{2^j,\lambda}(\frac{n}{2^{j+1}}\epsilon^2).$$

If we desire the tolerance ϵ_j^* that allows us to retain the true parameter value, in expectation, say, 25% of the time, it is straightforward to solve

$$\begin{aligned}
F_{2^j,\lambda}(\frac{n\epsilon_j^{*2}}{2^j})&=0.25\text{, or}\\
\epsilon_j^*&=\sqrt{\frac{2^j}{n}F_{2^j,\lambda}^{-1}(0.25)}
\end{aligned}$$

for each of $j = 0, 1, \ldots, \ell$. □

Because the noncentrality parameter λ depends on the data $X_1, \ldots, X_n$, the needed tolerances ϵ_j^* (for any j) will vary for any particular instance of the data. For one such instance, with $n = 32$, the needed ϵ_j^* are given in Table 2.1.

dimension of $S(\cdot)$	1	2	4	8	16	32
ϵ_j^* needed	0.0622	0.2017	0.5715	1.2945	2.8561	6.4198

Table 2.1: Tolerances needed for 25% retention of the true parameter value for one particular instance of observed data $X_1, \ldots, X_{32}$.

Now that we have calculated ϵ_j^* for $j = 1, \ldots, \ell$, we can fix ϵ_j^* at each level and consider the retention probability $p_j(\tau)$ as τ varies to depart from μ. By arguments very similar to the justification given in the preceding proof, the retention probability as a function of τ is given by

$$p_j(\tau) = F_{2^j, \lambda_j(\tau)}\left(\frac{n}{2^j}{\epsilon_j^*}^2\right)$$

where $F_{2^j, \lambda_j(\tau)}$ is the CDF for a $\chi^2_{2^j, \lambda_j(\tau)}$ random variable, and $\lambda_j(\tau) = \frac{n}{2^j}\sum_{i=1}^{2^j}(\overline{X}_k^j - \tau)^2$. It is then straightforward to evaluate this function on a grid of τ values.

At left in Figure 2.2, we compare, for various extents of summary, $p_j(\tau)$ to the true likelihood $\mathcal{L}(\tau; X_1, \ldots, X_{32})$ – i.e., the probability of $X_1, \ldots, X_{32}$ having been drawn from a $\mathcal{N}(\tau, 1)$ model – with all functions normalized to integrate to 1 for the sake of comparison. We see that, when $n = 32$ and the tolerance is fixed so that the true μ is retained with probability 0.25, summarizing via the minimal sufficient statistic $\overline{X}$ (red curve) will retain τ almost exactly in proportion to $\mathcal{L}(\tau; X_1, \ldots, X_{32})$ (black dotted curve).

By contrast, summarizing via statistics that are sufficient but not minimal (using the same 0.25 level of retaining the true μ) will not retain τ proportionally to $\mathcal{L}(\tau; X_1, \ldots, X_{32})$; we see that this behavior gets worse as the extent of grows more coarse. If the retention probability disagrees with the likelihood, then the distribution $\pi_\epsilon(\mu|X_1, \ldots, X_n)$ of the resulting ABC sample will be a poor approximation to $\pi(\mu|X_1, \ldots, X_n)$.

In principle, as long as the summary statistic chosen is sufficient for μ, we can guarantee that $\pi_\epsilon(\mu|X_1,\ldots,X_n)$ is an adequate approximation to $\pi(\mu|X_1,\ldots,X_n)$ simply by choosing ϵ small enough. However, this asymptotic guarantee is of little practical use on its own; requiring a too-small value of ϵ may render the probability of retaining any candidate τ so low that the number of candidates needed to retain a reasonable number may grow infeasibly large.

We can quantify this by approaching the problem from a slightly different angle. Until now, we have fixed the probability of retaining the true μ at 0.25 and compared different extents of summary j by finding the ϵ_j^* for each that achieves that probability; this results in retention behavior for each j that varies in quality as a proxy for the true likelihood. As an alternative, we can quantify how well the retention behavior using $\overline{X}$ matches the desired likelihood and then, for each other summary level j, calculate what value of ϵ would be needed to achieve that same quality of approximation.

Using the notation that $p_j(\tau,\epsilon)$ is the probability of retaining τ at summary level j and using tolerance ϵ, we quantify the quality of approximation via the integrated squared difference between the true likelihood and the normalized retention probability, given by

$$\text{ISE}(j,\epsilon) = \int_{-\infty}^{\infty} (\mathcal{L}(\tau; X_1,\ldots,X_n) - p_j(\tau,\epsilon))^2 d\tau$$

For computational purposes, we approximate the above via a discrete sum over plausible τ values. For the same realization of $X_1, ...X_{32}$ we have cited throughout, $\text{ISE}(0,\epsilon_0^*) = 0.000495$, where ϵ_0^* is the value for ϵ_0 that yields probability 0.25 of retaining the true μ. For each other j, we can compute (via numerical optimization) the value $\epsilon_j^{L_2}$ that achieves that same ISE. That $\epsilon_j^{L_2}$, in turn, dictates a probability of retaining the true μ, $p(0,\epsilon_j^{L_2})$. These probabilities are given in Table 2.2.

dimension of $S(\cdot)$	1	2	4	8	16	32
ϵ needed	0.0622	0.1014	0.1755	0.3197	0.6061	1.1809
$\mathbb{P}$(0 is retained)	0.25	0.07	0.00375	1.94e-05	3.96e-10	1.83e-20

Table 2.2: Tolerance values and associated retention probability at the true μ

From these results, we take away the message that if we use the full data set as our summary statistic, in order to achieve roughly the same quality of approximation as we would achieve by using the overall mean, the tolerance ϵ would need to be such that even the true parameter value is retained with probability less than 2×10^{-20}. We remark that the ϵ values from Table 2.2 are not directly comparable to each other because of the changing dimension.

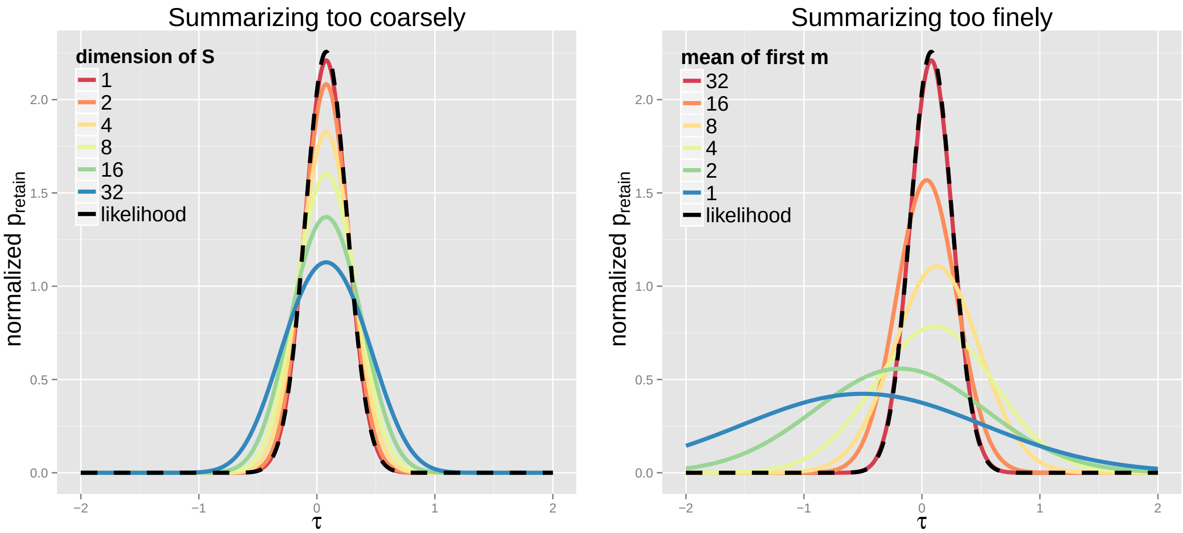

Figure 2.2: For one particular realization of data $X_1, \ldots, X_{32}$, the normalized probability of retaining τ — ideally proportional to $\mathcal{L}(\tau; X_1, \ldots, X_{32})$ — for ABC summary statistics in the case of summarizing too coarsely (left) and too finely (right).

2.2.2 Summarizing too finely

We can also consider what happens if we summarize the data too finely, in the sense that the summary statistic is no longer sufficient for the parameter; in other words, relevant information is being discarded.

We reiterate that the mean of all n observations, $\overline{X}$, is known to be the minimal sufficient statistic for μ. Define the partial mean $\overline{X}_m$ as the mean of the first m observations, i.e.,

$$\overline{X}_m = \frac{1}{m} \sum_{i=1}^{m} X_i, \text{ for } m \leq n.$$

Note that $\overline{X}_n$ is just our familiar $\overline{X}$, while $\overline{X}_1$ is just X_1 by itself.

Proposition 2.2.2. *Suppose that we have already observed the data $X_1, \ldots, X_n$ so that they are not considered to be random, and further that we are in the setting where we have drawn $\tau = 0$ (the true value of μ) from a prior distribution. Then the tolerance ϵ_m^* needed to achieve a retention probability of 0.25 can be calculated as*

$$\epsilon_m^* = \sqrt{\frac{1}{m} F_{1,\lambda_m}^{-1}(0.25)}$$

where F_{1,λ_m} is the CDF for the χ^2 distribution with one degree of freedom and noncentrality parameter $\lambda_m = m(\overline{X}_m)^2$.

Proof. Each Y_i is drawn from a $\mathcal{N}(0,1)$ distribution. Defining $\overline{Y}_m$ analogously to $\overline{X}_m$ as the mean of the first m observations,

$$\begin{aligned} \overline{Y}_m &\sim \mathcal{N}(0, \frac{1}{m}) \\ (\overline{X}_m - \overline{Y}_m) &\sim \mathcal{N}(\overline{X}_m, \frac{1}{m}) \\ m(\overline{X}_m - \overline{Y}_m)^2 &\sim \chi^2_{1,\lambda_m} \text{ where } \lambda_m = m(\overline{X}_m)^2. \end{aligned}$$

Denote by F_{1,λ_m} the CDF of a χ^2_{1,λ_m} distribution. Then the probability of retaining 0, the true parameter value when using a tolerance ϵ is given by

$$p(0,\epsilon) = \mathbb{P}(\sqrt{(\overline{X}_m - \overline{Y}_m)^2} \leq \epsilon) = \mathbb{P}(m(\overline{X}_m - \overline{Y}_m)^2 \leq m\epsilon^2) = F_{1,\lambda_m}(m\epsilon^2).$$

We can invert this relationship to calculate the ϵ_m^* needed to achieve a desired acceptance probability of the true μ, e.g., 0.25:

$$\epsilon_m^* = \sqrt{\frac{1}{m} F_{1,\lambda_m}^{-1}(0.25)}.$$

□

These ϵ_m^* values, for the same specific realization of the data $X_1, \ldots X_{32}$ as was used earlier, are given in Table 2.3.

m	32	16	8	4	2	1
ϵ_m^* needed	0.0622	0.0807	0.1200	0.1632	0.2332	0.3612

Table 2.3: Tolerances needed for 25% retention of the true parameter value when summarizing too much, i.e., keeping only $X_1, \ldots, X_m$, for a particular realization of the data $X_1, \ldots, X_{32}$.

Similarly to the case of summarizing too coarsely, we can now vary τ and use these ϵ_m^* values to calculate the probability of retaining τ for different values of m. When $\tau \neq 0$, $m(\overline{X}_m - \overline{Y}_m)^2 \sim \chi^2_{1,\lambda_m(\tau)}$, where now $\lambda_m(\tau) = m(\overline{X}_m - \tau)^2$. Hence,

$$p_m(\tau) = \mathbb{P}(\sqrt{(\overline{X}_m - \overline{Y}_m)^2} \leq \epsilon_m^*) = \mathbb{P}(m(\overline{X}_m - \overline{Y}_m)^2 \leq m{\epsilon_m^*}^2) = F_{1,\lambda_m(\tau)}(m{\epsilon_m^*}^2). \quad (2.5)$$

Again, similarly to the preceding section, we can evaluate this distribution function on a grid of τ values. Recall that in order for the distribution of the ABC sample to successfully approximate the true posterior distribution, the retention probability at τ should be roughly proportional to $\mathcal{L}(\tau; X_1, \ldots, X_{32})$. These retention probabilities, as a function of τ and normalized to integrate to 1 for the sake of comparison, are displayed at right in Figure 2.2.

Note that the summary statistics $\overline{X}_m$ for $m < n$ are not sufficient for μ, so in this case, there is no asymptotic guarantee that the true posterior distribution can be approximated to arbitrarily high accuracy simply by taking ϵ very small.

In Section 4.1, we examine the quality of ABC-derived posterior approximations when using the summary statistics described above (as well as those resulting from a method we introduce in Chapter 3) using data simulated according to this toy model. These simulation results are consistent with our understanding that the minimal sufficient statistic is most preferred as an ABC summary statistic; summary statistics that are not minimal result in slightly worse ABC posteriors, while summary statistics that are not sufficient result in far worse ABC posteriors.

Chapter 3

Common space mapping method

3.1 Method introduction

Motivated by the desire for summary statistics that are both sufficient and of low dimension, we introduce a method which seeks to learn, from a training set, embeddings of both parameters and data into the same low-dimensional space. Heuristically, these mappings will capture the information from the training set in the data relevant for variation among the parameters, and vice versa. We propose that the mapping from the original data space to the shared lower-dimensional space will be an acceptable choice of ABC summary statistic, as it will have properties resembling sufficiency and will have user-controlled dimension. In what follows, we make these ideas formal.

3.1.1 Starting point: model structure assumption

Suppose the parameter of interest $\theta \in \mathbb{R}^p$ admits a low-dimensional representation, i.e., the mapping

$$\eta = (\eta_1(\theta), \eta_2(\theta), \ldots, \eta_J(\theta))$$

from $\mathbb{R}^p \to \mathbb{R}^J$ is (approximately, at least) one-to-one, in the sense that given η, it is possible to reconstruct (approximately) θ. This would be the case if θ is of low dimension in its native form, or if θ is of high dimension but has inherently low-dimensional structure, as is often the case in the cosmological inference problems of interest.

The representation η is certainly not unique. We propose that if we search over a sufficiently wide class of such η, it will be possible to find corresponding mappings on the data space T_j such that an adequately-fitting model of the form

$$T_j(X) = \eta_j(\theta) + \epsilon_j, \tag{3.1}$$

where ϵ_j are i.i.d. standard normal, can be found. At first glance, this may seem to be a particularly restrictive form, but with sufficient flexibility in the class of functions, it is hoped that an appropriate one will be found.

Assuming the structure of Equation (3.1), the conditional density $f\left(T(X)|\eta(\theta)\right)$ is given by

$$\begin{aligned}
f\left(T(x)|\eta(\theta)\right) &= \left(\frac{1}{\sqrt{2\pi}}\right)^J \exp\left(\sum_{j=1}^{J} -\frac{1}{2}\left(T_j(x) - \eta_j(\theta)\right)^2\right) \\
f\left(T(x)|\eta(\theta)\right) &= \left(\frac{1}{\sqrt{2\pi}}\right)^J \exp\left(\sum_{j=1}^{J} -\frac{1}{2}\left[(T_j(x))^2 - 2T_j(x)\eta_j(\theta) + (\eta_j(\theta))^2\right]\right) \\
\log f\left(T(x)|\eta(\theta)\right) &= \sum_{j=1}^{J}\left[T_j(x)\eta_j(\theta) - \frac{1}{2}(T_j(x))^2 - \frac{1}{2}(\eta_j(\theta))^2 - \frac{1}{2}\log(2\pi)\right]
\end{aligned}$$

Then, if we assume that θ is distributed according to some prior distribution $\pi(\theta)$, then the joint log-likelihood of a transformed data-parameter pair would be given by

$$\sum_{j=1}^{J}\left[T_j(x_i)\eta_j(\theta_i) - \frac{1}{2}(T_j(x_i))^2 - \frac{1}{2}(\eta_j(\theta_i))^2 - \frac{1}{2}\log(2\pi)\right] + log(\pi(\theta_i)). \tag{3.2}$$

Taking the derivative of this expression with respect to $T_j(x_i)$ and setting equal to zero, we find that the joint log likelihood $\log f(T(x_i), \eta(\theta_i))$ is maximized when $T_j(x_i) = \eta_j(\theta_i)$. This suggests that desirable mappings $\widehat{\eta}_j(\theta)$ and $\widehat{T}_j(x)$ would be such that when θ and x correspond, $\widehat{\eta}_j$ and $\widehat{T}_j$ should (respectively) map them close together. According to the structure we have outlined, T_j and η_j share a codomain; thus, we refer to our attempt to learn optimal $\widehat{\eta}_j$ and $\widehat{T}_j$ mappings as the *common-space mapping method.*

3.1.2 Obtaining common-space mappings via diffusion maps

Given that our aim is to construct summary statistics for use in ABC, it is safe to assume that can relatively easily simulate parameter values from the prior distribution and realize data from a forward model under those parameter values. We can thus build a training set of N_T pairs (θ_i, x_i) which will inform our understanding of the parameter-data relationship. The step of constructing a training step has parallels in the approaches of Jiang et al. (2015) and the *semi-automatic ABC* of Fearnhead & Prangle (2012); we acknowledge that for any such approach, the informativeness of the training set is limited by the extent to which the model used to build it successfully embodies the data-generating phenomena.

Under the structural assumption in (3.1), the joint log-likelihood over the training set would be the sum of the contributions of (3.2) from each (θ_i, x_i) pair:

$$\sum_{i=1}^{N_T} \left[\sum_{j=1}^{J} \left[T_j(x_i)\eta_j(\theta_i) - \frac{1}{2}(T_j(x_i))^2 - \frac{1}{2}(\eta_j(\theta_i))^2 - \frac{1}{2}\log(2\pi) \right] + log(\pi(\theta_i)) \right]. \tag{3.3}$$

As the contributions from each (θ_i, x_i), pair are maximized by mappings where $T_j(x_i) = \eta_j(\theta_i)$, their sum will attain larger values when the resulting values of $T_j(x_i) = \eta_j(\theta_i)$ are similar; hence, we aim to construct mappings $\widehat{T}$ and $\widehat{\eta}$ which will have this property. However, because simply maximizing the likelihood over a training set may lead to overfitting, we choose to leave the extent of pulling together $\widehat{T}$ and $\widehat{\eta}$ as a parameter to be tuned.

Pursuing mappings $\widehat{T}$ and $\widehat{\eta}$ with this property, we appeal to diffusion mapping (Coifman & Lafon, 2006), a dimension-reduction method which finds lower-dimensional structure based on information about the similarity between observations. Formally, diffusion mapping utilizes a kernel similarity matrix $\mathbf{K}$ where $\mathbf{K}_{uv} = k(u, v)$ for some kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ that quantifies some notion of similarity between two observations u and v. These similarities suggest a symmetric graph, where the nodes are observations and the transition probabilities are the (suitably normalized) similarities between observations.

The central insight of diffusion mapping is that the eigenvectors of the resulting transition matrix are meaningful dimension reduction coordinates. Diffusion mapping is particularly successful at identifying local low-dimensional structure which may be nonlinear (as opposed to, e.g., principle components analysis); additionally, among nonlinear dimension reduction techniques, it is robust to the presence of noise added to the lower-dimensional structure.

To make use of diffusion mapping, we conceive of the parameters and data existing in the same space and quantify the similarity between pairs of parameters, pairs of data, and corresponding parameter-data pairs. To this end, we construct kernel similarity submatrices $\mathbf{K}_\theta$ and $\mathbf{K}_x$, which capture the similarities within, respectively, the training parameter set and the training data set. For $\mathbf{K}_\theta$ and $\mathbf{K}_x$, we choose k to be the exponential kernel, $k_h(u,v) = \exp(-\frac{\|u-v\|}{h})$, where we use the Euclidean norm $\|\cdot\|$. We remark that it may well be appropriate to use different bandwidths for the $\mathbf{K}_\theta$ and $\mathbf{K}_x$ submatrices; our default is to use the 0.05 quantile of the set of Euclidean distances between parameters and separately the set of Euclidean distances between data. We note further that the bandwidth h could optionally be a varied as another tuning parameter, and that different forms of the kernel (e.g. Gaussian, Epanechnikov) may be acceptable.

All of this similarity information is encoded in a $2N_T \times 2N_T$ block kernel similarity matrix $\mathbf{K}$,

$$\mathbf{K} = \begin{bmatrix} \mathbf{K}_\theta & \lambda \mathbf{I}_{N_T} \\ \lambda \mathbf{I}_{N_T} & \mathbf{K}_x \end{bmatrix}$$

Crucially, λ is the tuning parameter that represents, heuristically, the affinity between a parameter element and its corresponding data element. We remark that the parameter kernel matrix $\mathbf{K}_\theta$ can and should be constructed using a different bandwidth h from that used to construct the $\mathbf{K}_x$, and that these bandwidths are additional tuning parameters which can provide some more flexibility in learning a mapping.

Following the diffusion map approach, a spectral decomposition of this $\mathbf{K}$ matrix yields eigenvectors $\widehat{\psi}_j$ for $j = 1, \ldots, N_T$. Our approach is to take $[\widehat{\eta}_j \; \widehat{T}_j]^T = \widehat{\psi}_j$ for $j = 1, \ldots, J$, retaining the first J eigenvectors. We can thus obtain a value of $\widehat{T}_j(x_i)$ for each x_i in the training set. However, the value of using the common-space mappings as ABC summary statistics hinges upon the ability to project out-of-sample points (i.e., both observed and simulated data) into the common space. Fortunately, for a new data point $\tilde{x}$, the value of the mapping $\widehat{T}(\tilde{x})$ can be constructed via a variant[1] of the Nyström extension:

$$\widehat{T}_j(\tilde{x}) = \frac{\sum_{i=1}^{N} k(\tilde{x}, x_i)\widehat{T}_j(x_i)}{\lambda_j \sum_{i=1}^{N} k(\tilde{x}, x_i)}.$$

The resulting $\widehat{T}_j(\tilde{x})$ is thus a weighted average of the $\widehat{T}_j(x_i)$ for the x_i in the training set — with weights given by the new point's similarity to each point in the training set — and rescaling by the inverse of the eigenvalues λ_j of the kernel similarity matrix $\mathbf{K}$, as in Coifman et al. (2008).

$\widehat{T}$ will be sensitive to the affinity tuning parameter λ; using a large value for λ will prioritize the relationship between parameter values and the data that generated them, relative to the relationship between points close together in the parameter space or between points close together in the data space. Figure 3.1 depicts, for illustrative purposes, the results of the common-space mappings for different λ values in a simple model setting.

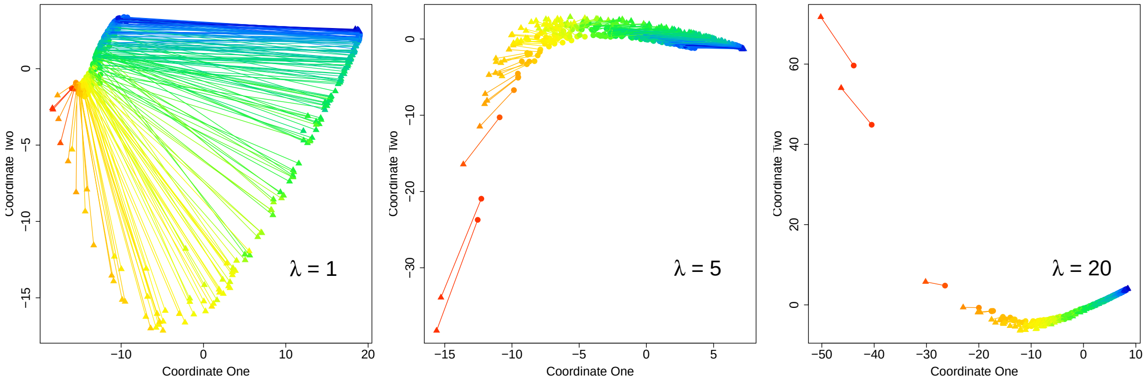

Figure 3.1: For a simple modeling example, the first two common-space mapping coordinates, visualized for three values of λ, with corresponding mapped pairs of $\widehat{T}$ (bullets) and $\widehat{\eta}$ (triangles) joined.

[1] The standard formulation of the Nyström extension would use the entire $2N$-length eigenvector $\widehat{\psi}_j$, but doing so here would require weights $k(\tilde{\theta}, \theta_i)$ for each θ_i in the training set. Obtaining these would require knowledge of $\tilde{\theta}$, the parameter value corresponding to the new data point $\tilde{x}_i$, which is the unknown parameter of interest.

We notice that as λ increases, the corresponding bullets (representing $\widehat{T}(x_i)$ values) and triangles (representing $\widehat{\eta}(\theta_i)$ values) are pulled closer together relative to their separation from other shapes of the same kind. Hence, the affinity λ governs the priority which the CSMM will afford to the correspondence between a parameter value θ_i and the data point x_i generated under it. In Section 3.2, we discuss an approach for choosing the tuning parameter λ to yield approximately sufficient ABC summary statistics.

3.1.3 Connection to exponential families

We mention briefly that the distribution $f(x|\theta)$ is said to belong to an exponential family if

$$f(x|\theta) = h(x)\exp\left(\eta(\theta)\cdot T(x) - A(\theta)\right) \tag{3.4}$$

for some functions $A(\theta)$ and $h(x)$ and possibly vector-valued functions $\eta(\theta)$ and $T(x)$. If this assumption, which is stronger than the one made above in (3.1), holds, then $T(x)$ is a sufficient statistic for θ.

In this setting, if θ_i is drawn from a prior $\pi(\theta)$, then the joint log-likelihood for a parameter-data pair (θ_i, x_i) would be $\log f(x_i, \theta_i) = \log\left(\pi(\theta_i) f(x_i|\theta_i)\right)$, which assuming (3.4) can be written as

$$\log f(x_i, \theta_i) = \log \pi(\theta_i) + \log h(x_i) + \eta(\theta_i)\cdot T(x_i) - A(\theta_i) \tag{3.5}$$

and the joint log-likelihood of a training data set, of the sort described in Section 3.1.2, would be given by

$$\sum_{i=1}^{N_T} \left[\log \pi(\theta_i) + \log h(x_i) + \eta(\theta_i)\cdot T(x_i) - A(\theta_i)\right]. \tag{3.6}$$

With respect to choices of $\widehat{\eta}$ and $\widehat{T}$, an estimator of the above will attain large values when the dot product $\widehat{\eta}(\theta_i)\cdot\widehat{T}(x_i)$ is made larger, which will happen when, for corresponding θ_i and x_i, $\widehat{\eta}_j(\theta_i)$ and $\widehat{T}_j(x_i)$ have the same sign.

Thus, when $f(x|\theta)$ belongs to an exponential family, there is an additional justification for seeking mappings $\widehat{\eta}$ and $\widehat{T}$ which map corresponding data-parameter pairs close together in the common space; in fact, a $\widehat{T}$ mapping constructed thus should be an estimate of the exponential family sufficient statistic $T(\cdot)$. However, it is worth noting that performing an ABC analysis may well be computational overkill in the setting of exponential family models; in those settings, more conventional likelihood-based approaches should be adequate.

3.2 Choice of tuning parameter

The common space mapping $\widehat{T}$ and the usefulness of any ABC summary statistic derived from that mapping will be sensitive to the choice of the affinity tuning parameter λ. In this suggestion, we motivate and present criteria to guide the choice of λ. Our approach will be to minimize, over candidate values of λ, an estimate of the *sufficiency gap*, an information theoretic quantity that quantifies the extent to which a statistic is not sufficient. Precedent for considering approximate sufficiency exists in the literature in form of the *deficiency* of Le Cam (1964), which is explored in an information-theoretic context by Raginsky (2011).

3.2.1 Mutual information and sufficiency

First, we recall (see, e.g., Cover & Thomas, 2012) that for continuous random variables X and Y with marginal densities $f(x)$ and $f(y)$ and joint density $f(x, y)$, $I(X, Y)$ is defined as

$$I(X,Y) = \int f(x,y) \log \frac{f(x,y)}{f(x)f(y)} \, dx \, dy$$

It follows immediately that if X and Y are independent, $I(X, Y) = 0$, since $\log \frac{f(x,y)}{f(x)f(y)} = 0$ throughout; the converse is also true. Hence, $I(X, Y)$ quantifies, in some sense, the strength of the (possibly nonlinear) relationship between X and Y. We recall further that $I(X, Y)$ is nonnegative.

In the setting of Bayesian inference, both parameters θ and data X are considered to be random. A basic result in information theory, the *data processing inequality*, states that for any function $S(\cdot)$ of the data,

$$I(\theta; S(X)) \leq I(\theta; X),$$

with equality if and only if $S(\cdot)$ is sufficient for θ.

Simply rearranging the above gives

$$I(\theta; X) - I(\theta; S(X)) \geq 0, \tag{3.7}$$

where, again, equality will hold if and only if $S(\cdot)$ is sufficient for θ. Thus, in some sense, the term on the left side of (3.7), which we will call the *sufficiency gap* in what follows, quantifies departure from sufficiency, and summary statistics $S(\cdot)$ that make that term small will be approximately sufficient. $I(\theta; X)$ cannot change with the choice of $S(\cdot)$, so if the goal is to reduce the sufficiency gap, $S(\cdot)$ should be chosen to make $I(\theta; S(X))$ as large as possible.

This idea guides the choice of the affinity parameter λ for the CSMM-derived $\widehat{T}$ summary statistic. In the case where θ and $\widehat{T}$ are of very low dimension, an acceptable approach is to simply estimate $I(\theta; \widehat{T}(X))$ for the $\widehat{T}(X)$ mappings resulting from various values of λ and choose the value of λ that maximizes $\widehat{I}(\theta; \widehat{T}(X))$. In order to estimate mutual information, we rely on the k-nearest neighbor approach of Kraskov et al. (2004) as implemented in the NPEET Python toolbox (Ver Steeg, 2014). $I(\theta; S(X))$ can be estimated using the values of θ in the simulated training set used to construct the common space mapping and the corresponding values of $\widehat{T}$ that result from that mapping. In the simulation examples of sections 4.1, 4.2, and 5.2, λ is chosen via this approach of empirically maximizing $\widehat{I}(\theta; \widehat{T}(X))$ over a set of candidate λ values.

We claim that choosing the tuning parameter λ thus will lead, in expectation over random data, to a more accurate estimate of the posterior distribution, as measured by Kullback-Leibler loss.

Proposition 3.2.1. *The sufficiency gap introduced in* (3.7) *is equivalent, in expectation* *over random X, to the Kullback-Leibler divergence between the true posterior distribution $\pi(\theta|X)$ and $\pi(\theta|S(X))$, the posterior distribution conditional on the possibly non-sufficient summary statistic $S(\cdot)$. Formally,*

$$I(\theta; X) - I(\theta; S(X)) = \mathbb{E}_X \left[KL \left(\pi(\theta|X) \| \pi(\theta|S(X)) \right) \right]. \tag{3.8}$$

Proof. We assume that $\theta \in \mathbb{R}^p$ and $X \in \mathbb{R}^d$ are continuous random variables with densities $f(\theta)$ and $f(x)$, respectively. We will begin by considering the sufficiency gap $I(\theta; X) - I(\theta; S(X))$. We recall that mutual information can be rewritten in terms of entropy and conditional entropy; namely, that for any random variables X and Y, $I(X; Y) = H(X) - H(X|Y)$. Hence,

$$\begin{aligned} I(\theta; X) - I(\theta; S(X)) &= H(\theta) - H(\theta|X) - [H(\theta) - H(\theta|S(X))] \\ &= -H(\theta|X) + H(\theta|S(X)) \\ &= \int f(\theta, x) \log \pi(\theta|x) d\theta dx - \int f(\theta, s) \log \pi(\theta|s) d\theta ds. \end{aligned}$$

Because X partitions Ω more finely than $S(X)$, we can rewrite the second integral (which is over $\theta \times S$) as an integral over $\theta \times X$:

$$\begin{aligned} &= \int f(\theta, x) \log \pi(\theta|x) d\theta dx - \int f(\theta, x) \log \pi(\theta|s(x)) d\theta dx \\ &= \int f(\theta, x) \log \frac{\pi(\theta|x)}{\pi(\theta|s(x))} d\theta dx \\ &= \int f(x) \pi(\theta|x) \log \frac{\pi(\theta|x)}{\pi(\theta|s(x))} d\theta dx. \\ &= \int f(x) \left[KL \left(\pi(\theta|X) \| \pi(\theta|S(X)) \right) \right] dx, \end{aligned}$$

which is exactly the right-hand side of (3.8). □

We remark that both sides of (3.8) are equal to zero when $S(\cdot)$ is sufficient.

3.2.2 Pairwise information criterion

While the preceding approach is acceptable in settings where θ and $\widehat{T}$ are of very low dimension, the nearest-neighbor estimator of mutual information will suffer from the curse of dimensionality if θ or $\widehat{T}(X)$ are of even moderately high dimension. Hence, in settings where it may not be appropriate to estimate $I(\theta; \widehat{T}(X))$ directly, we present an alternative, related criterion based on notions of pairwise separation.

We consider two observations, X_i and X_j, generated from the simulation process parameterized by θ_i and θ_j, respectively. We claim that a desirable attribute of a summary statistic $S(\cdot)$ is that it should respect separation between θ_i and θ_j; that is, if $\rho_\theta(\theta_i, \theta_j)$ is large for some distance ρ_θ, then $\rho_S\left(S(X_i), S(X_j)\right)$ will also be large.

We remark that in a loose conceptual sense, this intuition is related to the set of conditions which Frazier et al. (2015) states are necessary for ABC to achieve Bayesian consistency. That paper defines $\mathbf{b}(\theta)$ as the limiting value of the summary statistic based on data simulated under θ. Theorem 1 requires, among other conditions, that the map $\theta \to \mathbf{b}(\theta)$ be deterministic, continuous, and one-to-one in θ. Explicitly, the one-to-one condition means that if $\theta_i \neq \theta_j$, then $\lim_{d\to\infty} S(X_i^d) \neq \lim_{d\to\infty} S(X_j^d)$, where d represents here the dimension of the data. Put another way, if there exists some $\epsilon_\theta > 0$ such that $\rho_\theta(\theta_i, \theta_j) > \epsilon_\theta$, then there also exists some $\epsilon_S > 0$ and no D such that $\rho_S\left(S(X_i^d), S(X_j^d)\right) > \epsilon_S$ for all $d > D$.

One way to quantify the extent to which a statistic $S(\cdot)$ respects this separation is to enumerate pairs (θ_i, θ_j) and measure the strength of the relationship between $\rho_\theta(\theta_i, \theta_j)$ and $\rho_S\left(S(X_i), S(X_j)\right)$. The relationship need not be linear, so measures that quantify potentially non-linear correspondence may be appropriate. In particular, there is precedent in the literature (see, e.g. Reshef et al., 2011) for considering mutual information as a notion of measuring nonlinear dependence.

Choosing the last of these, we define the *mutual information of pairwise separation*

(MIPS) as

$$I\left(\rho_\theta(\theta_i, \theta_j), \rho_S\left(S(X_i), S(X_j)\right)\right),$$

which we can estimate via the same estimators as we use to estimate $I(\widehat{T}_\lambda, \theta)$. There are two advantages of the pairwise criterion: first, $\rho_\theta(\theta_i, \theta_j)$ and $\rho_S\left(S(X_i), S(X_j)\right)$ are one-dimensional, so the estimator of mutual information will not encounter the curse of dimensionality, and additionally, the pairwise criterion allows the user to specify ρ_θ in the event that a particularly natural notion of distance exists in the parameter space.

For the estimated pairwise information criterion to be truly useful as a guide to choosing λ, it would ideally attain a maximum at a similar region of λ to where the direct information criterion is maximized. In simulation studies that follow, where the dimension of θ and $\widehat{T}$ are low enough to allow for estimation of both criteria, we find (see, e.g., Figure 4.5) that the criteria seem to be related and are largest in some intermediate range of λ values.

Chapter 4

Applications: toy problems

In each of the following two toy simulation examples, we use ABC to obtain an approximate posterior distribution for one of the parameters of the normal distribution (taking the other parameter as fixed). In each case, we draw from a prior distribution conjugate to the normal likelihood so that, conditional on the value of the observed data, the true posterior distribution $\pi(\theta|x_{obs})$ is known analytically.

4.0 Quantifying posterior loss

In the simulation examples that follow, the product of each ABC analysis will be a sample of retained parameter values $\tau_1, \ldots, \tau_T$ and corresponding weights $w_1, \ldots, w_T$. We will assess the performance of an ABC summary statistic $S(\cdot)$ according to two notions of discrepancy between a posterior estimate (conditioned on $S(\cdot)$ and the true distribution.

4.0.1 CDF L^2 loss

As the parameter in these examples is one-dimensional, it is straightforward to evaluate the empirical weighted CDF for any value of θ, which is given by

$$\widehat{F}_T(\theta) = \sum_{i=1}^{T} w_i \, \mathbb{I}(\tau_i \leq \theta).$$

This empirical weighted CDF can be compared to the known true posterior CDF $F(\theta)$ by taking the mean of the squared differences between those the two CDFs evaluated on a lattice of parameter values:

$$\frac{1}{T}\sum_{j=1}^{T}\left(\widehat{F}_T(\theta_j) - F(\theta_j)\right)^2, \tag{4.1}$$

where the θ_j values are spaced uniformly in the central interval containing, e.g., 99.9% of the prior density. This quantity will serve as a numerical approximation to the L^2 distance

$$\int |\widehat{F}_T(\theta) - F(\theta)|^2 d\theta$$

between the true posterior CDF and the empirical weighted CDF obtained from the ABC sample.

4.0.2 Kullback-Leibler loss

The second notion of discrepancy will be (a sample estimate of) the Kullback-Leibler loss function

$$KL(\theta|x_{obs}\|\theta_{ABC}|s(x_{obs})) = \int \pi(\theta|x_{obs}) \log \frac{\pi(\theta|x_{obs})}{\pi_{ABC}(\theta|s(x_{obs}))} d\theta \tag{4.2}$$

with respect to the true known posterior distribution. We remark that, in expectation over the randomness in x_{obs}, this quantity should be minimized by choosing the tuning parameter according to the method described in Section 3.2.

Quantifying the Kullback-Leibler loss between a known distribution and a weighted sample is not trivially straightforward. Our approach is to use the nearest-neighbor based estimator of Wang et al. (2009) as implemented in the NPEET Python toolbox (Ver Steeg, 2014), which computes an estimate of $KL(P\|Q)$ given samples from each of the distributions P and Q. In order to estimate (4.2) via this estimator, we draw 1000 samples from the known distribution $\pi(\theta|x_{obs})$ and 1000 samples (with replacement) from the weighted ABC particle sample. In practice, we find that the resulting estimates of (4.2) are relatively stable under repeated resampling.

4.1 Unknown mean, known variance

We return to the toy example first outlined in Section 2.2, using ABC to obtain the posterior distribution of μ given n independent observations from a $\mathcal{N}(\mu, 1)$ random variable. Formally, in the setting where

$$\mu \sim \mathcal{N}(\mu_0, 1), \text{ and}$$
$$X_1, \ldots X_n \overset{IID}{\sim} \mathcal{N}(\mu, 1),$$

then the posterior distribution for the parameter μ conditional on the data is given by

$$\mu | X_1, \ldots X_n \sim \mathcal{N}(\frac{1}{n+1}\mu_0 + \left(\frac{n}{n+1}\right)\overline{X}, \frac{1}{n+1}).$$

For our simulations, we set $\mu_0 = 2$ and $n = 32$. In each of our 100 simulations, we use $\mu = 0$ as the true mean to generate an observed data vector; obtain an ABC SMC posterior sample using each of a set of summary statistics in turn; and calculate the posterior CDF L^2 loss and KL loss, according to the approach in Section 4.0.

We obtain ABC SMC posterior samples for μ using six different summary statistics. Four are described in section 2.2: the overall mean $\overline{X}$; the means-within-halves $(\overline{X}_1^1, \overline{X}_2^1)$; the means-within-quarters $(\overline{X}_1^2, \overline{X}_2^2, \overline{X}_3^2, \overline{X}_4^2)$; and the mean $\overline{X}_{16}$ of only the first half of the data. To these we add the median and the CSMM-derived $\widehat{T}$. In each simulation and for each summary statistic, we use 100 particles and iterate for 8 ABC SMC steps.

To train the CSMM, we build a training set of 200 (θ_i, x_i) pairs (where, in this example, the parameter θ is the mean μ), by drawing 200 parameter values from the prior $\pi(\theta)$ and, for each one, generating data from the forward model. We then construct the similarity matrices for parameters and data, quantifying distance for parameters via $d(\theta_u, \theta_v) = |\theta_u - \theta_v|$ and for data via $d(x, y) = \sqrt{\sum_{i=1}^n (x_{(i)} - y_{(i)})^2}$, the Euclidean distance between the order statistics of the data. (We compare order statistics because the independent and identical distribution of the $X_1, \ldots, X_n$ means their individual indices have no intrinsic meaning.) We elect to retain only the first dimension of the $\widehat{T}$ mapping, according to the intuition, suggested in Fearnhead & Prangle (2012), that the approximately sufficient statistic should be have dimension no larger than that of the parameter.

The affinity λ is chosen to maximize the estimated information criterion $\widehat{I}(\theta; \widehat{T}(X))$ within a set of candidate values of λ. Accordingly (see the left panel of Figure 4.5), we choose an affinity value of $\lambda = 8$ when comparing the CSMM-derived $\widehat{T}$ against other summary statistics in this example. The ABC posterior estimates for one example set of simulated data are shown in Figure 4.1, with (left) the empirical weighted CDFs and (right) kernel estimates of the PDFs. For practical reasons, we do not show all of the summary statistics considered. The results of this particular simulation run are fairly typical in the sense that the aggregated KL losses due to each summary statistic are similar to that statistic's average KL loss across 100 simulations. We observe that, for this one data set, the non-sufficient median is substantially inferior to the minimally sufficient $\overline{X}$ in terms of recovering the true posterior; the CSMM-derived $\widehat{T}$ performs comparably to the non-minimal but sufficient means-within-quarters.

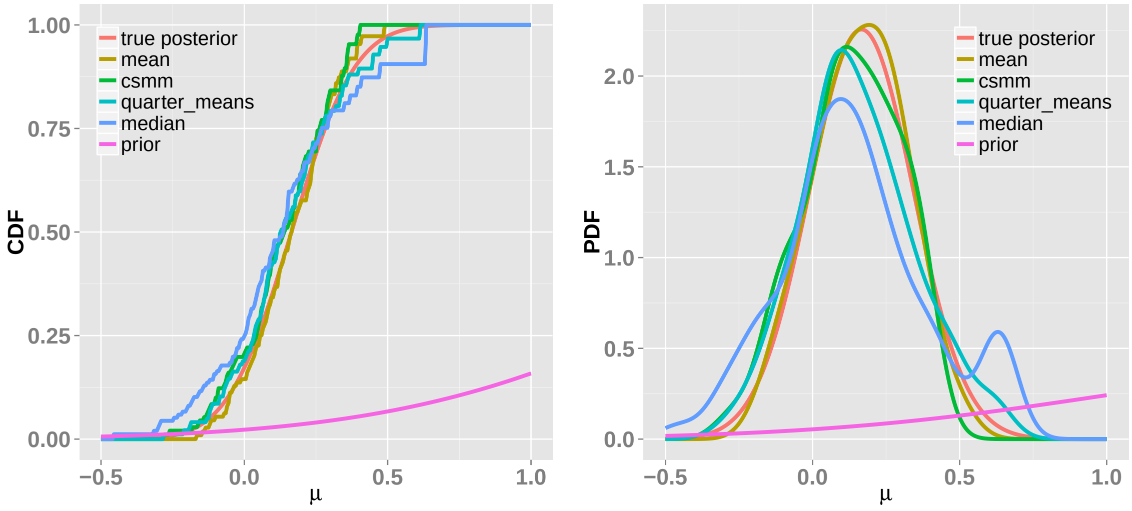

Figure 4.1: Example ABC posterior estimates of μ from one particular set of $X_1, \ldots, X_{32}$ data.

The violin plots in Figures 4.2 and 4.3 display the distributions of posterior CDF L^2 loss and KL loss across the 100 simulations for ABC posteriors using each summary statistic. Looking at the violin plots of L^2 loss, we remark that the CSMM-derived $\widehat{T}$, the minimal

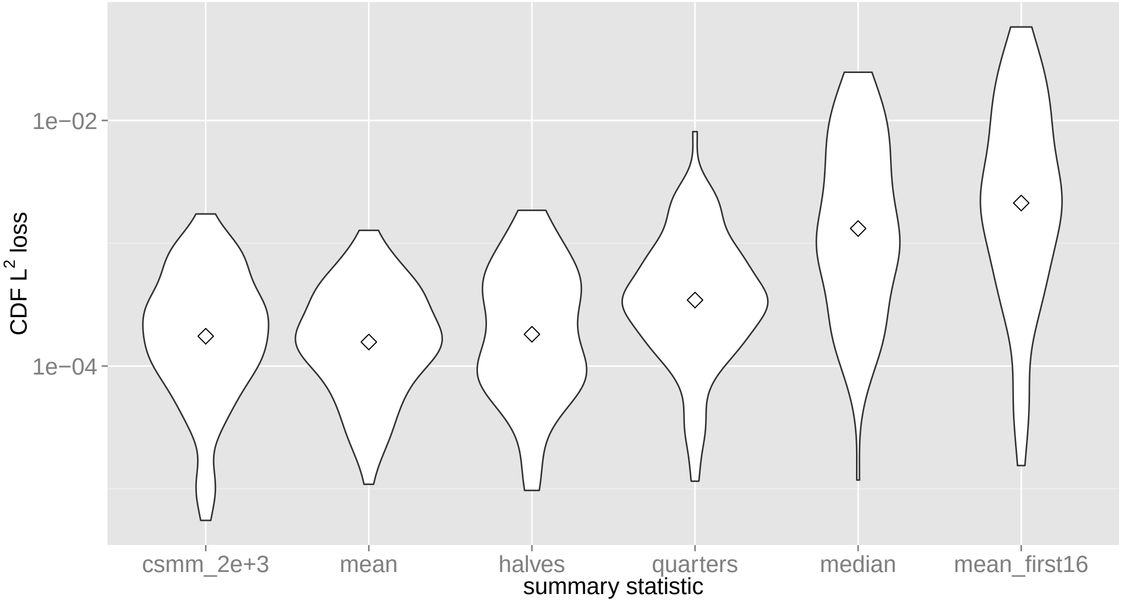

Figure 4.2: Distributions of CDF L^2 loss, across 100 simulations, when estimating the normal-normal posterior for μ via ABC using different summary statistics. Note that the y-axis is on a logarithmic scale.

sufficient statistic $\overline{X}$, and the sufficient-but-not-minimal statistics all perform comparably, with the means-within-quarters statistic exhibiting slightly inferior behavior to the others. The performance of the two non-sufficient statistics (at far right of each figures) is markedly inferior to that of the sufficient statistics. For $\overline{X}_{16}$, the mean of only the first half of the data, this is not terribly surprising; it is slightly more surprising for the median, which can be shown to be a consistent estimator of the mean for normally distributed data.

Measuring discrepancy according to posterior KL loss (Figure 4.3), we see similar broad trends, in the sense that the CSMM-derived $\widehat{T}$ performs comparably to the sufficient statistics, and the non-sufficient statistics perform considerably worse. The main difference is that the sufficient but not minimal means-within-halves and means-within-quarters statistics do not perform worse with respect to KL loss; we remark that this phenomenon may be, to some extent, an artifact of the noise in our nearest-neighbor estimator of KL divergence. We comment further that the relatively high value of KL loss obtained using even the minimal sufficient statistic $\overline{X}$ is likely due to the Monte Carlo error associated with exploring a continuous posterior based on a finite sample — in this case, 100 particles.

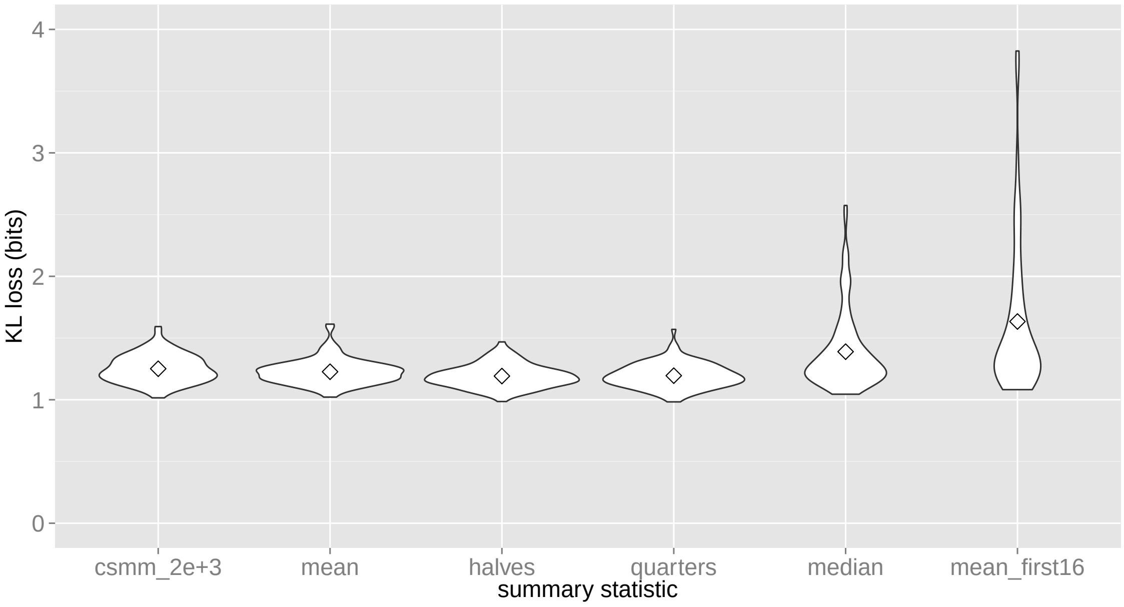

Figure 4.3: Distributions of posterior KL loss, across 100 simulations, when estimating the normal-normal posterior for μ via ABC using different summary statistics.

4.1.1 Sensitivity of posterior KL loss to the tuning parameter λ

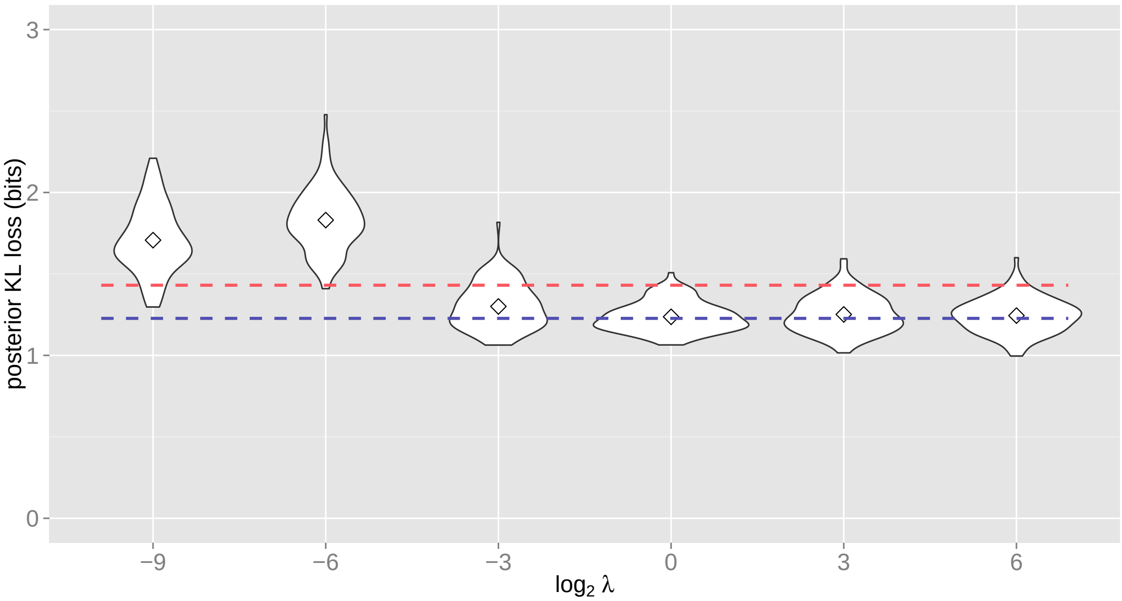

Figure 4.4: Distributions of posterior KL loss, across 100 simulations, when estimating the normal-normal posterior for μ via ABC using the CSMM-derived $\widehat{T}$ for different values of λ. The blue dotted line represents the average error using the mean as a summary statistic, while the red dotted line represents the same using the median.

Given the connection between $I(\theta; S(X))$ and $KL\left(\pi(\theta|X)\|\pi(\theta|S(X))\right)$ developed in Section 3.2, we would expect that for values of λ where an empirical estimate of $I(\theta; \widehat{T}_\lambda)$ is larger, the posterior KL loss from conditioning on $\widehat{T}_\lambda$ should be smaller when averaged across many simulations.

Figure 4.4 displays the distribution of KL losses across 100 simulations for six different values of λ, logarithmically spaced between 2^{-9} and 2^{6}. For values of λ below 2^{-3}, the CSMM-derived summary statistics perform poorly — comparably to the non-sufficient median — while for larger values of λ, the performance is comparable to that of the minimally sufficient mean. This behavior is consistent with the trends we see in $\widehat{I}(\theta; \widehat{T}_\lambda)$ in Figure 4.5, in the sense that the estimated information criteria are also lower for smaller values of λ.

We do not, however, see the expected correspondence on the higher end of the λ range, where performance appears to remain adequate despite a decaying information criterion. It is worth mentioning that for especially large values of λ, the ABC algorithm runs more slowly, requiring (in some cases, prohibitively) many more candidate data sets generated. This run time phenomenon, combined with the larger estimated information criteria, suggests that an intermediate value of λ is still desirable.

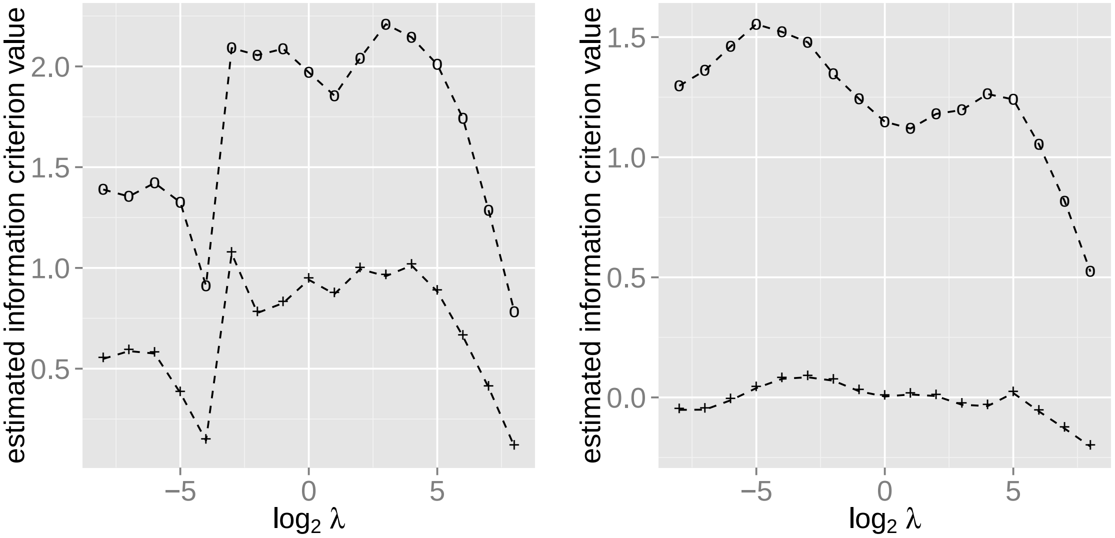

Figure 4.5: For a set of candidate values of the CSMM affinity λ (displayed on a logarithmic scale), the estimated information criteria in the unknown mean simulations of Section 4.1 (left) and the unknown precision simulations of Section 4.2 (right). Open circles represent the direct criterion $\widehat{I}(\theta; \widehat{T}(X))$, while crosses represent the pairwise criterion $\widehat{I}\left(\rho_\theta(\theta_i, \theta_j), \rho_S\left(T(X_i), T(X_j)\right)\right)$.

4.2 Known mean, unknown precision

The preceding example demonstrates that the CSMM-derived $\widehat{T}$, when used as a summary statistic in an implementation of ABC, exhibits performance comparable to that of the minimal sufficient statistic in terms of estimating the posterior distribution of a location parameter μ. The simulation results of this section suggest that the same is true when estimating the posterior distribution of a scale parameter.

To that end, we consider the conjugate Gamma-normal model for the precision σ^{-2}.

Formally, in the setting where

$$\sigma^{-2} \sim Gamma(\alpha_0, \beta_0), \text{ and}$$
$$X_1, \dots X_n \overset{IID}{\sim} \mathcal{N}(\mu, \sigma^2), \text{ for some known } \mu,$$

then the posterior distribution for the parameter σ^{-2} conditional on the data is given by

$$\sigma^{-2}|X_1, \dots X_n \sim Gamma\left(\alpha_0 + \frac{n}{2}, \beta_0 + \frac{1}{2}\sum_{i=1}^{n}(X_i - \mu)^2\right).$$

For our simulations, we set $\alpha_0 = 1$, $\beta_0 = 1$, $\mu = 0$, and $n = 32$. In each of our 100 simulations, we use $\sigma^{-2} = 0.25$ as the true precision to generate an observed data vector; obtain an ABC SMC posterior sample using each of a set of summary statistics in turn; and calculate the posterior CDF L^2 loss and KL loss, according to the approach in Section 4.0.

We implement our ABC analysis using four summary statistics. The first of these, $\sum_{i=1}^{n} X_i^2$, can easily be shown to be the minimal sufficient statistic for σ^{-2}. For the sake of comparison, we also consider the interquartile range, a measure of spread that is not sufficient for σ^{-2} but nonetheless is (up to scaling) a consistent estimator of the standard deviation σ in the normal model. Further, we consider the mean $\overline{X}$, which should not contain information relevant for inference on σ^{-2}. Finally, we consider the CSMM-derived mapping $\widehat{T}$.

We note that the mechanics of the procedure for learning this mapping are very much the same as in the preceding example involving μ. The affinity is, as before, chosen so as to maximize $\widehat{I}(\widehat{T}_\lambda, \theta)$ from among a set of candidate λ values, with the $\widehat{I}(\widehat{T}_\lambda, \theta)$ criterion for various values of λ shown in the right panel of Figure 4.5. The set of ABC posteriors resulting from the various summary statistics for one typical set of simulated data are shown in Figure 4.6. We note that, as before, the minimal sufficient statistic is very successful at recovering the true posterior, while non-sufficient statistics perform considerably worse.

We inspect the distributions of posterior estimation error, measured in terms of L^2 loss

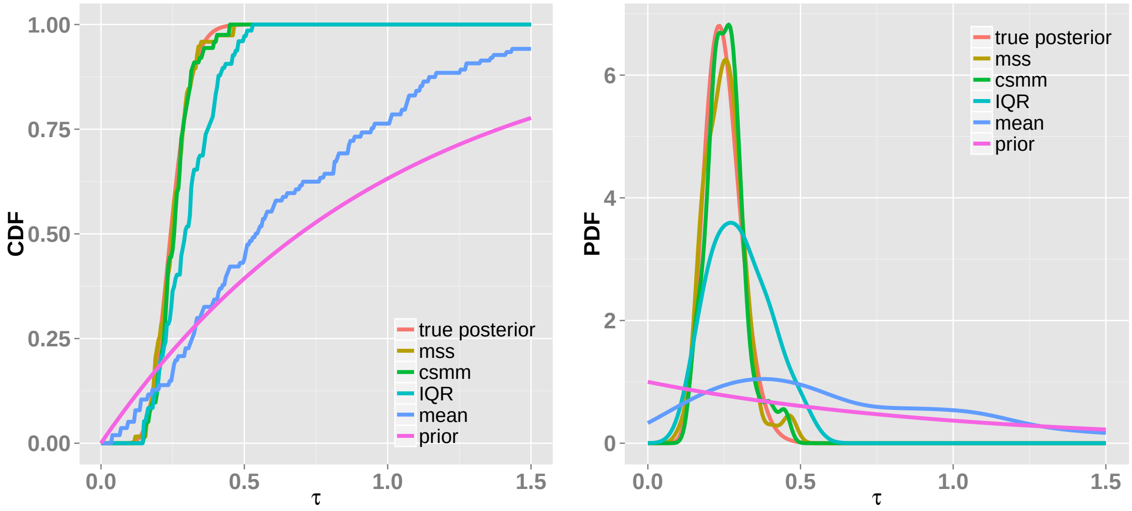

Figure 4.6: Example ABC posterior estimates of σ^{-2} from one particular set of $X_1, \ldots, X_{32}$ data.

(Figure 4.7) and KL loss (Figure 4.8). Directionally at least, the results are consistent across the two notions of discrepancy and also with our intuitions. Using the minimal sufficient statistic tends to yield a small KL loss, using the intuitively irrelevant $\overline{X}$ tends to result in a large error, and using the non-sufficient IQR tends to result in an intermediate error value. We remark that in these simulations, the CSMM-derived $\widehat{T}$ results in a posterior error comparable to the error using the minimal sufficient statistic and thus vastly outperforms the non-sufficient statistics.

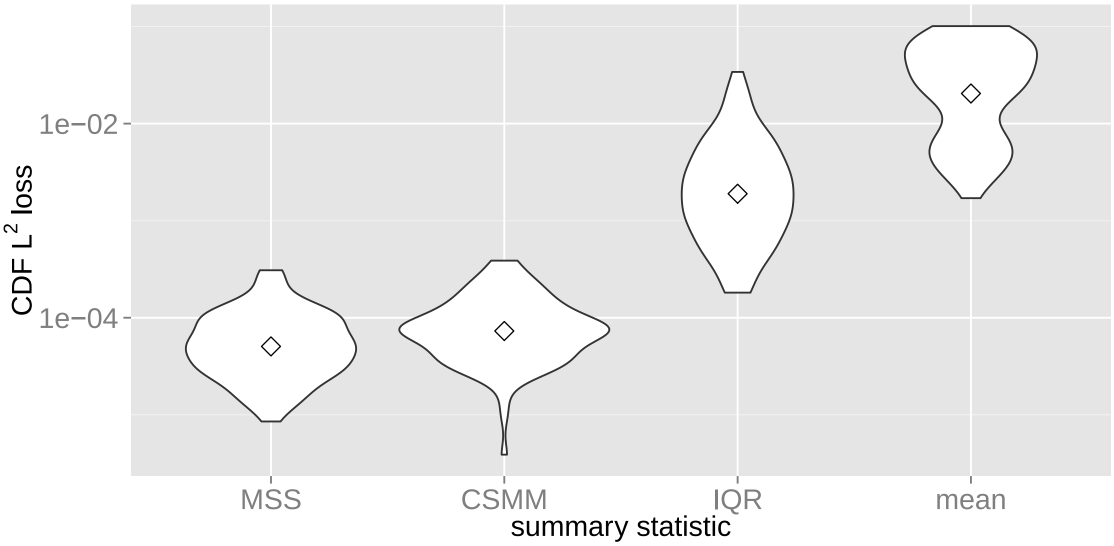

Figure 4.7: Distributions of CDF L^2 loss when estimating the gamma-normal posterior for σ^{-2} via ABC using different summary statistics. Note that the y-axis is on a logarithmic scale.

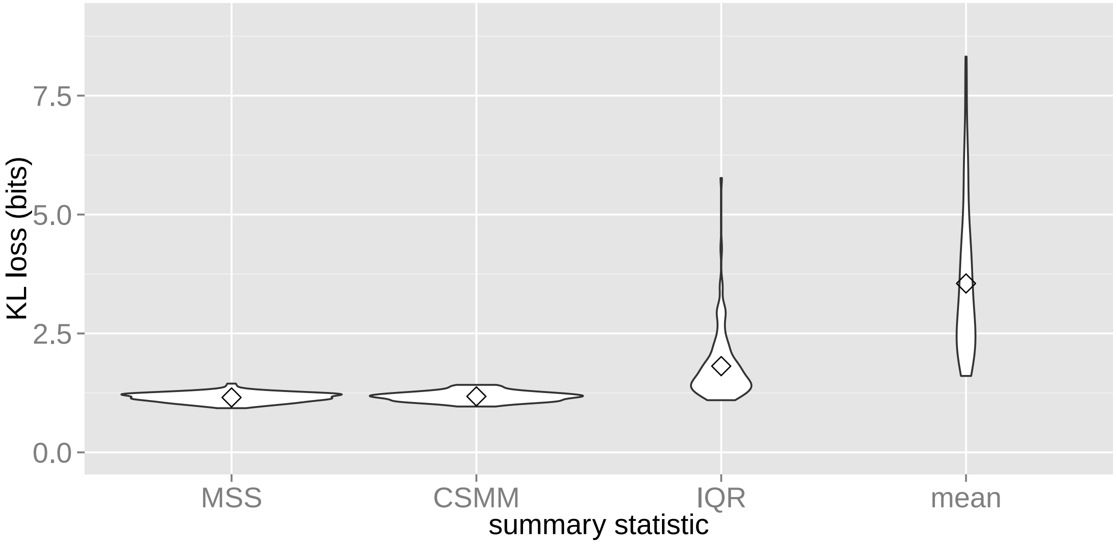

Figure 4.8: Distributions of posterior KL loss when estimating the gamma-normal posterior for σ^{-2} via ABC using different summary statistics.

Chapter 5

Applications: weak gravitational lensing

In this section, we consider the problem of cosmological parameter inference via weak gravitational lensing, which provided us with the original motivation for developing approximately sufficient low-dimensional statistics for use in likelihood-free inference. We give here a brief overview of weak lensing, deferring to reviews (e.g., Hoekstra & Jain, 2008; Munshi et al., 2008) for a more complete exposition of the phenomenon and associated inference techniques.

5.1 Overview of weak gravitational lensing

Prevailing cosmological models posit the existence of dark matter, i.e., matter which neither emits nor absorbs light, because the behavior of galaxies does not accord with predictions based only on the amount of luminous matter. Weak gravitational lensing is a phenomenon which permits inference on parameters in cosmological models (e.g., Ω_M, the dark matter density, and σ_8, the matter power spectrum normalization, in the ΛCDM model). General relativity predicts that the path of light traveling from distant galaxies to an observer should be bent by intervening matter, resulting in a circular object being observed as an ellipse (Dodelson, 2003); the spatial autocorrelation in the amount of shear of galaxies is related to the parameters mentioned above.

Distant galaxies already have some intrinsic ellipticity, so that the result of weak lensing is a slight modification of their ellipticity, roughly on the order of one percent of intrinsic ellipticity (Mandelbaum et al., 2014). Although this shear signal is incredibly faint for each individual galaxy, the shearing of individual galaxies is less relevant for cosmological parameter constraint than the ensemble shear behavior. Underlying smoothness in the dark matter structure dictates that nearby galaxies should exhibit similar shear effects. Hence, any parameter inference from weak lensing must incorporate information about the shapes of large numbers of galaxies.

This requires an accurate method for measuring galaxy shapes, a task made more difficult by contamination due to atmospheric and detector effects, commonly referred to as the point spread function (PSF). The recent series of GREAT (Gravitational lEnsing Accuracy Testing) challenges (see Mandelbaum et al., 2014, and references therein) has attempted to attract computational researchers across various disciplines to the development of shape-measurement methodology precise enough to be applicable to upcoming surveys. For the purposes of this work, we will assume that a method exists to measure galaxy shear to a precision which can itself be quantified.

Given a galaxy catalogue with associated positions, shape measurements, and estimates of shape measurement error, a standard approach to parameter constraint is to first compute a sample estimate of the two-point shear correlation function $\xi_\pm$ or its Fourier transform, the shear power spectrum C_ℓ. Unbiased estimates of the $\xi_\pm$ are given by

$$\widehat{\xi}_\pm(\theta) = \frac{\sum\limits_{(i,j)} w_i w_j \left(\varepsilon_i^{++}\varepsilon_j^{++} \pm \varepsilon_i^{\times\times}\varepsilon_j^{\times\times}\right)}{\sum\limits_{(i,j)} w_i w_j} \tag{5.1}$$

where ε_i^{++} and $\varepsilon_i^{\times\times}$ are estimates of the tangential and cross components of galaxy ellipticity for galaxy i; w_i is a weight associated with the precision of the ε_i measurement; and the sums $\sum_{(i,j)}$ are over pairs of galaxies (i, j) separated by angular distance θ (up to some binning in θ). Some analyses consider other statistics, e.g., aperture mass dispersion (Schneider et al., 2002), ring statistics (Schneider & Kilbinger, 2007), and COSEBIs (Schneider et al., 2010) which can be calculated from these $\widehat{\xi}_\pm(\theta)$ estimates and are intended to isolate informative characteristics of the correlation functions.

The standard inference paradigm proceeds by assuming that the so-called *data vector* (either $\xi_\pm$ or some transformation thereof) has a multivariate normal distribution. Then, the likelihood of a cosmological parameter set θ is computed via

$$\mathcal{L}(\theta; X) = \frac{1}{\sqrt{(2\pi)^p|\mathbf{C}|}} \exp\left(-\frac{1}{2}(X - g(\theta))^T \mathbf{C}^{-1}(X - g(\theta))\right) \tag{5.2}$$

where X is the empirically computed data vector, $g(\theta)$ is the theoretical value of the data vector given parameters θ, and $\mathbf{C}$ is the covariance matrix of the data vector. This likelihood can be evaluated either as part of a frequentist maximum likelihood analysis or, as is more often the case, in a Bayesian framework in conjunction with a posterior sampling scheme, as in, e.g., Fu et al. (2014).

Additionally, some analyses (e.g., Heymans et al., 2013) attempt to mitigate bias due to the intrinsic alignment of galaxies by dividing galaxies into bins in redshift z as well as angular separation θ and estimating shear auto- and cross-correlations with respect to those redshift bins. This increases the size of the data vector by a multiplicative factor of $N_t + \binom{N_t}{2}$, where N_t is the number of redshift bins. Still other analyses (e.g. Fu et al., 2014) augment the data vector with estimates of three- and four-point correlation functions (or their Fourier space analogues, the bispectrum and trispectrum, respectively) in order to incorporate information not captured by two-point correlation function estimates.

Some recent work has called into question the validity of the assumption on the form of the likelihood in (5.2). Keitel & Schneider (2011) demonstrate analytically that the distribution of two-point correlation function estimates is noticeably non-Gaussian even in the simplified setting where lensing shear is distributed according to a Gaussian random field model. Investigating the impact of using a Gaussian likelihood approximation on parameter constraints, Hartlap et al. (2009) find that an approach based on ICA yields considerably narrower credible intervals.

Additionally, there is not universal consensus on the optimal approach to obtaining the covariance matrix $\mathbf{C}$ in (5.2) above. Standard approaches include estimating the covariance matrix by resampling from the data; calculating covariance matrices analytically;

and using large numbers of simulations to estimate $\mathbf{C}$. Each approach has advantages and disadvantages; resampling requires making strong assumptions known to be invalid, while analytical covariances are difficult to calculate in all but the simplest cases.

The procedure of estimating the covariance matrix from simulations seems promising but is not without its shortcomings. Performing enough simulations to estimate a different covariance matrix at many different points in parameter space would be computationally prohibitive, so parameter values are typically fixed at a fiducial cosmology value for the covariance matrix simulations. In the context of the simulation-based covariance matrix approach, Dodelson & Schneider (2013) quantify the relationship between the number of simulations used and parameter uncertainty due to covariance matrix uncertainty. Sellentin & Heavens (2015) demonstrate that, for likelihood analyses using a covariance matrix estimated from large number of simulations, the appropriate distribution of the data vector is no longer Gaussian but rather a multivariate t-distribution derived in the paper.

In addition, any approach (e.g., Markov Chain Monte Carlo) involving the evaluation of a Gaussian likelihood, up to a normalizing constant, actually requires the inverse covariance or *precision* matrix $\mathbf{C^{-1}}$. The standard approach (also the maximum-likelihood estimate) is to simply invert an estimate of the covariance matrix; however, in addition to known biases (Hartlap et al., 2007), any estimator requiring the inversion of a random matrix may encounter issues with singularity (requiring a psuedo-inverse) and/or numerical stability.

Fortunately, forward simulation models are available to generate data realizations given cosmological input parameters; these simulation models range from comparatively simple Gaussian random fields to, e.g., the SUNGLASS simulations of Kiessling et al. (2011) that incorporate the effects of non-linear evolution. This suggests weak lensing as a natural area of application for ABC, as for these more complex models it is impossible to analytically specify a likelihood function. However, including the higher-order correlation functions and tomographic binning necessary to preserve the rich information from these simulations will increase the natural dimension of the summary statistic $S(\cdot)$. Thus, developing a method for reducing the dimension of the summary statistic while preserving the information relevant for inference will be a crucial step in implementing an ABC approach to weak lensing.

5.2 2D simulation study

In this section, we present an application of the common space mapping method to construct summary statistics for ABC inference of cosmological parameters from simulated weak lensing data. We employ a relatively simple simulation mechanism wherein galaxy shears are distributed according to a Gaussian random field whose autocorrelation function is dictated by the values of (Ω_M, σ_8) and our data vector (before any dimension reduction) contains simply the empirically estimated 2PCFs $\widehat{\xi}_{\pm}$.

5.2.1 Structure

Our aim is inference on the posterior distribution of (Ω_M, σ_8) conditional on some observed data (which is itself simulated). We generate this so-called "observed" data using input parameters $\Omega_M = 0.25$ and $\sigma_8 = 0.8$, fixing other parameters $h = 0.70$, $\Omega_b = 0.045$, and using a single redshift $z = 0.7$. After evaluating the power spectrum using the NICAEA engine (Kilbinger et al., 2009), as wrapped by the `lenstools` python library (Petri, 2015), we generate a cosmic shear grid realization using the `GalSim` toolkit (Rowe et al., 2015) and add shape noise. We then summarize the shear grid by calculating $\widehat{\xi}_{\pm}(\theta)$ using the `TreeCorr` package, which implements the algorithms described in Jarvis et al. (2004). We construct the observed data vector thus; the same process can be used to produce any number of simulated data vectors for different input (Ω_M, σ_8) values.

For our ABC summary statistic, we use the first two dimensions of $\widehat{T}$, the common space mapping, with the tuning parameter λ chosen to maximize the approximate sufficiency criterion $\widehat{I}(\widehat{T}_\lambda, \theta)$ described in Section 3.2.1. As a distance metric, we use Euclidean distance between observed and simulated $\widehat{T}$ values. We implement the ABC SMC algorithm, drawing particles initially from a uniform prior distribution on each parameter: $\Omega_M \in [0.1, 0.8], \sigma_8 \in [0.5, 1.0]$, with all other parameters held fixed.

5.2.2 Results

Figure 5.1 shows ABC SMC samples (after 10 SMC steps) from a simulated lensing study as described above. In this case, we simulate a 100 × 100 shear grid spaced 0.1° apart, i.e., a 10° × 10° patch of sky, and add mean-zero i.i.d. normal shape noise with $\sigma = 0.00811$, corresponding to a value of 20 galaxies per square arcminute. We compare the results using two different summary statistics: the estimated 2PCFs $\widehat{\xi}_{\pm}$ (left) and the CSMM-derived $\widehat{T}$ (right), retaining only the first two dimensions. For the CSMM-derived $\widehat{T}$, we choose the tuning parameter $\lambda = 1$, as it maximizes $\widehat{I}(\widehat{T}_{\lambda}, \theta)$ over a set of logarithmically-spaced candidate λ values. We note that the dots are drawn with area corresponding to the ABC-SMC weights.

The orange dotted line in each plot represents degeneracy curve ${\Omega_M}^{\alpha}\sigma_8$ corresponding to the input parameter values, plotted with the value $\alpha = 0.55$ following the lead of Petri et al. (2015), who put forth a derivation for the degeneracy exponent's optimal value, which they deem "consistent with what is found in the literature." We see from the bottom row of Figure 5.1 that after 9 ABC SMC steps, there is little difference between the ABC posterior samples obtained using the two different summary statistics, as both sets of samples lie close to the input degeneracy curve; this suggests that summarizing via the common-space mapping method does not diminish the quality of inference in this example. However, examining the top row, we see that after only 3 ABC SMC iterations, the ABC sample drawn using $\widehat{\xi}_{\pm}$ is quite diffuse, while the sample drawn using the CSMM-derived $\widehat{T}$ is more concentrated on the input degeneracy curve — in fact, it is already almost as concentrated as it eventually becomes. This suggests that the dimension reduction achieved by the CSMM has helped to speed up the convergence of the ABC algorithm.

5.2.3 Comparison to alternative approaches

The appeal of using ABC for parameter inference is that it permits the relaxation of the assumption, made in traditional approaches to parameter constraint via weak lensing, that the distribution of the data vector $\widehat{\xi}_{\pm}$, and thus the form of the likelihood, is multivariate

Gaussian. To investigate the impact of making or relaxing this assumption, we employ two alternative approaches: MCMC using the Metropolis-Hastings algorithm, and ABC making the same multivariate Gaussian assumption.

The first of these involves using MCMC to sample from the posterior distribution, which when using a uniform prior is simply proportional to the likelihood given in (5.2). To obtain the needed precision matrix $\mathbf{C^{-1}}$, we simulate many (i.e., 10^5) realizations of the $\widehat{\xi}_\pm$ data vector and invert their sample covariance matrix $\widehat{\mathbf{C}}$. As simulating enough data to compute an inverse sample covariance matrix at each point in parameter space would be prohibitive, we follow the standard practice of using a $\mathbf{C}^{-1}$ calculated from data simulated under one fiducial cosmology. For the fiducial cosmology values, we use $\Omega_M = 0.308$, $\sigma_8 = 0.815$, following some of those found in Planck Collaboration et al. (2015). In principle, this practice of not allowing the precision matrix to vary across parameter space should introduce an additional source of error in posterior inference.

The second alternative approach constitutes something of a compromise between truly likelihood-free ABC and MCMC with the Gaussian likelihood assumption. This compromise approach employs the ABC framework but simplifies the step of simulating candidate data by instead adding noise to the theory-predicted value of the $\widehat{\xi}_\pm$ data vector. The noise is simulated from a multivariate Gaussian distribution with covariance matrix $\mathbf{C}$ obtained, similarly to that in the above MCMC approach, by calculating the sample covariance matrix $\widehat{\mathbf{C}}$ of many data vectors simulated under a fiducial cosmology. A key difference in this approach is that, as opposed to calculating a likelihood, the process of obtaining samples from a multivariate Gaussian distribution does not require matrix inversion, a somewhat numerically unstable operation; for sampling, it suffices to compute the Cholesky decomposition of the covariance matrix.

Figure 5.2 shows posterior samples obtained via these alternative approaches as compared to those (at right) obtained via ABC (using the full $\widehat{\xi}_\pm$ data vector). We note that the MCMC chain is subsampled so that the number of particles is comparable to the number of particles resulting from the ABC approaches. In this particular exercise, we find that the impact of the Gaussian likelihood assumption on the resulting posterior sample appears small. We do note, however, that the ABC approaches seem to result in posterior samples that more richly capture the range of values near the input parameter degeneracy, while the MCMC seems to concentrate heavily in a particular subregion.

5.3 Application to CFHTLens data

In this section, we present the parameter constraints resulting from an ABC analysis of real weak lensing survey data. We use data from the Canada-France-Hawaii Lensing Survey (Heymans et al. 2012), hereafter referred to as CFHTLenS. The CFHTLenS survey analysis combined weak lensing data processing with THELI (Erben et al., 2013) and shear measurement with *lensfit* (Miller et al., 2013). A full systematic error analysis of the shear measurements in combination with the photometric redshifts is presented in Heymans et al. (2013), with additional error analyses of the photometric redshift measurements presented in Benjamin et al. (2013).

For our observed data, we use the galaxies in the CFHTLens W2 field, which includes a $5^\circ \times 5^\circ$ patch of sky; we retain only galaxies from those camera pointings which passed the systematics tests of Heymans et al. (2013). Apart from the observed data, the structure of the analysis is the same as that of the simulation studies described in Section 5.2.1; this includes using the same prior distribution on (Ω_M, σ_8) and the same pipeline for simulating candidate data.

Figure 5.3 shows the final weighted posterior sample resulting from ABC SMC analyses without (left) and with (right) dimension reduction via the CSMM. The green curves represent weighted least-squares fits to the (Ω_M, σ_8) degeneracy, whose banana-like shape is consistent with previous weak lensing analyses. The fitted degeneracy curve is $\sigma_8\Omega_M^{-0.63} = 0.35$ for the full data vector analysis and $\sigma_8\Omega_M^{-0.58} = 0.36$ for the analysis after dimension reduction. The similarity between the left and right plots suggests that summarizing via the CSMM has not diminished the quality of posterior inference; hence, the lower-dimensional CSMM-derived statistic is, in some sense, close to sufficient.

Our development of a dimension reduction scheme was motivated by the concern that failing to summarize a high-dimensional observable could result in computation time that is substantially longer — perhaps prohibitively so. Even in the context of this relatively simple setup, we find that reducing the summary statistic dimension via our method results in an ABC posterior that concentrates more quickly.

For any weighted posterior sample, we measure concentration by fitting a two-parameter curve ($\sigma_8 \Omega_M^{\alpha} = c$) to the sample, minimizing the weighted sum of squared differences from the sample measured perpendicularly to the curve. The value of this weighted sum of squared differences ($wSSE$) to the best-fit degeneracy curve, changing as the ABC SMC algorithm updates the posterior sample particles, is shown in Figure 5.4; smaller values of $wSSE$ indicate a more concentrated posterior estimate. In this simple example, we note that after roughly 45 minutes of computation time (run on a 2012 MacBook Pro, using one 2.5GHz processor), the posterior estimate based on $\widehat{T}$ has concentrated about as much as it does after four hours, while the posterior estimate based on $\widehat{\xi}_{\pm}$ remains comparatively diffuse.

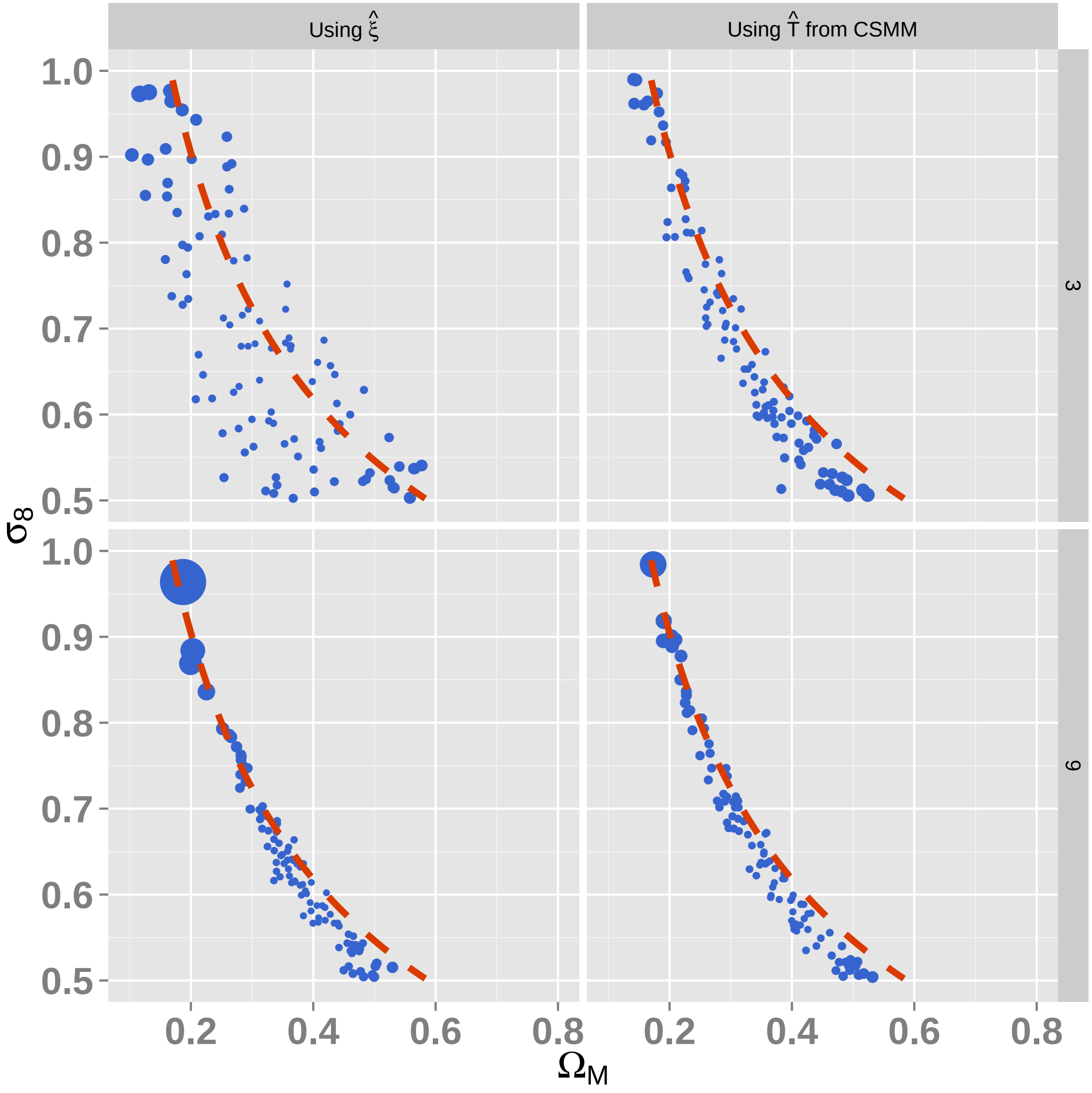

Figure 5.1: Samples resulting from an ABC SMC run (with size corresponding to particle weight) using the estimated 2PCFs $\widehat{\xi}_{\pm}$ (left column) and the CSMM-derived $\widehat{T}$ (right column) as summary statistic. The results are shown after 3 (top row) and 9 (bottom row) ABC SMC iterations.

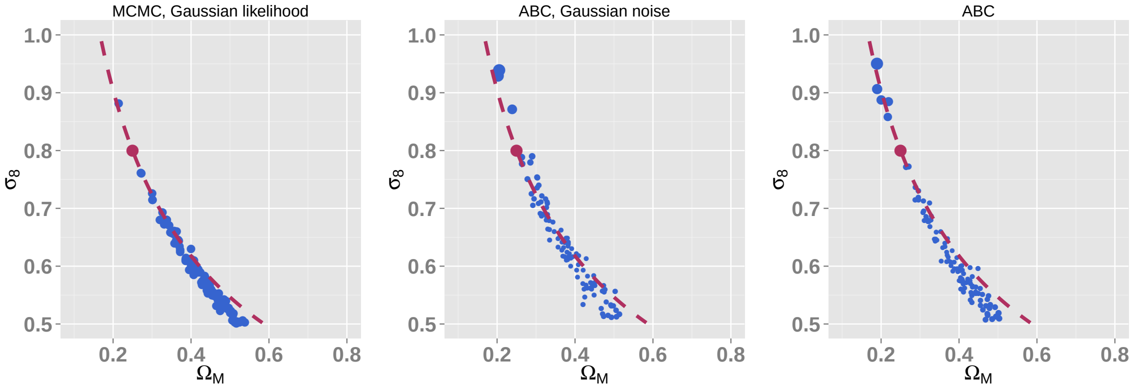

Figure 5.2: Comparison of posterior sampling approaches

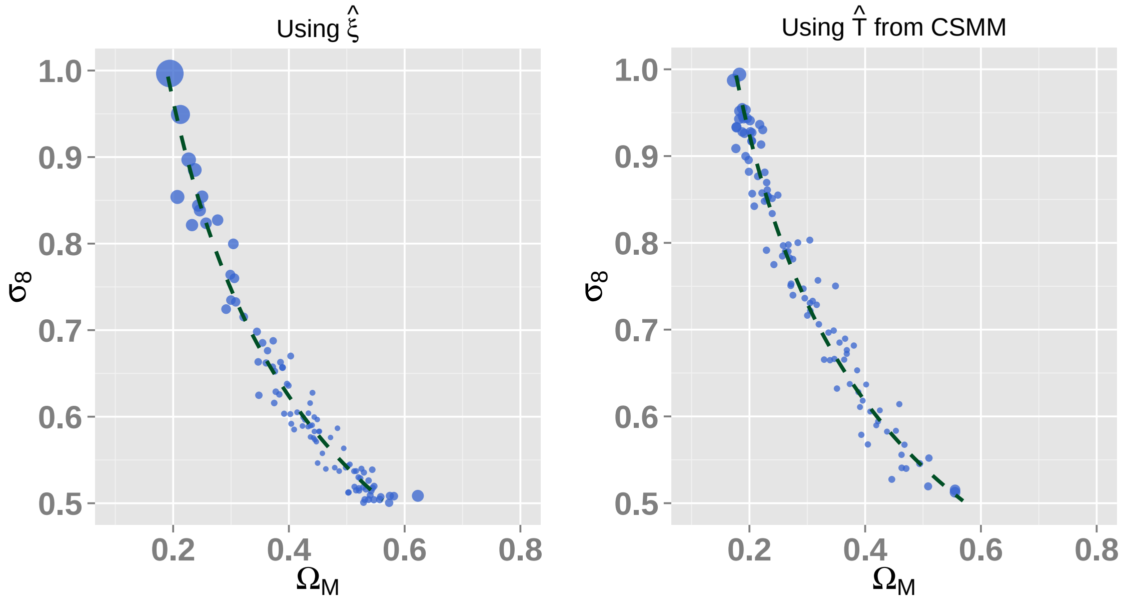

Figure 5.3: Final ABC SMC posterior sample conditional on CFHTLens W2 field data, using full $\widehat{\xi}_{\pm}$ data vector (left) and summarizing via the CSMM-derived $\widehat{T}$ (right). Also shown (dark green) are weighted least-squares curves fit to capture the (Ω_M, σ_8) degeneracy.

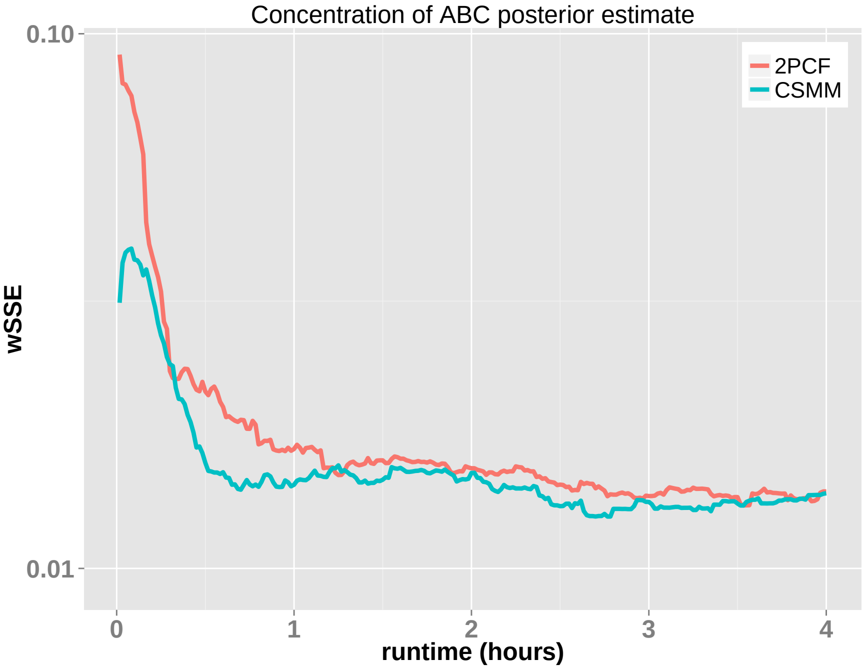

Figure 5.4: Evolution of posterior concentration over computation time, without (red) and with (teal) dimension reduction, where concentration is measured by the weighted sum of squared differences ($wSSE$) between the sample and the best-fit degeneracy curve. Note that $wSSE$ values are shown on a logarithmic scale.

Chapter 6

Software

In this chapter, we describe the Python module `csmm` which can be used to obtain the common space mappings introduced in Chapter 3 for use as ABC summary statistics.

6.1 Installation

The `csmm` module code can be obtain by downloading the `csmm.py` file from the repository located at `bitbucket.org/mcvespe/research_public/`. The user will also need to download the `entropy_estimators.py` file from the NPEET toolbox (Ver Steeg, 2014), available at `github.com/gregversteeg/NPEET`. The `csmm` module also depends on the standard Python libraries `NumPy` and `SciPy` for scientific computation and `matplotlib` for visualization.

Provided that the `csmm.py` and `entropy_estimators.py` files are in the appropriate working directory, the user will be able to `import csmm` in order to access the procedures described in the subsequent sections.

6.2 The `csmm` class

The `csmm` class handles the construction of the common-space mapping from a training set of N_T corresponding (θ_i, x_i) pairs.

6.2.1 Constructing a `csmm` object

To obtain the common space mapping, the first step is to construct an object of the `csmm` class, which requires the following inputs:

- `theta_train`: a matrix of training parameter values (of dimension $N_T \times p$)

- `x_train`: a matrix of training data (of dimension $N_T \times d$); each row should correspond to a row in theta_train
- `lambda_aff`: the value of the affinity tuning parameter λ
- `bw_quantile` (optional; default = 0.05): the quantile used to determine the bandwidth for the within-parameter and within-data kernel matrices $\mathbf{K}_\theta$ and $\mathbf{K}_x$
- `kernel_order` (optional; default = 1): the order of the exponent in the kernel similarity matrices $\mathbf{K}_\theta$ and $\mathbf{K}_x$; e.g., 1 for exponential, 2 for Gaussian
- `to_sort` (optional; default = `True`): whether to consider the order statistics of the data when evaluating similarity; should be `False` unless the different dimensions of the data are independent draws from the same distribution

A call to `csmm()` with the above inputs instantiates an object of the `csmm` class.

6.2.2 `csmm` object methods

A `csmm` object has three public methods: `nystrom`, `approx_suff`, and `plot_csmm`.

The `nystrom` method, which maps a new data point into the common space, is the vehicle for using the common space mapping as an ABC summary statistic. The method takes two arguments:

- `x_new`: a matrix of new data points (of dimension $N_{new} \times d$) to project into the common space
- `dims_keep`: a list of the (user-determined) dimension indices of the common space, e.g. `range(3)`

The `approx_suff` method returns an estimate of either the pairwise information criterion of Section 3.2.2 or the direct information criterion $\widehat{I}(\theta; \widehat{T}(X))$; these criteria are intended to guide the choice of the affinity tuning parameter λ. The method takes two arguments:

- `dims_keep`: a list of the (user-determined) dimensions of the common space, as in the `nystrom` method

- pairwise_mi (optional; default = False): whether to return the pairwise criterion

The plot_csmm method outputs, to a specified file, a diagnostic visualization of the common-space mapping. The method has three arguments:

- color_arg: a list (of length N_T) of values dictating the color of the points in the diagnostic plot

- file_arg: the file name where the diagnostic plot will be saved

- dims_arg (optional; default = range(2)): a list (of length 2) of the dimension indices of the common-space mapping to depict

Figure 6.1 shows four example plots produced by the plot_csmm function for a simulated weak lensing example. The plots show, for a particular two (user-specified) dimensions of the common space, the result of the $\widehat{\eta}(\theta)$, shown as triangles, and the $\widehat{T}(x)$ mappings, shown as squares. Corresponding common-space mapped $\widehat{\eta}(\theta_i)$ and $\widehat{T}(x_i)$ pairs are connected by gray lines and are colored according to a user-specified coloring scheme; in this example, it is the parameter degeneracy value ${\Omega_M}^{\alpha}\sigma_8$ using $\alpha = 0.55$.

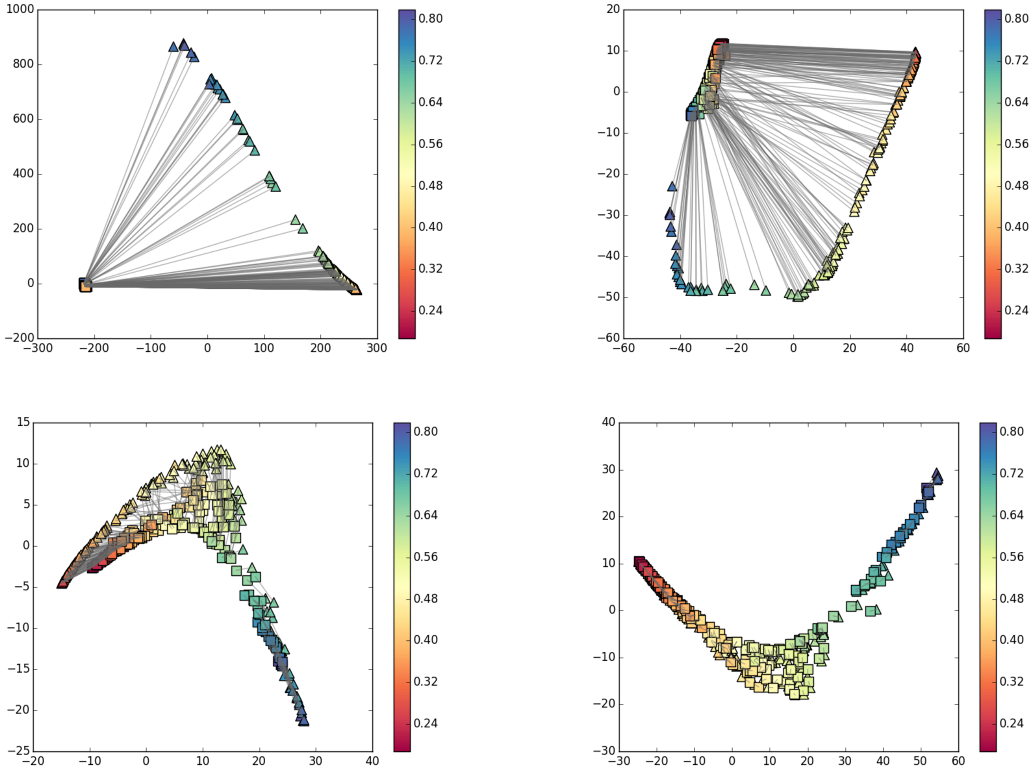

Figure 6.1: Four `plot_csmm` plots, showing the common space mappings for simulated weak lensing data, using $\lambda = 2^{-5}$ (top left), $\lambda = 2^{-2}$ (top right), $\lambda = 2^{1}$ (bottom left), and $\lambda = 2^{4}$ (bottom right).

The key feature of these plots is that as λ increases, the corresponding $\widehat{\eta}(\theta_i)$ and $\widehat{T}(x_i)$ pairs are pulled closer together. Heuristically, this suggests that the affinity parameter λ dictates the relative priority the mapping places on the connection between parameters and corresponding data.

6.3 Note on usage

Our approach in applying the method has been to use the same training set of parameters and data to fit many models, computing the associated "direct" information criterion for each of a set of candidate λ values, and ultimately choosing the value of λ that maximizes that information criterion. This has tended to yield mappings that perform reasonably well as ABC summary statistics in situations where the true posterior distribution is known.

Chapter 7

Concluding thoughts

7.1 Discussion

Any Bayesian parameter inference problem in which the data-generating process is more easily specified by a forward model than by an explicit likelihood lends itself to a natural application of ABC. Weak gravitational lensing belongs to this category; although simulation models allow for varying degrees of underlying complexity, the canonical approach to parameter constraint involves making a very restrictive assumption on the form of the likelihood so that it becomes tractable. Moreover, assuming a multivariate Gaussian likelihood can itself be computationally burdensome due to the simulations needed to produce the covariance matrix used therein.

While ABC holds some appeal for its relaxation of certain distributional assumptions, it introduces two other potential sources of error: approximation error, due to allowing the summarized data to differ by some tolerance ϵ, and error due to summarizing by a (possibly) non-sufficient statistic. In any setting where the data dimension is large, summarizing both observed and simulated data is unavoidable. (We remark that Monte Carlo error resulting from using a finite sample to infer a distribution, present in MCMC analyses, persists in ABC as well.)

Approximation error can be reduced (though perhaps not eliminated) by sequential resampling schemes that allow ϵ to shrink ever smaller. The error due to conditioning on a non-sufficient summary statistic can be addressed by choosing, as we do via the CSMM, a summary statistic that optimizes some criterion of approximate sufficiency. Optimizing this criterion guides the evaluation of candidate CSMM summary statistics on a relative basis but provides no absolute guarantee of near-sufficiency.

We also note that simply maximizing $\widehat{I}(\theta; \widehat{T_\lambda}(X))$ over candidate values of λ, while supported by theory, may not be the optimal approach for choosing λ. In the toy Gaussian simulations of Section 4.1, we found that the quality of posterior estimation (as measured by KL loss) does not diminish as λ is increased beyond the empirical maximizer of $\widehat{I}(\theta; \widehat{T_\lambda}(X))$, though the computation time increases noticeably.

To the extent that the common-space mapping method (as tuned via the sufficiency gap criterion) yields summary statistics that are approximately sufficient, there exists the potential for application beyond the ABC context in which we have presented it. In principle, any (frequentist or Bayesian) decision-theoretic procedure that relies on the calculation of a sufficient statistic could thus be rendered in approximate form.

Regarding the weak lensing applications of Chapter 5, we acknowledge that their contribution is primarily as a proof of concept; even the posterior estimates obtained from the real CFHTLens data should not be simply be accepted without further examination. The simulation process, adding i.i.d. shape noise to shears drawn from a two-dimensional Gaussian random field, is clearly oversimplified, neglecting to incorporate such phenomena as non-Gaussianity, redshift-shear dependence, intrinsic galaxy alignment, and cosmic variance. Additionally, the analyses we present explore only the space of Ω_M and σ_8, keeping other cosmological and nuisance parameters fixed.

We expect that any future ABC lensing studies would employ more sophisticated simulation methods and explore higher-dimensional parameter subspaces than the studies in this paper. However, we are encouraged that (a) the CSMM appears to accomplish dimension reduction without materially reducing the quality of estimation; and (b) that the ABC posterior samples obtained from the CFHTLens data appear roughly consistent (in terms of both shape and values) with previously obtained constraints, despite the considerable difference in approaches.

7.2 Future directions

Both the methodological and applied work contained in the preceding sections of this report offer opportunities for further inquiry and development.

7.2.1 Methodology

Common space dimension. Thus far, we have elected to set the dimension of the common space (which is at the user's discretion) equal to the dimension of the parameter space we are attempting to constrain. This is consistent with an intuition, explained as a consequence of Theorem 3 in Fearnhead & Prangle (2012), that the ideal summary statistic contains one component per parameter dimension. In the examples we have considered, the resulting common space mappings seem to perform adequately as ABC summary statistics. Further theoretical development could either serve to justify this practice or to guide a more principled choice of the common space dimension.

Information criteria. Our current approach for choosing the affinity parameter λ is to optimize, over a set of candidate values, an empirical estimate of the criterion $I(\theta; \widehat{T}(X))$ as laid out in Section 3.2. We remarked in that section that our approach has relied on nearest-neighbor estimators of mutual information which are adequate when the parameter and the common space are low-dimensional but may break down in higher dimensions due to the curse of dimensionality.

As an alternative in higher dimensions, we suggest a pairwise information criterion, which we justified heuristically via a connection to a one-to-one-ness condition, shown in Frazier et al. (2015) to be necessary in order for an ABC posterior estimate to achieve Bayesian consistency. We also observe that, in both toy problems and more complex settings, the pairwise information criterion seems to roughly track the direct information criterion for different regions of λ. While the two are not always maximized by the same value of λ, one tends to be larger in regions where the other is larger, and vice versa. With that said, further exploration of the theoretical properties of the pairwise criterion, and perhaps its relationship to the direct criterion, could be useful in providing a more rigorous justification for its use.

Moreover, closer examination of the theoretical properties of the sufficiency gap, particularly with regard to the connection to the Le Cam and Shannon deficiencies of Raginsky (2011), could yield additional insight into practical use of the information criteria.

7.2.2 Weak lensing applications

More realistic simulations. The lensing simulations of Section 5.2 present, as a proof-of-concept, the constraints on two parameters in the ΛCDM cosmological model that would result from observing data that is itself simulated. The simulation pipeline used therein (for generating both "observed" data and ABC candidate data) is quite simplistic compared to the state of the art, for a variety of reasons which we outline below. As a result, incorporating additional complexity into the simulation model would not only improve the parameter constraint procedure but also amplify the need for principled reduction of the data dimension. Indeed, in the existing simulations, it is computationally feasible to obtain an ABC posterior without any dimension reduction of the 2PCF data vector $\widehat{\xi}_{\pm}(\theta)$, albeit with more ABC SMC iterations required for convergence.

The current simulation pipeline uses the `GalSim` toolkit to generate a grid of shears from a Gaussian random field model, summarizing the data initially via the 2PCFs $\widehat{\xi}_{\pm}(\theta)$, which preserve the relevant cosmological information in the GRF setting. However, as cosmic shear is not perfectly described by a Gaussian random field model, there is additional information contained in the three-point correlation functions, denoted $\Gamma^{(0,1,2,3)}$. Efficient estimators, described by Jarvis et al. (2004) and implemented in the `TreeCorr` package, exist for these quantities, but the inferential value of incorporating them into the ABC data vector hinges on the capability of the simulation procedure to include non-Gaussian effects.

A similar situation arises in the context of so-called tomographic analyses, which take into account the cosmic shear of galaxies at varying redshifts. In our admittedly simplistic simulations, both the "observed" and ABC candidate data is simulated as a flat shear grid at a single redshift $z = 0.7$. Realistic survey data, of course, contain galaxies that vary in redshift; tomographic lensing analyses proceed by dividing galaxies into redshift bins and computing cross-correlation functions between pairs of redshift bins (and auto-correlation functions within those bins.) Computing these cross-correlations is not difficult, but as with 3PCFs, their inferential value is limited unless the simulation pipeline incorporates some structural dependence between shear and redshift.

In both situations, including either or both of the 3PCFs (for non-Gaussianity) and the tomographic cross-correlations in the data vector will substantially increase its dimension, to the point that an ABC analysis using the data vector as is would indeed be computationally prohibitive. Thus, in these more complex settings, a principled dimension reduction scheme, such as the CSMM, would be crucial for ABC shear analyses to be feasible.

References

Barber, S., Voss, J., & Webster, M. (2013). The rate of convergence for approximate bayesian computation. *arXiv preprint arXiv:1311.2038*.

Beaumont, M. A., Cornuet, J.-M., Marin, J.-M., & Robert, C. P. (2009). Adaptive approximate bayesian computation. *Biometrika*, (p. asp052).

Benjamin, J., Van Waerbeke, L., Heymans, C., Kilbinger, M., Erben, T., Hildebrandt, H., Hoekstra, H., Kitching, T. D., Mellier, Y., Miller, L., et al. (2013). Cfhtlens tomographic weak lensing: quantifying accurate redshift distributions. *Monthly Notices of the Royal Astronomical Society*, *431*(2), 1547–1564.

Blum, M. G. (2012). Approximate bayesian computation: a nonparametric perspective. *Journal of the American Statistical Association*.

Blum, M. G., Nunes, M. A., Prangle, D., Sisson, S. A., et al. (2013). A comparative review of dimension reduction methods in approximate bayesian computation. *Statistical Science*, *28*(2), 189–208.

Cameron, E., & Pettitt, A. (2012). Approximate bayesian computation for astronomical model analysis: a case study in galaxy demographics and morphological transformation at high redshift. *Monthly Notices of the Royal Astronomical Society*, *425*(1), 44–65.

Coifman, R. R., Kevrekidis, I. G., Lafon, S., Maggioni, M., & Nadler, B. (2008). Diffusion maps, reduction coordinates, and low dimensional representation of stochastic systems. *Multiscale Modeling & Simulation*, *7*(2), 842–864.

Coifman, R. R., & Lafon, S. (2006). Diffusion maps. *Applied and computational harmonic analysis*, *21*(1), 5–30.

Cover, T. M., & Thomas, J. A. (2012). *Elements of information theory*. John Wiley & Sons.

Csilléry, K., Blum, M. G., Gaggiotti, O. E., & François, O. (2010). Approximate bayesian computation (abc) in practice. *Trends in ecology & evolution*, *25*(7), 410–418.

Dodelson, S. (2003). *Modern cosmology*. Academic press.

Dodelson, S., & Schneider, M. D. (2013). The effect of covariance estimator error on cosmological parameter constraints. *Phys. Rev. D*, *88*, 063537.
URL `http://link.aps.org/doi/10.1103/PhysRevD.88.063537`

Drovandi, C. C., Pettitt, A. N., & Faddy, M. J. (2011). Approximate bayesian computation using indirect inference. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *60*(3), 317–337.

Drovandi, C. C., Pettitt, A. N., & Lee, A. (2015). Bayesian indirect inference using a parametric auxiliary model. *Statistical Science*, *30*(1), 72–95.

Erben, T., Hildebrandt, H., Miller, L., van Waerbeke, L., Heymans, C., Hoekstra, H., Kitching, T., Mellier, Y., Benjamin, J., Blake, C., et al. (2013). Cfhtlens: the canada–france–hawaii telescope lensing survey–imaging data and catalogue products. *Monthly Notices of the Royal Astronomical Society*, *433*(3), 2545–2563.

Fearnhead, P., & Prangle, D. (2012). Constructing summary statistics for approximate bayesian computation: semi-automatic approximate bayesian computation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *74*(3), 419–474.

Frazier, D. T., Martin, G. M., & Robert, C. P. (2015). On consistency of approximate bayesian computation. *arXiv preprint arXiv:1508.05178*.

Fu, L., Kilbinger, M., Erben, T., Heymans, C., Hildebrandt, H., Hoekstra, H., Kitching, T. D., Mellier, Y., Miller, L., Semboloni, E., et al. (2014). Cfhtlens: Cosmological constraints from a combination of cosmic shear two-point and three-point correlations. *Monthly Notices of the Royal Astronomical Society*, *441*(3), 2725–2743.

Gourieroux, C., & Monfort, A. (1997). *Simulation-based econometric methods*. Oxford University Press.

Gourieroux, C., Monfort, A., & Renault, E. (1993). Indirect inference. *Journal of applied econometrics*, *8*, S85–S85.

Hartlap, J., Schrabback, T., Simon, P., & Schneider, P. (2009). The non-gaussianity of the cosmic shear likelihood or how odd is the chandra deep field south? *Astronomy and Astrophysics*, *504*, 689–703.

Hartlap, J., Simon, P., & Schneider, P. (2007). Why your model parameter confidences might be too optimistic. unbiased estimation of the inverse covariance matrix. *Astronomy & Astrophysics*, *464*(1), 399–404.

Heymans, C., Grocutt, E., Heavens, A., Kilbinger, M., Kitching, T. D., Simpson, F., Benjamin, J., Erben, T., Hildebrandt, H., Hoekstra, H., et al. (2013). Cfhtlens tomographic weak lensing cosmological parameter constraints: Mitigating the impact of intrinsic galaxy alignments. *Monthly Notices of the Royal Astronomical Society*, *432*(3), 2433–2453.

Hoekstra, H., & Jain, B. (2008). Weak gravitational lensing and its cosmological applications. *Annual Review of Nuclear and Particle Science*, *58*, 99–123.

Jarvis, M., Bernstein, G., & Jain, B. (2004). The skewness of the aperture mass statistic. *Monthly Notices of the Royal Astronomical Society*, *352*(1), 338–352.

Jiang, B., Wu, T.-y., Zheng, C., & Wong, W. H. (2015). Learning summary statistic for approximate bayesian computation. *arXiv preprint arXiv:1510:02175*.

Joyce, P., & Marjoram, P. (2008). Approximately sufficient statistics and bayesian computation. *Statistical Applications in Genetics and Molecular Biology*, *7*(1).

Keitel, D., & Schneider, P. (2011). Constrained probability distributions of correlation functions. *Astronomy & Astrophysics*, *534*, A76.

Kiessling, A., Taylor, A., & Heavens, A. (2011). Simulating the effect of non-linear mode coupling in cosmological parameter estimation. *Monthly Notices of the Royal Astronomical Society*, *416*(2), 1045–1055.

Kilbinger, M., Benabed, K., Guy, J., Astier, P., Tereno, I., Fu, L., Wraith, D., Coupon, J., Mellier, Y., Balland, C., et al. (2009). Dark-energy constraints and correlations with systematics from cfhtls weak lensing, snls supernovae ia and wmap5. *Astronomy & Astrophysics*, *497*(3), 677–688.

Kraskov, A., Stögbauer, H., & Grassberger, P. (2004). Estimating mutual information. *Physical review E*, *69*(6), 066138.

Le Cam, L. (1964). Sufficiency and asymptotic sufficiency. *Ann. Math. Statist*, *35*, 1419–1455.

Lin, C.-A., & Kilbinger, M. (2015). A new model to predict weak-lensing peak counts-ii. parameter constraint strategies. *Astronomy & Astrophysics*, *583*, A70.

Mandelbaum, R., Rowe, B., Bosch, J., Chang, C., Courbin, F., Gill, M., Jarvis, M., Kannawadi, A., Kacprzak, T., Lackner, C., et al. (2014). The third gravitational lensing accuracy testing (great3) challenge handbook. *The Astrophysical Journal Supplement Series*, *212*(1), 5.

Marin, J.-M., Pudlo, P., Robert, C. P., & Ryder, R. J. (2012). Approximate bayesian computational methods. *Statistics and Computing*, *22*(6), 1167–1180.

Marjoram, P., Molitor, J., Plagnol, V., & Tavaré, S. (2003). Markov chain monte carlo without likelihoods. *Proceedings of the National Academy of Sciences*, *100*(26), 15324–15328.

Miller, L., Heymans, C., Kitching, T., Van Waerbeke, L., Erben, T., Hildebrandt, H., Hoekstra, H., Mellier, Y., Rowe, B., Coupon, J., et al. (2013). Bayesian galaxy shape measurement for weak lensing surveys–iii. application to the canada–france–hawaii telescope lensing survey. *Monthly Notices of the Royal Astronomical Society*, *429*(4), 2858–2880.

Munshi, D., Valageas, P., Van Waerbeke, L., & Heavens, A. (2008). Cosmology with weak lensing surveys. *Physics Reports*, *462*(3), 67–121.

Nunes, M. A., & Balding, D. J. (2010). On optimal selection of summary statistics for approximate bayesian computation. *Statistical applications in genetics and molecular biology*, *9*(1).

Petri, A. (2015). The lenstools computing package.
URL `http://dx.doi.org/10.5281/zenodo.32462`

Petri, A., Liu, J., Haiman, Z., May, M., Hui, L., & Kratochvil, J. M. (2015). Emulating the cfhtlens weak lensing data: Cosmological constraints from moments and minkowski functionals. *Physical Review D*, *91*(10), 103511.

Planck Collaboration, et al. (2015). Planck 2015 results. xiii. cosmological parameters. *arXiv preprint arXiv:1502.01589*.

Prangle, D. (2015). Summary statistics in approximate bayesian computation. *arXiv preprint arXiv:1512.05633*.

Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A., & Feldman, M. W. (1999). Population growth of human y chromosomes: a study of y chromosome microsatellites. *Molecular Biology and Evolution*, *16*(12), 1791–1798.

Raginsky, M. (2011). Shannon meets blackwell and le cam: Channels, codes, and statistical experiments. In *Information Theory Proceedings (ISIT), 2011 IEEE International Symposium on*, (pp. 1220–1224). IEEE.

Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P. J., Lander, E. S., Mitzenmacher, M., & Sabeti, P. C. (2011). Detecting novel associations in large data sets. *science*, *334*(6062), 1518–1524.

Rowe, B., Jarvis, M., Mandelbaum, R., Bernstein, G., Bosch, J., Simet, M., Meyers, J., Kacprzak, T., Nakajima, R., Zuntz, J., et al. (2015). Galsim: The modular galaxy image simulation toolkit. *Astronomy and Computing*, *10*, 121–150.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, *12*(4), 1151–1172.

Ruli, E., Sartori, N., & Ventura, L. (2015). Approximate bayesian computation with composite score functions. *Statistics and Computing*.

Schneider, P., Eifler, T., & Krause, E. (2010). Cosebis: Extracting the full e-/b-mode information from cosmic shear correlation functions. *A&A*, *520*, A116.

Schneider, P., & Kilbinger, M. (2007). The ring statistics-how to separate e-and b-modes of cosmic shear correlation functions on a finite interval. *Astronomy and Astrophysics*, *462*, 841–849.

Schneider, P., van Waerbeke, L., Kilbinger, M., & Mellier, Y. (2002). Analysis of two-point statistics of cosmic shear. i. estimators and covariances. *Astronomy and Astrophysics*, *396*, 1–19.

Sellentin, E., & Heavens, A. F. (2015). Parameter inference with estimated covariance matrices. *arXiv preprint arXiv:1511.05969*.

Sisson, S. A., Fan, Y., & Tanaka, M. M. (2007). Sequential monte carlo without likelihoods. *Proceedings of the National Academy of Sciences*, *104*(6), 1760–1765.

Toni, T., Welch, D., Strelkowa, N., Ipsen, A., & Stumpf, M. P. (2009). Approximate bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface*, *6*(31), 187–202.

Ver Steeg, G. (2014). Non-parametric entropy estimation toolbox (npeet). [Online; accessed 5-October-2015].
URL `https://github.com/gregversteeg/NPEET/blob/master/npeet_doc.pdf`

Wang, Q., Kulkarni, S. R., & Verdú, S. (2009). Divergence estimation for multidimensional densities via k-nearest-neighbor distances. *IEEE TRANSACTIONS ON INFORMATION THEORY*, *55*(5).

Weyant, A., Schafer, C., & Wood-Vasey, W. M. (2013). Likelihood-free cosmological inference with type ia supernovae: approximate bayesian computation for a complete treatment of uncertainty. *The Astrophysical Journal*, *764*(2), 116.

Wickham, H. (2009). *ggplot2: elegant graphics for data analysis*. Springer New York. URL `http://had.co.nz/ggplot2/book`

Wilkinson, R. D. (2013). Approximate bayesian computation (abc) gives exact results under the assumption of model error. *Statistical applications in genetics and molecular biology*, *12*(2), 129–141.

Wood, S. N. (2010). Statistical inference for noisy nonlinear ecological dynamic systems. *Nature*, *466*(7310), 1102–1104.