Sufficient statistics and likelihood-free inference with applications to weak gravitational lensing

Thesis Proposal

Michael Vespe

Contents

1 Introduction

Recent years have seen the introduction of likelihood-free inference to the applied statistical literature. Likelihood-free inference, a class of methods that includes approximate Bayesian computation (ABC), is appropriate in any setting of Bayesian parametric inference wherein the evaluation of a likelihood function of a parameter value may be computationally or analytically intractable, but a forward process exists to simulate data using that parameter value as its input. The end result of any ABC algorithm is a sample of parameter values whose distribution is an approximation to the desired posterior distribution.

A crucial step in any ABC algorithm involves comparing observed data to simulated data. When the dimension of the data is too high for straightforward comparison, an ABC analysis will reduce the dimension of the data via some summary statistic. Ideal summary statistics will be minimally sufficient for the parameter of interest; however, minimal sufficient statistics are rarely known in cases where the likelihood function is complex, and for some models, there may be no sufficient statistic with dimension lower than the data.

The approximate posterior sample resulting from an ABC analysis may be quite sensitive to the choice of summary statistic; this motivates the desire for summary statistics to be chosen in a principled fashion. Some recent methods focus on selecting subsets of summary statistics from among a larger pool of candidate statistics [13] or regressing parameter values on simulated data [8]. By contrast, we propose a method which, using a simulated training set, estimates mappings that embed both the parameters and data in the same lower-dimensional space. The resulting mappings from the data to the lower-dimensional space are then a natural choice for ABC summary statistics.

We present a simple toy example where the minimal sufficient statistic and the true posterior distribution are known. In this example, we demonstrate that the mapping resulting from our method, when used as the summary statistic in an ABC analysis, performs comparably to the minimal sufficient statistic, and favorably to other candidate statistics that are either not sufficient or not minimal.

We consider, as an area of application, the cosmological challenge of inferring key parameters from measurements of weak gravitational lensing. Weak lensing is the distortionary effect on the perceived shape of distant galaxies as their light passes through dark matter; inference about the structure of this distortionary effect permits the constraint of parameters in a cosmological model (e.g., the ΛCDM model). The setting is natural for application of ABC in conjunction with our method: given values for cosmological parameters, it is possible to simulate shear realizations of a form comparable to observed data, while specifying and evaluating a likelihood function for those parameter values would be impossible under realistic modeling assumptions.

2 Background and motivation

2.1 Likelihood-free inference

Likelihood-free methods allow for inference in Bayesian settings where it is difficult or impossible to evaluate a likelihood function for a given set of parameter values, but where a forward model is available to simulate data using those parameter values as inputs. This includes the approach of ABC, which was first introduced in 1999 [21] in a biological context. We give here a brief explanation of ABC, deferring to the literature (e.g., [18]) for more comprehensive exposition.

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:

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{f(x)} = \frac{f(x|\theta)\pi(\theta)}{\int f(x|\theta)\pi(\theta)d\theta} \propto f(x|\theta)\pi(\theta) \tag{1}$$

Direct evaluation of the $\pi(\theta|x)$ would require a means of evaluating both the likelihood $f(x|\theta)$ and the marginal distribution of $f(x)$. Methods in the broad Markov Chain Monte Carlo (MCMC) class 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.

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 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 of retained τ values will have 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, $S(\cdot) : \mathbb{R}^d \to \mathbb{R}^p$. This requires modifying lines 4-6 of Algorithm 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 $\pi(\theta|S(x) = S(x_{obs}))$. Note that if $S(\cdot)$ is sufficient for θ, then $\pi(\theta|S(x) = S(x_{obs}))$ is equal to $\pi(\theta|x_{obs})$; a justification is given in section A.1 of the appendix. If $S(\cdot)$ is not sufficient for θ, then the distribution of the retained τ values will be an approximation of the desired posterior.

However, depending on the distribution of $S(x)$, it may still be the case that the event that $S(y) = S(x_{obs})$ is extremely improbable. We modify the algorithm again, retaining τ if the difference between $S(y)$ and $S(x_{obs})$ is small in some metric $\rho : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$. The algorithm, modified to incorporate these changes, is given in Algorithm 2.

Algorithm 2 Basic ABC Rejection Algorithm, Modified

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

The distribution of the resulting sample of retained τ values is now $\pi_\epsilon(\theta|S(x) \approx S(x_{obs}))$, where the subscript ϵ refers to the tolerance up to which the summarized observed and simulated data may differ. If and only if $\epsilon = 0$ and $S(\cdot)$ is sufficient for θ, then $\pi_\epsilon(\theta|S(x) = S(x_{obs}))$ is equal to the desired $\pi(\theta|x_{obs})$. By contrast, as ϵ grows arbitrary 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 $\pi_\epsilon(\theta|S(x) \approx S(x_{obs}))$ 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.

Moreover, any ABC analysis will also be sensitive to the choice of summary statistic S. Asymptotic performance guarantees, such as those in [1], require of S only that it be sufficient for θ in order for $\pi_\epsilon(\theta|S(x) \approx S(x_{obs}))$ to converge to $\pi(\theta|x_{obs})$ as the size of the retained τ sample grows arbitrarily large. (Note that $S(x) = x$ satisfies trivially the condition that $S(\cdot)$ be sufficient for θ.) However, using ABC in practice requires that $S(\cdot)$ achieve some substantial dimension reduction in order to avoid the curse of dimensionality. In section 3.1, we demonstrate the effect of different choices of summary statistic on a toy example ABC analysis in which the true posterior distribution is known.

The ABC algorithm outlined here is only the most basic version; the literature contains numerous refinements and extensions of the method. One such extension is a generalization given in [4], where instead of simply accepting or rejecting τ, a weight is assigned via a smoothing kernel function $K(\cdot)$. (Indeed, taking the rectangular kernel $K_\epsilon(\rho(S(y), S(x_{obs}))) = \frac{1}{2}\mathbb{I}_{\{\rho(S(y),S(x_{obs}))\leq\epsilon\}}$ recovers Algorithm 2.) Additional extensions include a likelihood-free MCMC given in [19]; the Sequential Monte Carlo (SMC) approach of [26]; and an iterative scheme given in [2] that incorporates the importance sampling-based techniques of Population Monte Carlo. While these algorithms greatly enhance the computational feasibility of using ABC methods, the choice of an appropriate summary statistic still remains crucial.

Comparatively less attention is paid in the literature to the choice of summary statistic. [13] provides a scoring mechanism for selecting subsets of summary statistics from among a larger pool of candidate statistics, while [8] constructs estimated summary statistics via regression of parameter values on simulated data.

2.2 Weak gravitational lensing

We give here a brief overview of weak gravitational lensing in order to motivate the development of our methods, deferring to reviews (e.g., [12] and [20]) for a more complete exposition of the phenomenon and associated inference techniques.

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 [7]; the correlation 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 [17]. 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 [17] 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

$$\widehat{\xi}_\pm(\theta) = \frac{\sum\limits_{(i,j)} w_i w_j \left(\varepsilon_i^{++}\varepsilon_j^{++} \pm \varepsilon_i^{\times\times}\varepsilon_j^{\times\times}\right)}{\sum\limits_{(i,j)} w_i w_j} \tag{2}$$

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 [25], ring statistics [24], COSEBIs [23]) 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 $\boldsymbol{\theta}$ is computed via

$$\mathcal{L}(\boldsymbol{\theta}|X) = \frac{1}{\sqrt{(2\pi)^p|\mathbf{C}|}} \exp\left(-\frac{1}{2}(X - g(\boldsymbol{\theta}))^T\mathbf{C}^{-1}(X - g(\boldsymbol{\theta}))\right) \tag{3}$$

where X is the empirically computed data vector, $g(\boldsymbol{\theta})$ is the theoretical value of the data vector given parameters $\tilde{\theta}$, 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., [9].

Additionally, some analyses (e.g., [11]) 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 [9] 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 (3). Hartlap et al. [10] uses a large sample of simulations to estimate the likelihood and investigate the impact on parameter constraints of using a Gaussian approximation. Keitel and Schneider [14] 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.

Fortunately, forward simulation models are available to generate data realizations given cosmological input parameters; these simulation models range from overly simplified Gaussian random fields to, e.g., the SUNGLASS simulations of [15] 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.

3 Preliminary work

Focusing first on the ABC paradigm, we consider the issues that arise in a simple toy example in order to motivate the development of methods to estimate approximately sufficient statistics of low dimension.

3.1 Toy example: estimating a normal mean via ABC

We remark here 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, in the

ABC formulation, the posterior is approximated by

$$\pi(\tau)\mathbb{P}(y \text{ is retained} \mid y \sim M_\tau)$$

Consequently, in an ABC scheme, it would be desirable for 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 $X_1, \ldots, X_n \sim \mathcal{N}(\mu, 1)$, 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, 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, which we make formal.

3.1.1 Not summarizing enough

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 $\overline{X}$.)

At any level j, the collection $\overline{X}_1^j, \ldots \overline{X}_{2^j}^j$ is jointly sufficient for μ and thus is 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 f, so they are considered not to be random in what follows; and τ has already been drawn from the prior distribution to also be equal to 0, which is the true μ. In other words, our prior has guessed μ correctly, so to speak.

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

$$\epsilon_j^* = \sqrt{\frac{2^j}{n} F^{-1}_{2^j,\lambda}(0.25)} \tag{4}$$

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 $\lambda = \frac{n}{2^j}\sum_{i=1}^{2^j}(\overline{X}_k^j)^2$. The justification for this is given in Appendix A.2. 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 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 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 A.2, 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 distribution function 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 straightforward to evaluate this function on a grid of τ values.

In Figure 1, 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 everything 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 retain τ in a way that departs from being proportional to $\mathcal{L}(\tau; X_1, \ldots, X_{32})$; we see that this behavior gets worse as the undersummarizing grows more extreme. 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

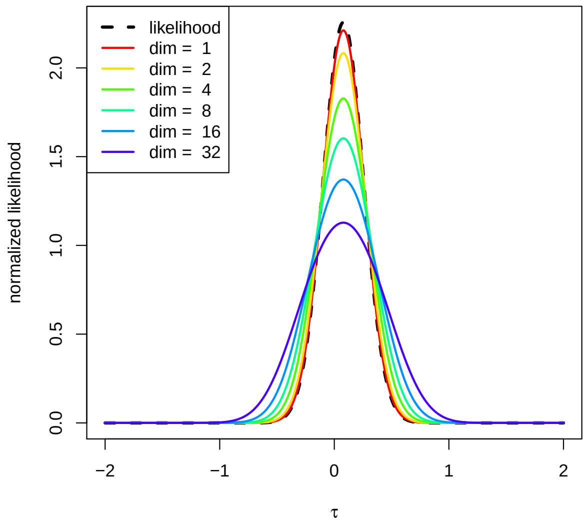

Figure 1: Normalized retention probability $p_j(\tau)$

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

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

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 use numerical optimization to compute the value $\epsilon_j^{L_2}$ that achieves that same ISE. That $\epsilon_j^{L_2}$ dictates a probability of retaining the true μ, $p(0, \epsilon_j^{L_2})$. These probabilities are given in Table 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: Tolerance values and associated retention probability at the true μ

These results tell us 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 if we used 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 are not directly comparable to each other because of the changing dimension.

3.1.2 Summarizing too much

We can also consider what happens if we summarize the data too much, in the sense that the summary statistic is no longer sufficient for the parameter; in other words, we would be discarding information.

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. Suppose again that we have already observed the data $X_1, \ldots, X_n$ so that it is 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 distribution function for the χ^2 distribution with 1 degree of freedom and noncentrality parameter $\lambda_m = m(\overline{X}_m)^2$; a justification for this is given in Appendix A.3. These ϵ_m^* values, for the same values of $X_1, \ldots X_{32}$ as were used earlier, are given in Table 3.

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

Table 3: Tolerances needed for 25% retention of the true parameter value when summarizing too much, i.e., keeping only $X_1, \ldots, X_m$. As before, these values pertain to a specific realization of the data $X_1, \ldots, X_{32}$.

Similarly to the case of not summarizing enough, 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,

$$p_m(\tau) = \mathbb{P}(\sqrt{(\overline{X}_m - \overline{Y}_m)^2} \leq \epsilon_m^*) = \mathbb{P}(m(\overline{X}_m - \overline{Y}_m)^2 \leq m{\epsilon_m^*}^2) = F_{1,\lambda_m(\tau)}(m{\epsilon_m^*}^2). \quad (5)$$

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

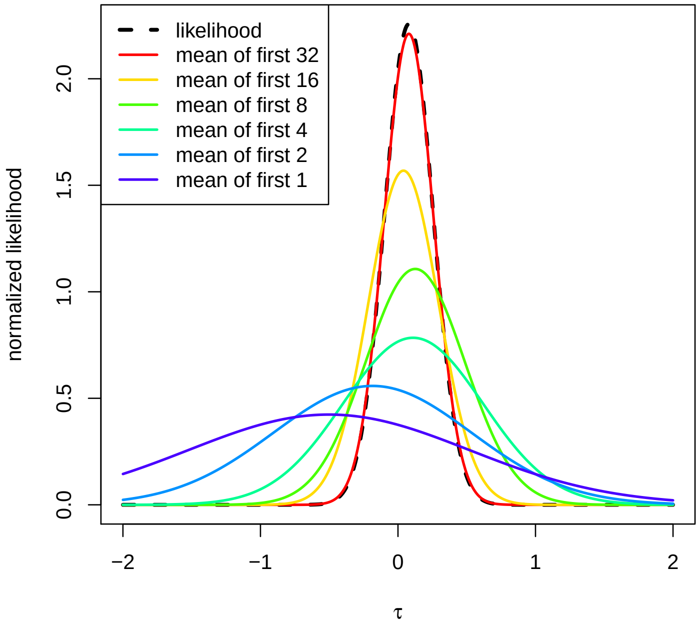

Figure 2: Normalized retention probability of τ when summarizing too much

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 in Figure 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 quality simply by taking ϵ very small.

3.1.3 Simulations

One motivation for considering this simple toy example was that the true posterior distribution $\pi(\mu|X_1, \ldots, X_n)$ can easily be derived analytically. Specifically, if we use a $\mathcal{N}(2, 1)$ prior for μ, $n = 32$, and $X_i|\mu \overset{\text{IID}}{\sim} \mathcal{N}(\mu, 1)$, then the posterior distribution $\pi(\mu|X_1, \ldots X_n)$ is $\mathcal{N}(\frac{1}{33}2 + \frac{32}{33}\overline{X}, \frac{1}{33})$.

Each of the simulations to follow will yield an ABC-derived posterior sample of retained parameter values $\tau_1, \ldots, \tau_t$. From this sample, it is straightforward to evaluate the empirical CDF for any value of μ, which is given by

$$\widehat{F}_t(\mu) = \frac{1}{t} \sum_{i=1}^{t} \mathbb{I}(\tau_i \leq \mu).$$

This empirical CDF can be compared to the known true posterior CDF $F(\mu)$ by taking the mean of the squared differences between those the two CDFs at the sample points,

$$\frac{1}{t}\sum_{i=1}^{t}\left(\widehat{F}_t(\tau_i)-F(\tau_i)\right)^2 . \tag{6}$$

In Figure 3, we present the results of ABC analyses using five different choices of summary statistic: the mean, the median, the half-means $(\overline{X}_1^1, \overline{X}_2^1)$, the quarter-means $(\overline{X}_1^2, \overline{X}_2^2, \overline{X}_3^2, \overline{X}_4^2)$, and the mean of the first half of the data $(\overline{X}_{16})$. The violin plot shows the distributions of CDF estimation errors, as computed via (6), across 500 simulations. In each of these simulations, new $X_1, \ldots, X_{32}$ data were generated, along with 10,000 candidate τ values, of which 1% were retained to form a posterior sample.

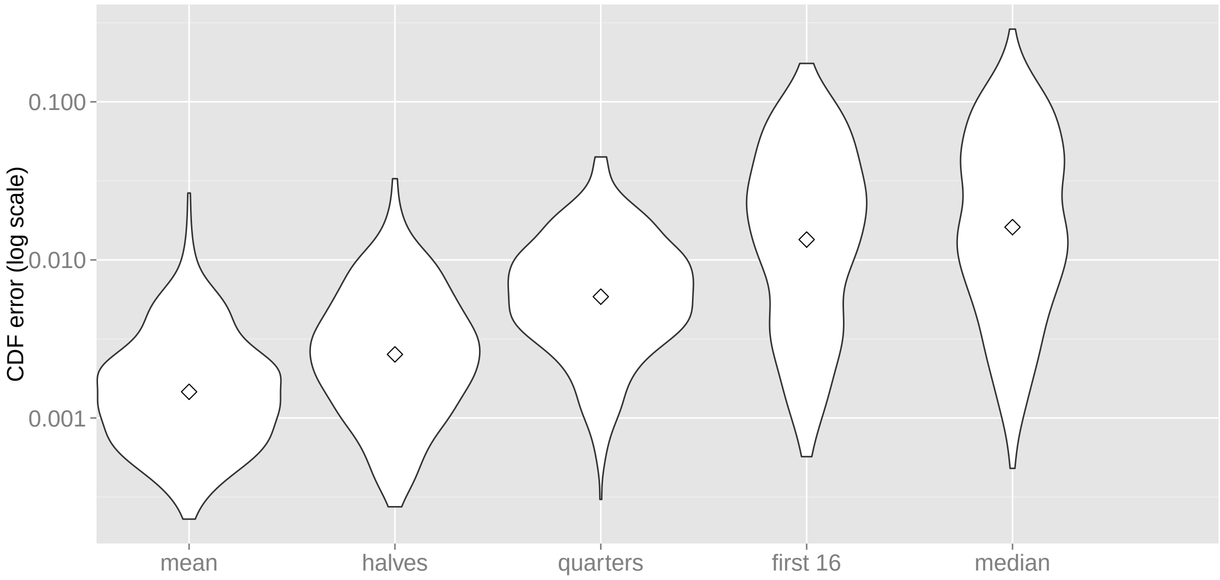

Figure 3: Distributions of error across 500 simulations of an ABC analysis using each of five summary statistics; diamonds indicate mean CDF error.

These simulation results confirm our previous understanding that the minimal sufficient statistic, in this case the mean of all of the data, results in the lowest CDF error across simulations. Moreover, summary statistics that are sufficient but not minimal (the half-means and quarter-means) typically fare slightly worse than the mean. Summary statistics that are minimal but not sufficient (the median and the mean of only half of the data) tend to perform far worse.

3.2 Common space mapping method

Motivated by the desire for summary statistics that are both sufficient and of low dimension, we introduce a training set method which seeks to find 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.

Assume $\boldsymbol{\theta}$ admits a low-dimensional representation, i.e.

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

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

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

$$T_j(X) = \eta_j(\boldsymbol{\theta}) + \epsilon_j, \tag{7}$$

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 sufficiently flexibility in the class of functions, it is hoped that an appropriate one will be found.

3.2.1 Fitting the model

First, consider that the conditional density $f(\mathbf{T}(\mathbf{X})|\boldsymbol{\eta}(\boldsymbol{\theta}))$ is given by

$$f(\mathbf{T}(\mathbf{X})|\boldsymbol{\eta}(\boldsymbol{\theta})) = \left(\frac{1}{\sqrt{2\pi}}\right)^J \exp\left(\sum_{j=1}^{J} -\frac{1}{2}\left(T_j(x) - \eta_j(\theta)\right)^2\right)$$

$$f(\mathbf{T}(\mathbf{X})|\boldsymbol{\eta}(\boldsymbol{\theta})) = \left(\frac{1}{\sqrt{2\pi}}\right)^J \exp\left(\sum_{j=1}^{J} -\frac{1}{2}\left[(T_j(x))^2 - 2T_j(x)\eta_j(\theta) + (\eta_j(\theta))^2\right]\right)$$

$$\log f(\mathbf{T}(\mathbf{X})|\boldsymbol{\eta}(\boldsymbol{\theta})) = \sum_{j=1}^{J}\left[T_j(x)\eta_j(\theta) - \frac{1}{2}(T_j(x))^2 - \frac{1}{2}(\eta_j(\theta))^2 - \frac{1}{2}\log(2\pi)\right]$$

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

$$\log f(T(x_i), \eta(\theta_i)) = \sum_{j=1}^{J}\left[T_j(x_i)\eta_j(\theta_i) - \frac{1}{2}(T_j(x_i))^2 - \frac{1}{2}(\eta_j(\theta_i))^2 - \frac{1}{2}\log(2\pi)\right] + log(\pi(\theta_i))$$

Taking the derivative of this with respect to $T_j(x_i)$ and setting equal to zero, we find that the joint log likelihood $\log f(\mathbf{T}(\mathbf{X}), \boldsymbol{\eta}(\boldsymbol{\theta}))$ is maximized when $T_j(x_i) = \eta_j(\theta_i)$. This suggests a scheme in which we try to find mappings $\widehat{T}_j(x)$ and $\widehat{\eta}_j(x)$ whose results are pulled close together.

Given a training set of N_T (θ_i, x_i) pairs, the joint log likelihood of those parameters and data would be

$$\log f(T(\boldsymbol{x}), \eta(\boldsymbol{\theta})) = \sum_{i=1}^{N_T} \left[\sum_{j=1}^{J} \left[T_j(x_i)\eta_j(\theta_i) - \frac{1}{2}(T_j(x_i))^2 - \frac{1}{2}(\eta_j(\theta_i))^2 - \frac{1}{2}\log(2\pi) \right] + log(\pi(\theta_i)) \right]$$

Our approach will be to learn mappings $\widehat{T}_j(x)$ and $\widehat{\eta}_j(\theta)$ that embed the parameters and data in the same lower-dimensional space. To this end, we appeal to a class of spectral dimension-reduction methods that find local structure in complex data.

These methods rely on a user-specified kernel 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. A common choice for k is the Gaussian kernel, $k_h(u, v) = \exp(\frac{-\|u-v\|^2}{h})$, where the bandwidth h is a tuning parameter. We form the $\mathbf{K}$ suitable for use in such a dimension-reduction method via the block matrix

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

where $\mathbf{K}_\theta$ and $\mathbf{K}_x$, are kernel matrices capturing the similarities within, respectively, the training parameter set and the training data set, and λ is a tuning parameter that represents, at least 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 provide some additional flexibility in learning a mapping.

Following the diffusion map approach of [6], a spectral decomposition of this $2N_T \times 2N_T$ $\mathbf{K}$ 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. These yield a value of $\widehat{T}_j(x_i)$ for each x_i in the training set. For a new data point $\tilde{x}$, the value of the mapping $\widehat{T}(\tilde{x})$ could be constructed by taking a weighted average of the $\widehat{T}_j(x_i)$ for the x_i in the training set. This can be accomplished by weighting by the same similarity function as was used for the $\mathbf{K}_x$ matrix, i.e.,

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

Note that $\widehat{T}$ will be sensitive to the 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.

We remark that the distribution $f(\mathbf{x}|\boldsymbol{\theta})$ is said to belong to an exponential family if

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

for some functions $A(\boldsymbol{\theta})$ and $h(\mathbf{x})$ and possibly vector-valued functions $\boldsymbol{\eta}(\boldsymbol{\theta})$ and $\mathbf{T}(\mathbf{x})$. If this assumption, which is stronger than the one made above in (7) holds, then $\mathbf{T}(\mathbf{x})$ is a sufficient statistic for $\boldsymbol{\theta}$; otherwise it will be an approximation to one. This suggests the utility of the $\widehat{T}$ resulting from the common space mapping method as an ABC summary statistic.

3.2.2 Normal mean toy example, revisited

We return to the toy example of Section 3.1, using ABC to obtain the posterior distribution of μ given n independent observations from a $\mathcal{N}(\mu, 1)$ random variable. We evaluate the performance of the summary statistic $\widehat{T}$ resulting from the CSMM in comparison to other summary statistics.

We first build a training set for the CSMM by simulating 1000 (μ_i, x_i) pairs, where each μ_i is drawn from a $Unif[-5, 5]$ distribution, and the x_i are vectors of length 32, each of whose entries are drawn independently from a $\mathcal{N}(\mu_i, 1)$ distribution.

We then construct the similarity matrices for parameters and data, quantifying distance for parameters via $d(\mu_u, \mu_v) = |\mu_u - \mu_v|$ and for data via $d(x_u, x_v) = \sqrt{\sum_{i=1}^{n}(x_{u_{(i)}} - x_{v_{(i)}})^2}$, 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.) For various values of the tuning parameter λ, the CSMM yields a mapping $\widehat{T}_\lambda$.

Our simulations have the same general structure as those in Section 3.1.3. We calculate the CDF error from (6) when using the various $\widehat{T}_\lambda$ mappings, showing their error distributions in Figure 4, with those for the mean, median, and half-means shown for reference to the right.

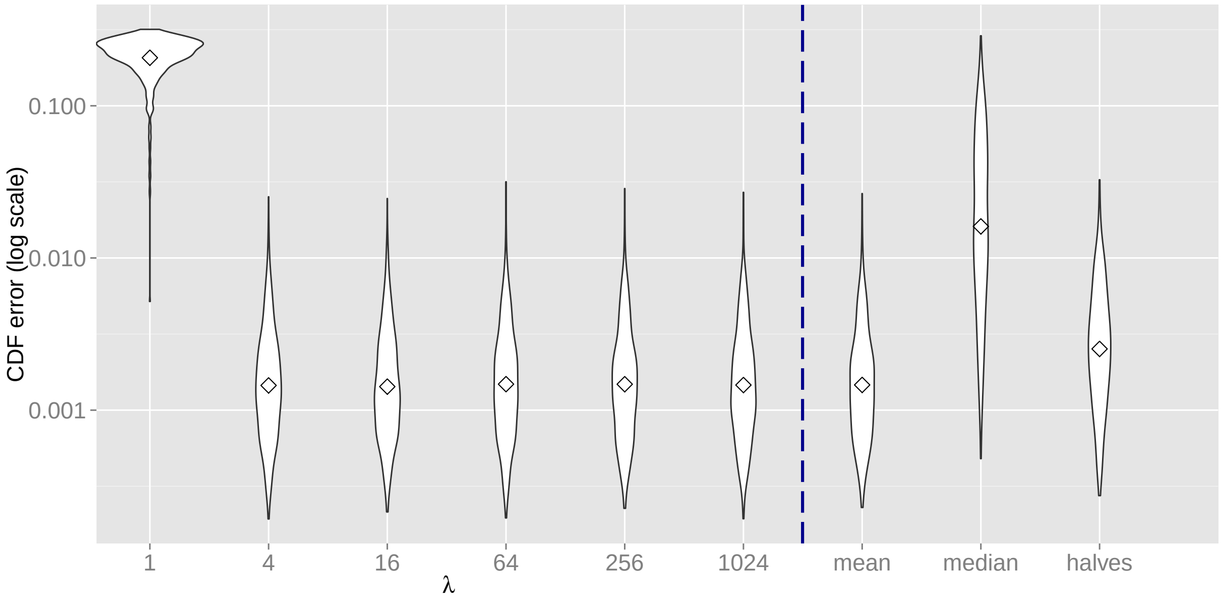

Figure 4: CDF error distributions when using the CSMM-derived $\widehat{T}_\lambda$ for various values of λ; at right, those for the mean, median, and half-means.

We observe that, in this example, the performance of the CSMM is not particularly sensitive to the value of λ, provided that it is not too small. Moreover, for adequately large

λ, $\widehat{T}_\lambda$ appears to perform comparably to the mean and noticeably better than the median or the half-means. This behavior suggests that, at least in this toy case, the CSMM is learning a mapping $\widehat{T}_\lambda$ that is in some sense approximately minimally sufficient. In this particular example, increasing λ (at least up to 1024) does not appear to have an adverse effect on the performance of $\widehat{T}_\lambda$. In section 4.1, we consider further the implications of the choice of λ.

3.3 Simple weak lensing application

We present an application of this method in a simple weak lensing analysis, simulating data from known inputs. Specifically, we generate a random realization of a shear field using input cosmological parameters $\Omega_M = 0.25, \sigma_8 = 0.8$, and all other inputs (including survey redshift distribution) chosen to replicate those in the simulation exercises of [16]. In this case, we model lensing shear as a Gaussian random field on a grid of pixels. We add i.i.d. shape noise ($\sigma_{\text{int}} = 0.37$) to represent the effect of intrinsic galaxy ellipiticity.

In applying the CSMM to learn a mapping $\widehat{T}_\lambda$, we must choose a value for the tuning parameter λ. In Figure 5, we examine the $\widehat{T}_\lambda$ mappings produced for three possible values of λ. Our heuristic interpretation, discussed further in 4.1, is that too large values of λ will result in too high priority given to the relationship between corresponding parameters and data in the training set, that is, overfitting to the training set. Thus, we aim for a value of λ for which $\widehat{T}_\lambda$ maps a (θ_i, x_i) pair close together but not quite to the same point. For this simple weak lensing example, we choose $\lambda = 5$.

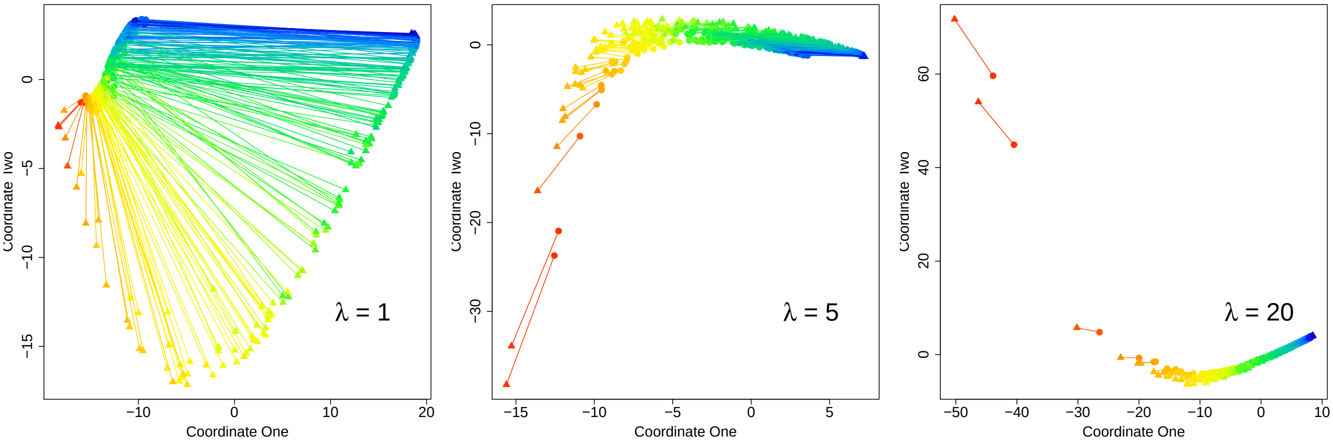

Figure 5: For three values of λ, the first two coordinates of the $\widehat{T}_\lambda$ (bullets) and $\widehat{\eta}_\lambda$ (triangles) mappings in the simple weak lensing example. Points (and lines connecting corresponding data and parameters) are colored according to the value defining the degeneracy, $\Omega_M^{0.7}\sigma_8$.

For our ABC analysis, we sample candidate parameter values $(\tilde{\Omega}_M, \tilde{\sigma}_8)$ from a uniform prior distribution on the rectangle $[0.1, 0.8] \times [0.5, 1]$. In Fig. 6 we compare samples from the approximate posterior distributions using two summary statistics: at left, estimates $\widehat{\xi}_\pm(\theta)$ of the two-point correlation functions – evaluated in eight logarithmically spaced bins – and, at right, the first two coordinates of $\widehat{T}$ learned via the common space mapping method. In each case, we take ρ to be standard Euclidean distance and we choose ϵ so that 5% (blue) and 10% (green) of the candidate samples are retained for each case.

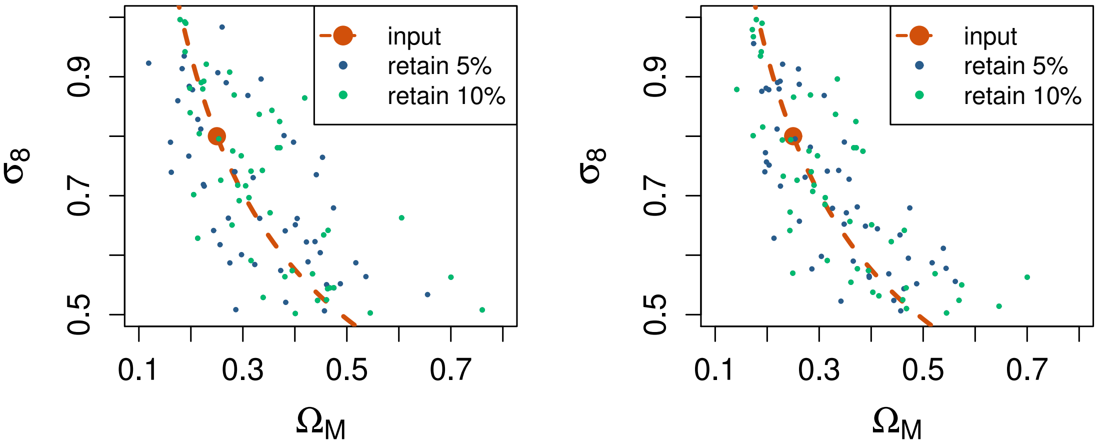

Figure 6: ABC-derived posterior samples using $\widehat{\xi}_\pm(\theta)$ (left) and $\widehat{T}_1, \widehat{T}_2$ (right).

As Ω_M and σ_8 are known to be degenerate (these data can only distinguish the value of $\Omega_M^{0.7}\sigma_8$), we display the degeneracy curve corresponding to the input parameters in orange. Simple inspection suggests that using $\widehat{T}$ as learned via the CSMM as the summary statistic is preferable to simply using $\widehat{\xi}_\pm(\theta)$, because the samples from the former assemble more tightly around the true degeneracy curve than those from the latter.

4 Research plan

4.1 Methodology

We observed in section 3.2.2 on the normal mean toy example that, above some low threshold, varying λ did not appear to affect the behavior of $\widehat{T}_\lambda$ as an ABC summary statistic. We posit that this is true here because the $\mathcal{N}(\mu, 1)$ model is identifiable; there is no redundant information contained in the parameter value μ. One of the strengths of our proposed method is that it will yield not only summary statistic $\widehat{T}(x)$ suitable for ABC but also a lower-dimensional reparameterization $\widehat{\eta}(\boldsymbol{\theta})$ that isolates the information in $\boldsymbol{\theta}$ that is relevant to the distribution of x.

We suspect that for parameterizations where there is some redundancy, choosing λ too large will, in some sense, overfit to the redundancy in the training set. The weak lensing case presents one seemingly natural domain, as $\boldsymbol{\theta}$ can be thought of either as some small collection of (e.g., two) ΛCDM parameters or as a collection of infinite-dimensional correlation functions. We plan to develop theory to make this more rigorous and demonstrate the effect of such overfitting on some examples. We hope that this work would lead to a principled scheme for choosing the tuning parameter λ in novel applications.

Beyond the choice of λ, we will work toward a better understanding of the theoretical properties of the common space mapping method. We will aim to explain more rigorously why the method seems to work in the cases where it does work, and conversely, what conditions on the structure of the parameters and the data would cause it to perform poorly. For one specific example, the structural assumption laid out in (7) would admit some very trivial $\mathbf{T}$ and $\boldsymbol{\eta}$ mappings, but in practice, the method tends to yield nontrivial $\widehat{T}$ and $\widehat{\eta}$ mappings.

4.2 Application

Our primary motivation for exploring the CSMM was the problem of weak gravitational lensing, as outlined in Section 2.2. We will refine our application of the method to this problem, attempting to incorporate more complex, and computationally costly, simulation mechanisms (e.g., [15]) into the forward modeling process. The end goal of these refinements is application to real data, both from existing and upcoming surveys, to improve upon current cosmological parameter constraint.

We will also consider other applied settings where the structure of the data and parameters does not lend itself to an obvious means of summarizing the data without discarding information. One such scenario would be the case of Gaussian mixture models, where there is no sufficient statistic of lower dimension than the number of observations. Canonically, inference in this setting proceeds via the E-M algorithm [3]. Other approaches include a Bayesian treatment given in [22] and the Population Monte Carlo (PMC) approach of [5]. Especially in the case where no sufficient lower-dimensional summary statistic is known to exist, an approximately sufficient dimension reduction would be a crucial ingredient in any ABC approach.

Finally, we will work toward a fast, efficient implementation of the method, which we plan to make publicly available as an `R` package.

References

[1] S. Barber, J. Voss, and M. Webster. The rate of convergence for approximate bayesian computation. *arXiv preprint arXiv:1311.2038*, 2013.

[2] M. A. Beaumont, J.-M. Cornuet, J.-M. Marin, and C. P. Robert. Adaptive approximate bayesian computation. *Biometrika*, page asp052, 2009.

[3] J. A. Bilmes et al. A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. *International Computer Science Institute*, 4(510):126, 1998.

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

[5] O. Cappé, A. Guillin, J.-M. Marin, and C. P. Robert. Population monte carlo. *Journal of Computational and Graphical Statistics*, 13(4), 2004.

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

[7] S. Dodelson. *Modern cosmology.* Academic press, 2003.

[8] P. Fearnhead and D. Prangle. 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, 2012.

[9] L. Fu, M. Kilbinger, T. Erben, C. Heymans, H. Hildebrandt, H. Hoekstra, T. D. Kitching, Y. Mellier, L. Miller, E. Semboloni, et al. 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, 2014.

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

[11] C. Heymans, E. Grocutt, A. Heavens, M. Kilbinger, T. D. Kitching, F. Simpson, J. Benjamin, T. Erben, H. Hildebrandt, H. Hoekstra, et al. 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, 2013.

[12] H. Hoekstra and B. Jain. Weak gravitational lensing and its cosmological applications. *Annual Review of Nuclear and Particle Science*, 58:99–123, 2008.

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

[14] D. Keitel and P. Schneider. Constrained probability distributions of correlation functions. *Astronomy & Astrophysics*, 534:A76, 2011.

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

[16] H. Lin, S. Dodelson, H.-J. Seo, M. Soares-Santos, J. Annis, J. Hao, D. Johnston, J. M. Kubo, R. R. Reis, and M. Simet. The sdss co-add: Cosmic shear measurement. *The Astrophysical Journal*, 761(1):15, 2012.

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

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

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

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

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

[22] K. Roeder and L. Wasserman. Practical bayesian density estimation using mixtures of normals. *Journal of the American Statistical Association*, 92(439):894–902, 1997.

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

[24] P. Schneider and M. Kilbinger. 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, 2007.

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

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

A Proofs of Claims

A.1 Summarizing by a sufficient statistic

Claim If $S(\cdot)$ is a sufficient statistic for θ, then $\pi(\theta|S(x)) = \pi(\theta|x)$.
Proof In what follows, let $f(\cdot)$ denote a density function.

Suppose that $S(X)$ is a sufficient statistic for θ, so that $f(x|\theta, S(x)) = f(x|S(x))$. Consider the posterior distribution $\pi(\theta|x)$.

$$\pi(\theta|x) = \frac{f(\theta, x)}{f(x)} = \frac{f(\theta, x, S(x))}{f(x, S(x))}$$

This last equality above is due to the fact that $X = x$ implies that $S(X) = S(x)$.

$$\pi(\theta|x) = \frac{f(x|\theta, S(x))f(\theta, S(x))}{f(x|S(x))f(S(x))} = \frac{f(x|S(x))f(\theta, S(x))}{f(x|S(x))f(S(x))} = \frac{f(\theta, S(x))}{f(S(x))}$$

This last quantity is simply $\pi(\theta|S(x))$, yielding the result.

A.2 Tolerance needed for summary statistic tiers in toy example

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

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

where $F_{2^j,\lambda}^{-1}$ is the inverse distribution function 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 varianceеs 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

$$\frac{n}{2^j}\sum_{k=1}^{2^j}(\overline{X}_k^j - \overline{Y}_k^j)^2 \tag{8}$$

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

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

for some chosen ϵ.

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

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

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

$S(Y_1, \ldots, Y_n)$, meaning that

$$\begin{aligned} p_j(\epsilon) &= \mathbb{P}\left(\sqrt{\sum_{k=1}^{2^j} (\overline{X}^j_k - \overline{Y}^j_k)^2} \leq \epsilon \right) \text{ or equivalently,} \\ p_j(\epsilon) &= \mathbb{P}\left(\sum_{k=1}^{2^j} (\overline{X}^j_k - \overline{Y}^j_k)^2 \leq \epsilon^2 \right), \text{ or finally,} \\ p_j(\epsilon) &= \mathbb{P}\left(\frac{n}{2^j} \sum_{k=1}^{2^j} (\overline{X}^j_k - \overline{Y}^j_k)^2 \leq \frac{n}{2^j} \epsilon^2 \right). \end{aligned} \tag{9}$$

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 $\lambda = \frac{n}{2^j} \sum_{i=1}^{2^j} (\overline{X}^j_k)^2$. 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

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

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

A.3 Tolerance needed when undersummarizing

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

Call the CDF of this last quantity F_{1,λ_m}. Then

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

Thus, the expression on the right hand side of the last equality gives the probability of retaining τ when $\tau = 0$. 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.