# Bayesian control variates for optimal covariance estimation with pairs of simulations and surrogates

Nicolas Chartier, Benjamin D. Wandelt

Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **515,** 1296–1315 (2022) https://doi.org/10.1093/mnras/stac1837
Advance Access publication 2022 July 4

# Bayesian control variates for optimal covariance estimation with pairs of simulations and surrogates

Nicolas Chartier[1,2]★ and Benjamin D. Wandelt[2,3]

[1]*Laboratoire de Physique de l'École Normale Supérieure, ENS, Universite PSL, CNRS, Sorbonne Université, Université Paris Cité, F-75005 Paris, France*
[2]*Institut d'Astrophysique de Paris, Sorbonne Université, CNRS, UMR 7095, 98 bis bd Arago, F-75014 Paris, France*
[3]*Center for Computational Astrophysics, Flatiron Institute, 162 5th Avenue, New York, NY 10010, USA*

Accepted 2022 June 28. Received 2022 June 28; in original form 2022 May 5

## ABSTRACT

Predictions of the mean and covariance matrix of summary statistics are critical for confronting cosmological theories with observations, not least for likelihood approximations and parameter inference. Accurate estimates require running costly *N*-body and hydrodynamics simulations. Approximate solvers, or *surrogates*, greatly reduce the computational cost but introduce biases, especially in the non-linear regime of structure growth. We propose 'CARPool Bayes' to solve the inference problem for both the means and covariances using a combination of simulations and surrogates. Our approach allows incorporating prior information for the mean and covariance. We derive closed-form solutions for *maximum a posteriori* covariance estimates that are efficient Bayesian shrinkage estimators, guarantee positive semidefiniteness, and can optionally leverage analytical covariance approximations. We discuss choices of the prior and propose a procedure for obtaining optimal prior hyperparameter values with a small set of test simulations. We test our method by estimating the covariances of clustering statistics of GADGET-III *N*-body simulations at redshift $z = 0.5$ using surrogates from a 100–1000× faster particle-mesh code. Taking the sample covariance from 15 000 simulations as the truth, and using an empirical Bayes prior with diagonal blocks, our estimator produces nearly identical Fisher matrix contours for ΛCDM parameters using only 15 simulations of the non-linear dark matter power spectrum. In this case, the number of simulations is so small that the sample covariance is degenerate. We show cases where even with a naïve prior our method improves the estimate. Our framework is applicable to a wide range of cosmological problems where fast surrogates are available.

**Key words:** methods: statistical – software: simulations – cosmological parameters – large-scale structure of Universe.

## 1 INTRODUCTION

To study the large-scale structure of the Universe and cosmic growth history in the era of data-driven cosmology, one needs to accurately model the statistical properties of observables in order to infer cosmological parameter constraints from surveys. The covariance matrix $\mathbf{\Sigma}$ of a summary statistics vector, such as the matter power spectrum across different wavenumbers, and most importantly its inverse – the *precision* matrix – are paramount to extracting low-dimensional summaries, building inference frameworks, or computing likelihood approximations from mock catalogues (Heavens, Jimenez & Lahav 2000; Eifler, Schneider & Hartlap 2009; Takahashi et al. 2009; Harnois-Déraps, Vafaei & Van Waerbeke 2012; Dodelson & Schneider 2013; Harnois-Déraps & Pen 2013; Blot et al. 2014; Joachimi & Taylor 2014; Percival et al. 2014; Taylor & Joachimi 2014; Alsing & Wandelt 2018; Harnois-Déraps, Giblin & Joachimi 2019; Hikage, Takahashi & Koyama 2020; Wadekar, Ivanov & Scoccimarro 2020; Giocoli et al. 2021).

The most trusted yet costly method to compute the covariance matrix of large-scale structure clustering statistics consists in generating mock realizations of survey observables with intensive *N*-body simulations – or even hydrodynamical simulations for certain applications – that mimic the conditions of observational data sampling in terms of redshift, sky area, volume, etc. We then use the samples to compute the unbiased and positive definite sample covariance estimator but getting a reliable estimate requires many realizations, especially if we need the precision matrix for the estimation of parameter confidence bounds.

To reduce the computational cost of generating simulation samples various parallel, distributed-memory *N*-body solvers have been developed sometimes with GPU-acceleration (Springel 2005, GADGET; Ishiyama, Fukushige & Makino 2009, GREEM; Warren 2013, 2HOT; Harnois-Déraps et al. 2013, CUBEP$^3$M; Garrison 2019, Abacus; Habib et al. 2016, HACC; Potter, Stadel & Teyssier 2017, PKDGRAV3; Yu, Pen & Wang 2018, Cheng et al. 2020, CUBE). Relying solely on massively parallel computing to tackle next-generation observational data sets appears impractical given our time, memory, and energy resources since thousands of simulations are needed to produce sufficiently accurate cosmological parameter constraints (see for instance, Blot et al. 2016).

For this reason, cosmologists have been searching for alternatives to running a large number of *N*-body simulations for a particular cosmological model.

On the theoretical side, analytical computations give covariance matrices that have little or no Monte Carlo noise but approximate and only valid for some assumptions on the data model. Such computations typically exploit the Gaussian limit and/or deviations

★ E-mail: nicolas.chartier412@gmail.com

from Gaussianity of the covariance (Li et al. 2019; Philcox & Eisenstein 2019; Philcox et al. 2020) or stem from perturbation theory (Mohammed & Seljak 2014; Mohammed, Seljak & Vlah 2017). For reviews of methods motivated by theoretical predictions, refer to Bernardeau et al. (2002) and Desjacques, Jeong & Schmidt (2018).

On the computational side, researchers have developed approximate solvers which are much faster than full $N$-body or hydrodynamical codes at the cost of coarser computations and simplifications that reduce the overall accuracy with respect to intensive solutions, especially at small scales. An important part of these approximate solvers use Lagrangian perturbation theory (LPT) within a low-fidelity particle-mesh (PM) framework: Tassev, Zaldarriaga & Eisenstein (2013, COLA); Tassev et al. (2015, sCOLA) implemented by Leclercq et al. (2020), Feng et al. (2016, FastPM) available in a distributed version by Modi, Lanusse & Seljak (2021a), White, Tinker & McBride (2014, QPM), and Kitaura, Yepes & Prada (2014, PATCHY), to name a few. Methods based on low-order LPT predictions provide numerous fast structure formation statistics for cosmology: Scoccimarro & Sheth (2002, PTHalos); Tassev & Zaldarriaga (2012) and Monaco et al. (2013) building upon the work of Taffoni, Monaco & Theuns (2002, PINOCCHIO), or Chuang et al. (2015, EZmocks).

An increasingly popular approach, based upon optimization, is to construct mathematical models – *emulators* – that directly predict summary statistics for specific cosmologies and parameters and of which the free parameters were previously determined through training with a specific loss function and, most importantly, simulation suites covering an appropriate range of the space of the upcoming input data (DeRose et al. 2019; Lucie-Smith, Peiris & Pontzen 2019; McClintock et al. 2019a, b; Zhai et al. 2019; Alsing et al. 2020; Angulo et al. 2021; Pedersen et al. 2021; Rogers & Peiris 2021; Kasim et al. 2022). A large proportion of the underlying mathematical models of emulators are trained neural network architectures (Alves de Oliveira et al. 2020; Lucie-Smith et al. 2020; Remy et al. 2020; Villaescusa-Navarro et al. 2021a; Spurio Mancini et al. 2022) that produce summary statistics, and some have been specifically designed to output matter density fields from input initial conditions, or even snapshots of low-resolution $N$-body simulations with particles positions and velocities (He et al. 2019; Dai & Seljak 2020; Kodi Ramanah et al. 2020). Recently, Modi et al. (2021b) proposed a solution to the inverse problem of estimating the initial density field of the early Universe: they combine a differentiable $N$-body solver with a recurrent neural network architecture (RNN) to build a tractable inference scheme. Also, Hassan et al. (2021, HIFLOW) trained an emulator and are able to produce 2D neutral Hydrogen maps conditioned on cosmology.

As a consequence of the growing enthusiasm for *Machine Learning* solvers, we have seen the production of massive simulation suites – Garrison et al. (2018), Villaescusa-Navarro et al. (2020, 2021b, 2022,) – that more and more often aim specifically at providing ways to train various emulators and models. Any trained model suffers from two main drawbacks: namely the need for many training simulations and the subsequent limitation of the model to generalize by the parameter range of the training set; and the absence of guarantee for unbiasedness of the predictions with respect to full $N$-body or hydrodynamical outputs.

All the fast solvers described above – which we will refer to collectively as *surrogates* – trade the accuracy of full $N$-body mocks, especially in the non-linear regime at small scales, for computational speed and memory gains. As a consequence, parameter constraints derived from surrogates only do not match the reliability and accuracy needed for upcoming surveys. For experiments, see the studies by Lippich et al. (2019), Blot et al. (2019), and Colavincenzo et al. (2019), where statistical biases in parameters estimation using covariance matrices from surrogates range up to $10-20$ per cent higher than with covariances computed from full $N$-body solvers.

Another approach is to attempt to reduce the number of needed simulations by modifying the statistical estimator of the covariance matrix. Numerous studies have been encouraging the use of new methods in order to deal with future surveys large data sets: *covariance tapering* in Paz & Sánchez (2015) who demonstrated the ability to reduce the confidence intervals of parameters without adding bias, fitting a theoretical model with mock samples (Pearson & Samushia 2016), jackknife resampling for the covariance (Escoffier et al. 2016; Favole et al. 2021), reducing the number of simulations by using both theoretical and simulated covariances (Hall & Taylor 2019), combining an empirical covariance with a simple target via (non-)linear shrinkage (Pope & Szapudi 2008; Joachimi 2017). As hinted at above, precision matrix estimation is the elephant in the room when it comes to undesirable effects – parameters shifts – of poor conditioning on to parameter constraints. Among the recent papers that deal with these limits and means to overcome (some of) them, the reader can refer to Taylor, Joachimi & Kitching (2013) who show how the accuracy of the precision matrix impacts parameter constraints in the case of Gaussian-distributed weak lensing power spectra, the *precision matrix expansion* method from Friedrich & Eifler (2018) and Sellentin & Heavens (2018) who show the limit of a Gaussian likelihood to derive parameter constraints, appendix B of Philcox et al. (2021) that details parameter shifts stemming from a noisy covariance estimate, Percival et al. (2022) who choose a specific covariance prior in a Bayesian framework, and also the Dark Energy Survey (DES) Year 3 results from Friedrich et al. (2021).

Variance reduction methods allow to exploit the accuracy of $N$-body solvers while dramatically lowering the number of required samples to compute robust moments estimators. Smith et al. (2021), for example, combined different lines of sight in redshift space and lowered the variance of the quadrupole estimator of the two-point clustering statistic by more than one third.

Pontzen et al. (2016), Angulo & Pontzen (2016), and Villaescusa-Navarro et al. (2018) discuss variance reduction with simulation pairs having special initial conditions. The technique allows to estimate the mean of statistics such as the power spectrum, the monopole and quadrupole of the redshift-space correlation functions or the halo mass function faster by a factor of more than 50. The induced bias, however, on certain higher order $N$-point functions renders the method not adapted to covariance estimation.

In Chartier et al. (2021) and Chartier & Wandelt (2021, CWAV20 and CW21 from now on), we developed the Convergence Acceleration by Regression and Pooling (CARPool) method, a general approach to reducing the number of simulations needed for low variance and explicitly unbiased estimates of clustering statistics moments. CWAV20 demonstrated a dramatic reduction of the number of simulations required to estimate the mean of a given statistic by exploiting the variance reduction principle known as *control variates*. The key idea is to combine a small number of costly simulations with a large number of correlated *surrogates*. Very recently, Ding et al. (2022) tested the CARPool principle to estimate the mean of the two-point and three-point clustering statistics of haloes, in order to prepare the high-resolution simulations needed for the Dark Energy spectroscopic Instrument (DESI). By pairing *AbacusSummit* suite (Maksimova et al. 2021) simulations with FastPM approximations, they found $\approx$100 times smaller variances with CARPool at scales $k \leq 0.3\,h\mathrm{Mpc}^{-1}$ than with high-resolution simulations alone.

Additionally, the extension of the method to different cosmologies (one or very few simulations of the cosmologies of interest paired with a 'primary cosmology' as the surrogate) resulted in an increase of the *effective volume* by ≈20 times. In CW21, we extended the principle to covariance estimation by applying the variance reduction approach to individual elements of a symmetric matrix, and we assessed the covariance estimates by deriving cosmological parameters confidence intervals with the Fisher matrix (using the precision). With this straightforward approach we found significant improvement in many cases, but a definite drawback was that positive-definiteness of the covariance estimate is not guaranteed. The main reason for this was because the covariance matrix was treated as a first-order statistic for the methods in CWAV20 to be directly applicable.

In this paper, to circumvent this drawback, we frame the problem as a Bayesian inference of simulation means and covariances when a (typically small) set of pairs of simulations and surrogates are available in addition to a (typically large) set of unpaired fast surrogates. We derive closed-form Maximum A Posteriori (MAP) estimators of the covariance of the simulation statistics that incorporate the information brought by the surrogates and the prior, and test the estimates by comparing the resulting confidence bounds for a ΛCDM cosmology with the true bounds. The results in this paper are very general and can apply to any summary statistics from simulations. For this reason, we motivate the study with an introductory example in Section 2 before explaining the notations and derivations in Section 3. We show several example applications to large-scale structure statistics in Section 4 and we conclude and discuss the implications of our work in Section 5.

## 2 ILLUSTRATIVE EXAMPLE

Imagine having a simulation code to compute the evolution of collisionless dark matter particles in an expanding ΛCDM universe, within a simulation volume mimicking the observational conditions of some future survey. We would like to ask: 'What amount of information the clustering statistics of the large-scale structure carry about the cosmological parameters? By which amount will we be able to constrain cosmological parameters with said statistics?' Let us say we try with the two-point correlation function in Fourier space, i.e. the (dark matter) power spectrum. For each of $n_s$ runs, with different random seeds for the initial conditions, labelled $i$, $1 \leq i \leq n_s$, the output is the vector $\boldsymbol{s}_i$ of $p_s = 158$ power spectrum bins up to $k_{\max} \approx 1.0\,h\mathrm{Mpc}^{-1}$. We will introduce the detailed notation in Section 3.1.

Under the hypothesis that the observable is sampled from a multivariate normal (MVN) distribution and that the covariance matrix does not depend on the parameters, the Fisher matrix for $d$ parameters is the symmetric matrix of size $(d, d)$

$$\mathcal{F}_{ij} = \left(\frac{\partial \boldsymbol{\mu}(\boldsymbol{\theta})}{\partial \theta_i}\right)^T \boldsymbol{\Sigma}_{yy}^{-1} \left(\frac{\partial \boldsymbol{\mu}(\boldsymbol{\theta})}{\partial \theta_j}\right) ; \tag{1}$$

hence the importance of having an accurate estimate of the covariance matrix and its inverse, the precision matrix. Then, for a parameter $\theta_i$, the Cramér–Rao inequality gives the lower bound, marginalized over the remaining parameters, for the variance of an unbiased estimator of $\theta_i$:

$$\sigma^2_{\theta_i} \geq \left[\mathcal{F}^{-1}\right]_{ii} . \tag{2}$$

To get an accurate estimate of the confidence bounds for the parameters requires an accurate estimate of the covariance matrix $\boldsymbol{\Sigma}_{ss}$. Using the standard sample covariance estimator (or maximum likelihood estimator), we would expect to need thousands of simulations costing $\mathcal{O}(10^7)$ CPU hours, much like in the *Quijote* suite (Villaescusa-Navarro et al. 2020).

But we have at our disposal a much faster surrogate solver that uses approximations from a Lagrangian fluid description of the dark matter field to produce fast but unfortunately biased approximations of this power spectrum. In this paper, we show how to leverage these fast surrogates to obtain accurate estimates of the means and covariance of the summary statistics while reducing the required number of simulations by orders of magnitude.

Fig. 1 illustrates the take-home message of this work. It shows the predicted marginal confidence regions of ΛCDM cosmological parameters[1] computed using different estimates of the power spectrum covariance. The case labelled 'Truth' uses the standard maximum likelihood estimate (MLE) of the covariance matrix from 15 000 full simulations. This 'Truth' case is hardly visible because the contours are nearly perfectly overlapped by the 'CARPool Bayes' case that uses only 15 simulations in combination with fast surrogates (noted as 10 + 5 simulations, the second term being the number of test simulations used to set a prior hyperparameter; see discussion in Section 3.5). This is one of the Bayesian covariance estimators we develop in this paper. These two cases are to be compared with the 'ML (sims only)' case showing the standard MLE of the covariance matrix from 200 simulations but without surrogates. The case labelled 'ML (surr. only)' illustrates that relying on the surrogates alone results in biased estimates of the size and orientation of the contours.[2]

Fig. 1 emphasizes the potential of the Bayesian formulation of the CARPool approach that we develop in detail in the following. Readers mostly interested in applications and numerical examples can skip to Section 4.

## 3 BAYESIAN INFERENCE OF COVARIANCE FROM SIMULATION-SURROGATE PAIRS

We wish to estimate the covariance matrix of the summary statistics $\boldsymbol{s}$, $\dim(\boldsymbol{s}) = p_s$ from accurate, expensive simulations. We also have access to a fast surrogate solver, $\boldsymbol{r}$, $\dim(\boldsymbol{r}) = p_r$, which we would not rely on alone. Inspired by CARPool, we build estimators to exploit both simulation and surrogate statistics, with the main goal of reducing the number of intensive simulations we have to run.

### 3.1 Definitions and notations

With simulation summary statistics samples $\boldsymbol{s}_i$, $i = 1, \ldots, n_s$ the standard approach to estimating the covariance matrix, defined as $\boldsymbol{\Sigma}_{ss} \equiv \mathbb{E}\left[(\boldsymbol{s} - \mathbb{E}[\boldsymbol{s}])(\boldsymbol{s} - \mathbb{E}[\boldsymbol{s}])^T\right]$ is to compute

$$\begin{aligned} \widehat{\boldsymbol{\Sigma}}_{ss} &= \frac{\gamma}{n_s} \sum_{i=1}^{n_s} (\boldsymbol{s}_i - \bar{\boldsymbol{s}})(\boldsymbol{s}_i - \bar{\boldsymbol{s}})^T \\ \bar{\boldsymbol{s}} &= \frac{1}{n_s} \sum_{i=1}^{n_s} \boldsymbol{s}_i , \end{aligned} \tag{3}$$

[1] We use $n_{\mathrm{spec}}$ as the spectral index not to induce confusion with the number of simulations $n_s$ used in the paper.

[2] We correct the bias of the precision matrix computed by inverting the standard sample covariance matrix estimator in equation (1) with the so-called 'Hartlap factor' (see Section 3.6.1 for a reminder) when using sample covariances, i.e. for 'ML (sims only)', 'ML (surr. only)' and also for the truth even if the correction is small. We do not use any correction when using the 'CARPool Bayes' estimate, a point which we discuss in Section 3.6.2.

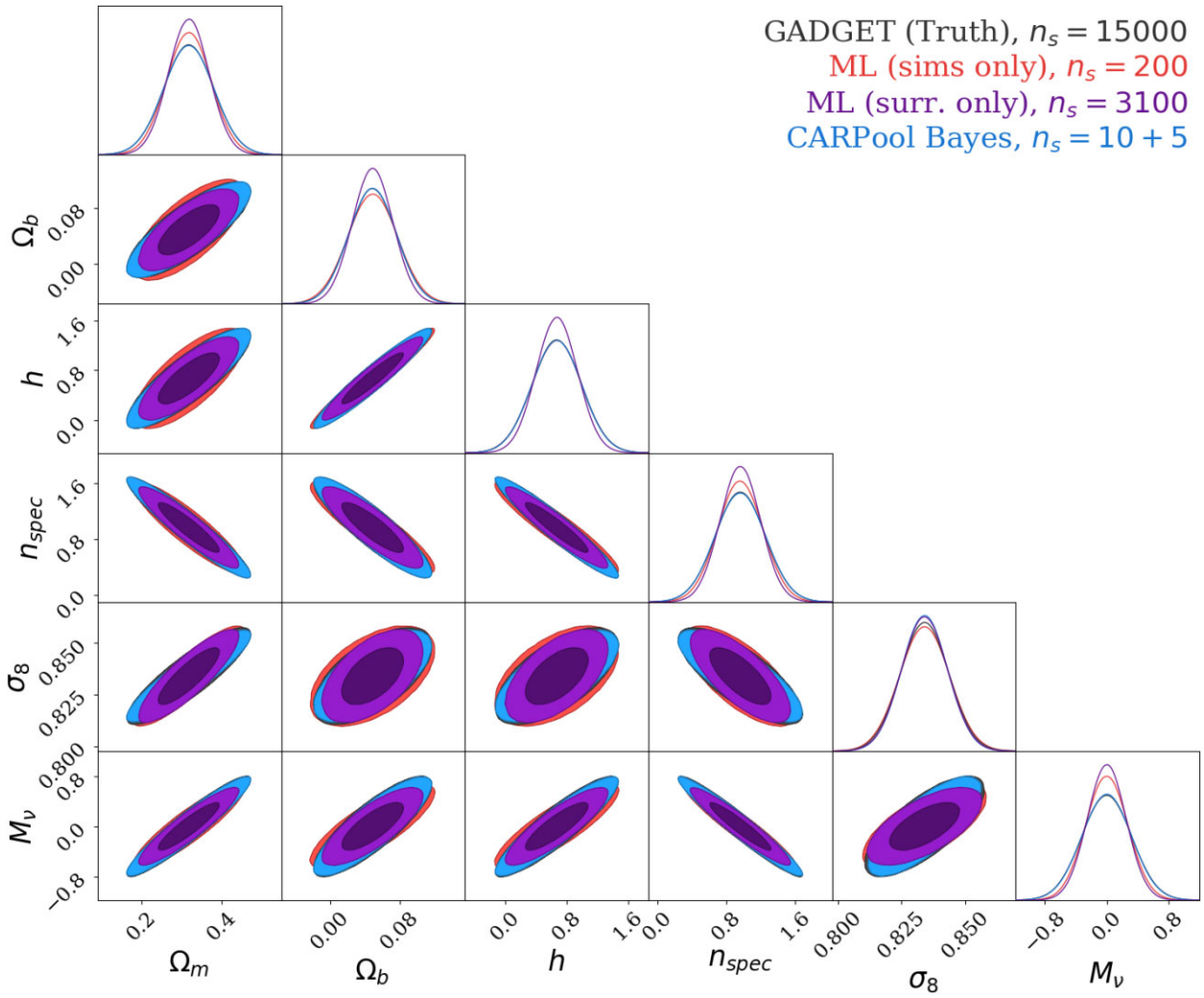


**Figure 1.** Illustrating the power of Bayesian control variates using the confidence contours of the cosmological parameters computed using the Fisher matrix based on the estimated matter power spectrum covariance matrix. The power spectrum has $p_s = 158$ bins. The 'truth' designates the confidence regions (black) from the sample covariance matrix of 15 000 $N$-body simulations, and the parameter means are set to the ΛCDM model used in the simulations. The contours are overlapped nearly perfectly by the light blue when the covariance in the Fisher matrix is computed using only 15 simulations with our CARPool Bayes MAP estimator (10 simulations and 5 for setting a prior hyperparameter, see Section 3.5). The sample covariance (ML) estimator based on many more simulations than ours gives less accurate contours. Contours based on 3100 COLA surrogates alone are rotated and too small showing that the surrogates alone are inaccurate. Detailed discussion in the text and in Section 4.

the ML estimator given an MVN likelihood function when $\gamma = 1$. To get an unbiased estimator, we use Bessel's correction factor $\gamma = n_s/(n_s - 1)$ in the ML estimator for the covariance. Equation (3) needs many samples to provide a high-quality estimate: as a matter of fact, the convergence of the smallest eigenvalues is slow (Bai & Yin 1993) and these eigenvalues will dominate the precision matrix and impact parameter parameter constraints (Taylor et al. 2013; Blot et al. 2016).

Now we add surrogates. The goal is to build a Bayesian model for the covariance of the simulations but including whatever information is provided by the surrogates. The set of surrogates $\boldsymbol{r}_j, j = 1, \ldots, n_s + n_r$ comprises $n_s$ samples that are paired with the simulations, i.e. they were computed using the same random numbers, and $n_r$ additional unpaired surrogates. We combine pairs of simulations and surrogates into a single vector

$$\boldsymbol{x} \equiv \left(\boldsymbol{s}, \boldsymbol{r}\right)^T \tag{4}$$

which implies a block matrix structure for the mean and covariance

$$\begin{aligned}\boldsymbol{\mu} &\equiv \mathbb{E}\left[\boldsymbol{x}\right] = \left(\boldsymbol{\mu}_s, \boldsymbol{\mu}_r\right)^T \\ \boldsymbol{\Sigma} &\equiv \begin{pmatrix} \boldsymbol{\Sigma}_{ss} & \boldsymbol{\Sigma}_{sr} \\ \boldsymbol{\Sigma}_{rs} & \boldsymbol{\Sigma}_{rr} \end{pmatrix}.\end{aligned}$$

Following the standard notation, we will denote the Schur complement as

$$\left(\boldsymbol{\Sigma}/\boldsymbol{\Sigma}_{rr}\right) \equiv \boldsymbol{\Sigma}_{ss} - \boldsymbol{\Sigma}_{sr}\boldsymbol{\Sigma}_{rr}^{-1}\boldsymbol{\Sigma}_{rs}. \tag{5}$$

$\mathcal{S}_p^+$ designates the space of symmetric positive-definite matrices, which is a subset of $\mathbb{R}^{p(p+1)/2}$.

For the $n_r$ *unpaired* surrogates $\boldsymbol{r}^*$, we introduce the unobserved (and in fact non-existent) corresponding simulations $\boldsymbol{s}^*$ as latent variables and then treat them as *missing data*. Again we combine

into a vector $\boldsymbol{x}^* \equiv (\boldsymbol{s}^*, \boldsymbol{r}^*)^T$ giving

$$\underbrace{s_1, \ldots, s_{n_s}}_{x_1,\ldots,x_{n_s}} \quad , \quad \underbrace{s_1^*, \ldots, s_{n_r}^*}_{x_1^*,\ldots,x_{n_r}^*}$$
$$\underbrace{r_1, \ldots, r_{n_s}}_{x_1,\ldots,x_{n_s}} \quad , \quad \underbrace{r_1^*, \ldots, r_{n_r}^*}_{x_1^*,\ldots,x_{n_r}^*}.$$

We will also distinguish the empirical counterparts of the surrogate moments according to whether they use all the $\boldsymbol{r}$ available or just the paired ones, i.e.

$\bar{\boldsymbol{r}}$, $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{rr}}$ $\longrightarrow$ estimated from the unpaired set only;
$\bar{\boldsymbol{r}}^\star$, $\widehat{\boldsymbol{\Sigma}}^\star_{\boldsymbol{rr}}$ $\longrightarrow$ estimated from both the paired and unpaired sets.

For instance,

$$\bar{\boldsymbol{r}}^\star = \frac{1}{n_s + n_r} \sum_{j=1}^{n_s+n_r} \boldsymbol{r}_j \,,$$

where we do not differentiate the paired and unpaired surrogates for simplicity ($\boldsymbol{r}_j = \boldsymbol{r}^*_{j-n_s}$ if $j \geq n_s + 1$).

We recall the well-known result that the best prediction $\widehat{\boldsymbol{s}^*}$ for any $\boldsymbol{s}^*$ given $\boldsymbol{r}^*$ with no constraints (i.e. we do not restrict the problem to the class of linear estimators) under the square loss of residuals coincides, when under an MVN distribution, with the linear regression:

$$\begin{aligned}
&\mathcal{P}(\boldsymbol{s}^*|\boldsymbol{r}^*, \boldsymbol{\Sigma}) = MVN(\boldsymbol{\mu}_{\boldsymbol{s}^*|\boldsymbol{r}^*}, \boldsymbol{\Sigma}_{\boldsymbol{s}^*|\boldsymbol{r}^*}) \\
&\widehat{\boldsymbol{s}^*} = \boldsymbol{\mu}_{\boldsymbol{s}^*|\boldsymbol{r}^*} = \boldsymbol{\Sigma}_{\boldsymbol{sr}} \boldsymbol{\Sigma}_{\boldsymbol{rr}}^{-1} (\boldsymbol{r}^* - \boldsymbol{\mu}_{\boldsymbol{r}}) + \boldsymbol{\mu}_{\boldsymbol{s}} \\
&\boldsymbol{\Sigma}_{\boldsymbol{s}^*|\boldsymbol{r}^*} = (\boldsymbol{\Sigma}/\boldsymbol{\Sigma}_{\boldsymbol{rr}})
\end{aligned} \tag{6}$$

The regression matrix of $\boldsymbol{s}$ given $\boldsymbol{r}$ will appear from now on as

$$\boldsymbol{B} \equiv \boldsymbol{\Sigma}_{\boldsymbol{sr}} \boldsymbol{\Sigma}_{\boldsymbol{rr}}^{-1}$$

For legibility, and without loss of generality, we will write all random vectors as zero-mean in the derivations such that for any sample $\boldsymbol{i}$

$$\boldsymbol{x}_i \leftarrow \boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{x}}.$$

The final equations serving as numerical recipes will include the means explicitly.

With these notations, we now turn to inferring the simulation block of the covariance $\boldsymbol{\Sigma}_{\boldsymbol{ss}}$ with the help of surrogates, given (multiple realizations of) $\boldsymbol{x}$ and $\boldsymbol{x}^*$.

### 3.2 Maximum-likelihood solution with surrogates

In a Gaussian model, the log-likelihood of $n_s$ *independent and identically distributed* (iid) samples of $\boldsymbol{x}$ and $n_r$ iid samples $\boldsymbol{x}^*$ of simulation-surrogate pairs, is

$$\begin{aligned}
-2 \ln \left[\mathcal{L}(\{\boldsymbol{x}\}, \{\boldsymbol{x}^*\}|\boldsymbol{\Sigma})\right] &= (n_s + n_r) \ln \left[\det(\boldsymbol{\Sigma})\right] \\
&+ \sum_{i=1}^{n_s} \boldsymbol{x}_i^T \boldsymbol{\Sigma}^{-1} \boldsymbol{x}_i + \sum_{i=1}^{n_r} \boldsymbol{x}_i^{*T} \boldsymbol{\Sigma}^{-1} \boldsymbol{x}_i^* + c_f \,,
\end{aligned} \tag{7}$$

where $c_f$ is the remaining constant of the likelihood for the full model including $\boldsymbol{x}$ and $\boldsymbol{x}^*$. Treating the simulations $\boldsymbol{s}^*$ in $\boldsymbol{x}^*$ as unobserved, latent variables we use the expectation maximization (EM) approach (Dempster, Laird & Rubin 1977). While EM is typically an iterative algorithm that can be slow to converge, we show in Appendix A that we can find the ML estimators of the mean and of the covariance from simulations and surrogates in closed form by computing the fixed point of the EM iterations. These are

$$\widehat{\boldsymbol{B}} = \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{sr}} \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{rr}}^{\ -1} \tag{8}$$

$$\widehat{\boldsymbol{\mu}_{\boldsymbol{s}|\boldsymbol{r}}} = \bar{\boldsymbol{s}} + \widehat{\boldsymbol{B}} \left(\bar{\boldsymbol{r}}^\star - \bar{\boldsymbol{r}}\right) \tag{9}$$

$$\begin{aligned}
\widehat{\boldsymbol{\Sigma}}^{\mathbf{ML}}_{\boldsymbol{ss}} &= \widehat{(\boldsymbol{\Sigma}/\boldsymbol{\Sigma}_{\boldsymbol{rr}})} + \widehat{\boldsymbol{B}} \widehat{\boldsymbol{\Sigma}}^\star_{\boldsymbol{rr}} \widehat{\boldsymbol{B}}^T \\
&= \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{ss}} + \widehat{\boldsymbol{B}} \left(\widehat{\boldsymbol{\Sigma}}^\star_{\boldsymbol{rr}} - \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{rr}}\right) \widehat{\boldsymbol{B}}^T \,,
\end{aligned} \tag{10}$$

where $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{ss}}$ is the sample covariance from equation (3) using simulations only. We provide a proof in Appendix A that as long as the covariance of the surrogate is positive definite the ML estimate $\widehat{\boldsymbol{\Sigma}}^{\mathbf{ML}}_{\boldsymbol{ss}}$ is guaranteed to be positive (semi)definite.[3]

As we will show in Section 4, this solution improves the estimated simulation covariance significantly with respect to the ML covariance computed from simulations only (equation 3). But the key ingredient for many applications is the precision matrix: computing optimal data combinations, least-squares estimators and optimal filtering. We will see that the dramatically underestimated smallest eigenvalues of the ML estimate of the covariance are critical.

Fortunately, the Bayesian approach allows us to include priors amounting to a form of regularization, as we will show now.

### 3.3 Inclusion of a prior information and maximum a posteriori (MAP) solutions

A convenient prior to choose for the block covariance $\boldsymbol{\Sigma}$, with $P \equiv p_s + p_r$, is the Inverse–Wishart ($\mathcal{W}^{-1}$) prior with hyperparameters $\boldsymbol{\Psi} \in \mathcal{S}_P^+$, the scale matrix, and $\nu$, the number of degrees of freedom. With $n_p \equiv \nu + P + 1$ then

$$\begin{aligned}
\mathcal{W}^{-1}(\boldsymbol{\Sigma}|\Psi, \nu) &= \frac{\det(\Psi)^{\nu/2}}{2^{\nu P/2} \Gamma_P(\frac{\nu}{2})} \det(\Sigma)^{-n_p/2} e^{-\frac{1}{2}\mathrm{tr}(\Psi \boldsymbol{\Sigma}^{-1})} \\
\boldsymbol{\Psi} &\equiv \begin{pmatrix} \boldsymbol{\Psi}_{\boldsymbol{ss}} & \boldsymbol{\Psi}_{\boldsymbol{sr}} \\ \boldsymbol{\Psi}_{\boldsymbol{rs}} & \boldsymbol{\Psi}_{\boldsymbol{rr}} \end{pmatrix} \,,
\end{aligned} \tag{11}$$

where $\Gamma_P$ is the multivariate Gamma function. $\mathcal{W}^{-1}(\boldsymbol{\Sigma}|\Psi, \nu)$ has mode $\boldsymbol{\Psi}/n_p$ for $n_p > 2P$. Its mean $\boldsymbol{\Psi}/(n_p - (2P + 2))$ exists if $n_p > 2P + 2$. In our problem, for any prior $\mathcal{P}(\boldsymbol{\Sigma})$, the mode of the posterior distribution is located at the MAP estimate

$$\widehat{\boldsymbol{\Sigma}}^{\mathbf{MAP}}_{\boldsymbol{ss}} = \underset{\boldsymbol{\Sigma}_{\boldsymbol{ss}} \in \mathcal{S}_P^+}{\mathrm{argmax}} \left[\mathcal{L}(\{\boldsymbol{x}\}, \{\boldsymbol{r}^*\}|\boldsymbol{\Sigma}) \times \mathcal{P}(\boldsymbol{\Sigma})\right] \tag{12}$$

In order to get a $\boldsymbol{\Sigma}_{\boldsymbol{ss}}$ MAP estimate, we chose to study two approaches: solving the MAP either for the whole $\boldsymbol{\Sigma}$ matrix and $\mathcal{W}^{-1}$ prior (Section 3.3.1), or for the 'regression' parameters used to infer the $\boldsymbol{\Sigma}_{\boldsymbol{ss}}$ block, which amounts to dealing with the problem solved in Anderson (1957) and reparametrizing the $\mathcal{W}^{-1}$ prior (Section 3.3.2).

#### *3.3.1 MAP with prior on the block covariance* $\boldsymbol{\Sigma}$

We take $\mathcal{P}(\boldsymbol{\Sigma}) = \mathcal{W}^{-1}(\boldsymbol{\Sigma}|\Psi, \nu)$. The derivation of the MAP estimator for the 'full' covariance, in this case, bears similarity to the well-known proof that the Inverse–Wishart distribution is a conjugate

[3] Anderson (1957) derived the same ML estimator by integrating out the $\boldsymbol{s}^*$ in equation (7) to obtain the marginal likelihood for the observed samples only

$$\begin{aligned}
-2 \ln \left[\mathcal{L}(\{\boldsymbol{x}\}, \{\boldsymbol{r}^*\}|\boldsymbol{\Sigma})\right] &= n_s \ln \left[\det(\boldsymbol{\Sigma})\right] + \sum_{i=1}^{n_s} \boldsymbol{x}_i^T \boldsymbol{\Sigma}^{-1} \boldsymbol{x}_i \\
&+ n_r \ln \left[\det(\boldsymbol{\Sigma}_{\boldsymbol{rr}})\right] + \sum_{j=1}^{n_r} \boldsymbol{r}_j^{*T} \boldsymbol{\Sigma}_{\boldsymbol{rr}}^{\ -1} \boldsymbol{r}_j^* + c_m \,,
\end{aligned}$$

with $c_m$ the remaining constant of the model with missing $\boldsymbol{s}^*$.

prior for the covariance matrix under an MVN likelihood (where $\boldsymbol{\Psi}$ becomes an additional factor of $\boldsymbol{\Sigma}^{-1}$ in the trace factorization of the log-likelihood). In the absence of additional unpaired surrogates in equation (7), the MAP estimator for $\boldsymbol{\Sigma}$ with the prior of equation (11) would match the classical result

$$\widehat{\boldsymbol{\Sigma}}^{\boldsymbol{\Delta}} = \frac{n_s \widehat{\boldsymbol{\Sigma}} + \boldsymbol{\Psi}}{n_s + n_p} \equiv \begin{pmatrix} \widehat{\boldsymbol{\Sigma}}^{\boldsymbol{\Delta}}_{ss} & \widehat{\boldsymbol{\Sigma}}^{\boldsymbol{\Delta}}_{sr} \\ \widehat{\boldsymbol{\Sigma}}^{\boldsymbol{\Delta}}_{rs} & \widehat{\boldsymbol{\Sigma}}^{\boldsymbol{\Delta}}_{rr} \end{pmatrix} \tag{13}$$

The unpaired surrogate samples, in our case, can be used in addition to the standard $\widehat{\boldsymbol{\Sigma}}^{\boldsymbol{\Delta}}_{ss}$:

$$\widehat{\boldsymbol{\Sigma}}^{\mathbf{MAP}}_{rr} = \frac{(n_s + n_r)\widehat{\boldsymbol{\Sigma}}^{\star}_{rr} + \boldsymbol{\Psi}_{rr}}{n_s + n_r + n_p} = \frac{n_r \widehat{\boldsymbol{\Sigma}}_{rr} + (n_s + n_p)\widehat{\boldsymbol{\Sigma}}^{\boldsymbol{\Delta}}_{rr}}{n_s + n_r + n_p} \tag{14}$$

$$\widehat{\boldsymbol{B}}_{\mathbf{MAP}} = \widehat{\boldsymbol{\Sigma}}^{\boldsymbol{\Delta}}_{sr} {\widehat{\boldsymbol{\Sigma}}^{\boldsymbol{\Delta}}_{rr}}^{-1} \tag{15}$$

$$\widehat{\boldsymbol{\mu}^{\mathbf{MAP}}_{s|r}} = \bar{\boldsymbol{s}} + \widehat{\boldsymbol{B}}_{\mathbf{MAP}} \left(\bar{\boldsymbol{r}}^{\star} - \bar{\boldsymbol{r}}\right) \tag{16}$$

$$\begin{aligned} \widehat{\boldsymbol{\Sigma}}^{\mathbf{MAP}}_{ss} &= \widehat{\boldsymbol{\Sigma}}^{\boldsymbol{\Delta}}_{s|r} + \widehat{\boldsymbol{B}}_{\mathbf{MAP}} \widehat{\boldsymbol{\Sigma}}^{\mathbf{MAP}}_{rr} \widehat{\boldsymbol{B}}^{T}_{\mathbf{MAP}} \\ &= \widehat{\boldsymbol{\Sigma}}^{\boldsymbol{\Delta}}_{ss} + \widehat{\boldsymbol{B}}_{\mathbf{MAP}} \left(\widehat{\boldsymbol{\Sigma}}^{\mathbf{MAP}}_{rr} - \widehat{\boldsymbol{\Sigma}}^{\boldsymbol{\Delta}}_{rr}\right) \widehat{\boldsymbol{B}}^{T}_{\mathbf{MAP}}, \end{aligned} \tag{17}$$

Notice that in the absence of a prior, equation (17) reduces to equation (10) and to the standard result $\widehat{\boldsymbol{\Sigma}}^{\boldsymbol{\Delta}}_{ss}$ with no unpaired surrogates.

Priors for the simulation and surrogate means could be trivially included as derived in appendix A in CWAV20.

Note that a simple limit of these equations exist for the case when the surrogate covariance is known exactly,

$$\widehat{\boldsymbol{\Sigma}}^{\mathbf{MAP},\boldsymbol{\Sigma}_{rr}}_{ss} = \widehat{\boldsymbol{\Sigma}}^{\boldsymbol{\Delta}}_{ss} + \widehat{B}_{\mathrm{MAP}} \left(\boldsymbol{\Sigma}_{rr} - \widehat{\boldsymbol{\Sigma}}^{\Delta}_{rr}\right) \widehat{B}^{T}_{\mathrm{MAP}}. \tag{18}$$

In Appendix A2, we show that this result can be obtained by taking the limit of equation (17) for infinite number of surrogates. In this case no unpaired surrogates need to be generated which can lead to significant savings when the computational expense for generating a large number of unpaired surrogates is not negligible compared to the simulation cost. In addition, any residual error in the estimate due to a limited number of surrogates is eliminated.

### *3.3.2 MAP with prior on the regression parameters*

A different approach is to solve the MAP for the parameters that allow to estimate $\boldsymbol{\Sigma}_{ss} = \boldsymbol{\Sigma}_{s|r} + \boldsymbol{B}\boldsymbol{\Sigma}_{rr}\boldsymbol{B}^{T}$, that is to say we use a prior for the joint distribution $\mathcal{P}(\boldsymbol{B}, \boldsymbol{\Sigma}_{s|r}, \boldsymbol{\Sigma}_{rr})$ which is a reparametrization of the $p_s(p_s + 1)/2 + p_r(p_r + 1)/2 + p_s p_r$ parameters of $\mathcal{P}(\boldsymbol{\Sigma})$. For that, we need the properties of the blocks of a covariance sampled from a $\mathcal{W}^{-1}(\boldsymbol{\Sigma}|\Psi, \nu)$ distribution. A quick outline of the derivation appears in Appendix B. With $\boldsymbol{B}_{\boldsymbol{\Psi}} \equiv \boldsymbol{\Psi}_{sr}{\boldsymbol{\Psi}_{rr}}^{-1}$ we get

$$\widehat{\boldsymbol{\Sigma}}^{\mathbf{MAP}}_{rr} = \frac{(n_s + n_r)\widehat{\boldsymbol{\Sigma}}^{\star}_{rr} + \boldsymbol{\Psi}_{rr}}{n_s + n_r + \nu - p_s + p_r + 1} \tag{19}$$

$$\begin{aligned} \widehat{\boldsymbol{B}}_{\mathbf{MAP}} = &\left[\boldsymbol{\Psi}_{sr} + \sum_{i=1}^{n_s} (\boldsymbol{s}_i - \boldsymbol{\mu}_s)(\boldsymbol{r}_i - \boldsymbol{\mu}_r)^T\right] \\ &\times \left[\boldsymbol{\Psi}_{rr} + \sum_{i=1}^{n_s} (\boldsymbol{r}_i - \boldsymbol{\mu}_r)(\boldsymbol{r}_i - \boldsymbol{\mu}_r)^T\right]^{-1} \end{aligned} \tag{20}$$

$$\widehat{\boldsymbol{\Sigma}}^{\mathbf{MAP}}_{s|r} = \frac{n_s \widehat{\boldsymbol{\Sigma}}_{s|r} + \boldsymbol{\Psi}_{s|r} + \left(\widehat{\boldsymbol{B}}_{\mathbf{MAP}} - \boldsymbol{B}_{\boldsymbol{\Psi}}\right) \boldsymbol{\Psi}_{rr} \left(\widehat{\boldsymbol{B}}_{\mathbf{MAP}} - \boldsymbol{B}_{\boldsymbol{\Psi}}\right)^{T}}{\nu + n_s + 2ps + 1} \tag{21}$$

$$\widehat{\boldsymbol{\Sigma}}^{\mathbf{MAP}}_{ss} = \widehat{\boldsymbol{\Sigma}}^{\mathbf{MAP}}_{s|r} + \widehat{\boldsymbol{B}}_{\mathbf{MAP}} \widehat{\boldsymbol{\Sigma}}^{\mathbf{MAP}}_{rr} \widehat{\boldsymbol{B}}^{T}_{\mathbf{MAP}}, \tag{22}$$

where both $\widehat{\boldsymbol{B}}_{\mathbf{MAP}}$ – rewritten explicitly as found in the derivation – and $\widehat{\boldsymbol{\mu}^{\mathbf{MAP}}_{s|r}}$ – intervening in $\widehat{\boldsymbol{\Sigma}}_{s|r} = 1/n_s \sum_{j=1}^{n_s} \left(\boldsymbol{s}_j - \widehat{\boldsymbol{\mu}^{\mathbf{MAP}}_{s|r}}\right)\left(\boldsymbol{s}_j - \widehat{\boldsymbol{\mu}^{\mathbf{MAP}}_{s|r}}\right)^{T}$ – estimators are identical to Section 3.3.1. And we have written $\boldsymbol{\Psi}_{s|r} = \boldsymbol{\Psi}_{ss} - \boldsymbol{\Psi}_{sr}{\boldsymbol{\Psi}_{rr}}^{-1}\boldsymbol{\Psi}_{rs}$. We have dropped the $P$ notation here since the reparametrization of the likelihood and prior in terms of the regression matrices instead of $\boldsymbol{\Sigma}$ makes $p_s$ and $p_r$ appear separately. As expected, the MAP estimator for $\boldsymbol{\Sigma}_{ss}$ in this approach differs from the one derived in Section 3.3.1 since the prior is not parametrization-invariant.[4]

## 3.4 Choice of the prior parameter $\boldsymbol{\Psi}$

How should we choose the form of the parameter matrix $\boldsymbol{\Psi}$? From now on, we consider that the surrogate and simulation summary statistics have the same dimension $p_s = p_r$, as this will be the case in Section 4. In the context of an Inverse–Wishart distribution, $\boldsymbol{\Psi}$ must be a $2p_s \times 2p_s$ symmetric positive-definite matrix.

Two generic choices we will present in the following with (1) blocks that are proportional to the identity matrix (the 'identity' prior) or (2) blocks that are diagonal matrices (the 'diagonal' prior). In both cases, the coefficients and covariances are estimated based on the simulation-surrogate pairs. Readers familiar with shrinkage estimators may recognize these as popular shrinkage targets (other common targets appear in table 2 from Schäfer & Strimmer 2005). We will find that $\boldsymbol{\Psi}$ appears in our estimators in an analogous way. For other particular applications, more tailored choices are of course possible. This may be the case when an approximate theoretical model for the covariances is available. As we will see in the numerical experiments in Section 4, even the choice of a 'diagonal' prior performs well and avoids the overfitting observed in the ML estimator as long as $n_p$ is chosen using the simple procedure described in Section 3.5. The 'identity' prior demonstrated improvement over the sample covariance of simulations for a much higher $n_s$ than the 'diagonal' one, thus we will only present in Section 4 computations with the 'diagonal' prior. We briefly describe the priors below.

### *3.4.1 'Identity' prior*

A common form adopted as a target for shrinkage estimates of covariance matrices is the 'identity' prior: the autocovariance of simulations and surrogates are proportional to identity matrices and the cross-covariance a diagonal matrix such that the correlation in each bin equals to $\rho\sigma_s\sigma_r$, with $\rho \in [0, 1[$.

$$\boldsymbol{\Psi}_{\mathrm{id}} \equiv \left(\begin{array}{cc|cc} \sigma_s^2 & 0 & \rho\sigma_r\sigma_s & 0 \\ & \ddots & & \ddots \\ 0 & \sigma_s^2 & 0 & \rho\sigma_r\sigma_s \\ \hline & & \sigma_r^2 & 0 \\ \boldsymbol{\Psi}_{sr}^{T} & & & \ddots \\ & & 0 & \sigma_r^2 \end{array}\right) \tag{23}$$

We require $\rho < 1$ for this matrix to be positive definite since $\det(\boldsymbol{\Psi}_{id}) = \left(\sigma_s^2\sigma_r^2(1 - \rho^2)\right)^{p_s}$. This choice of $\boldsymbol{\Psi}$ is very simple but still serves as a 'regularizer' of the estimators from Section 3.3. We adopt an empirical Bayes approach, where we estimate $\rho$ and the variances $\sigma_r^2$ and $\sigma_s^2$ directly from the simulation-surrogate pairs.

[4]We know that for two random vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ with $\boldsymbol{y} = h(\boldsymbol{x})$, if $h$ is differentiable, then for probability distributions $\mathcal{P}_y(\boldsymbol{y}) = \mathcal{P}_x(\boldsymbol{x}) \times \det(\boldsymbol{J}_{h^{-1}(y)})$ where $\boldsymbol{J}$ is the Jacobian matrix. So under a reparametrization, the two distributions have no reason to peak at the same coordinates.

The estimated variance of $y_i = s_i$ or $r_i$, $1 \leq i \leq p_s$, is $\sigma^2_{y_i} = \frac{1}{n_s-1}\sum_{j=1}^{n_s}\left(y_{i,j} - \overline{y_j}\right)^2$ and the estimated covariance between $s_i$ and $r_i$ is $\rho_i\sigma_{r_i}\sigma_{s_i} = \frac{1}{n_s-1}\sum_{j=1}^{n_s}\left(s_{i,j} - \overline{s_i}\right)\left(r_{i,j} - \overline{r_i}\right)$. In equation (23), each of the parameters $\sigma_s$, $\sigma_r$ and $\rho\sigma_r\sigma_s$ is the average of the $p_s$ corresponding quantities, indexed by $i$.

Our numerical experiments with dark matter clustering statistics strongly preferred the 'diagonal' prior we discuss next.

### 3.4.2 *'Diagonal' prior*

A natural choice to regularize the ML estimate for the covariance with simulations and surrogates is to use the estimated diagonal elements of $\mathbf{\Sigma_{ss}}$, $\mathbf{\Sigma_{sr}}$, and $\mathbf{\Sigma_{rr}}$.

$$\mathbf{\Psi}_{\rm emp} \equiv \left(\begin{array}{ccc|ccc} \sigma^2_{s_1} & & 0 & \rho_1\sigma_{r_1}\sigma_{s_1} & & 0 \\ & \ddots & & & \ddots & \\ 0 & & \sigma^2_{s_{p_s}} & 0 & & \rho_{p_s}\sigma_{r_{p_s}}\sigma_{s_{p_s}} \\ \hline & & & \sigma^2_{r_1} & & 0 \\ & \mathbf{\Psi_{sr}^T} & & & \ddots & \\ & & & 0 & & \sigma^2_{r_{p_r}} \end{array}\right) \tag{24}$$

The computation of each $\sigma^2_{s_i}$, $\sigma^2_{r_i}$ and $\rho_i\sigma_{s_i}\sigma_{r_i}$ is the same as from the 'identity' prior above.

While having a very simple structure, we can see this prior as a more adapted correction of the eigenvalues of the block matrix $\mathbf{\Sigma}$ based on the data, whereas $\mathbf{\Psi}_{\rm id}$ adds the same amount of correction on all the eigenvalues, regardless of the statistics at hand.

## 3.5 (Cross-)validation to choose the prior hyperparameter $n_p$

The hyperparameter $\nu$ (through $n_p = \nu + p_s + p_r + 1$) in equation (11) will be seen to determine the weight attributed to the prior in the closed-form solutions for $\widehat{\mathbf{\Sigma}}^{\rm MAP}_{ss}$. For different statistics, and in terms of the maximum number of simulations $n_s$ one is able to run, varying $n_p$ via $\nu$ can significantly impact the quality of the covariance, as we will discuss in Section 4.

We propose retaining a small set $\{\boldsymbol{s}_{test}\}$ of test simulations such that $n_s = n_s^{\rm cov} + n_s^{\rm test}$, where $n_s^{\rm cov}$ plays the role of the $n_s$ of the paired set in equations (13)–(22).

Consider the estimate $\widehat{\mathbf{\Sigma}}^{\rm MAP}_{ss}(n_p)$ as a function of $n_p$. This can be computed with the same $n_s$ simulations, $\mathbf{\Psi}$ prior and $n_r$ surrogates. Then an optimal $n_p$ can be computed by evaluating the MVN likelihood $\mathcal{L}\left(\{\boldsymbol{s}_{test}\}|\widehat{\mathbf{\Sigma}}^{\rm MAP}_{ss}(n_p)\right)$ which plays the role of a utility function. We find the $n_p$ that maximizes the likelihood on the test data[5]

In our tests, we allow $n_p \in [\![1, 4 * p_s + 1]\!]$, the upper bound being the smallest integer for which the Inverse–Wishart distribution is normalizable. While low $n_p$ values correspond to an improper prior we find in our numerical experiments that the likelihood rises quickly for small values of $n_p$, with corresponding improvements to the MAP covariance estimates. Then a plateau is reached, with a shallow peak or plateau and a slow decrease as $n_p$ increases. Within the shallow peak the covariance estimates are robust to the precise value of $n_p$ and we advise choosing small values once the shallow regions is reached. We interpret this preference for low values as being due to the fact that for the simple, generic priors we used (block covariance with diagonal blocks, see Section 3.4) and for the summary statistics at hand a minimum of regularization by the prior is nearly optimal when $n_s$ is small. If specifically motivated prior matrices are available larger $n_p$ could perhaps become advantageous.

We present a summary of the estimation process, for the case of the block covariance estimation of Section 3.3.1, in Algorithm 1. We obtained nearly identical results treating $n_p$ as a hyperparameter and introducing a (Jeffreys) scale prior for it before maximization.

**Algorithm 1:** MAP Estimator for $\mathbf{\Sigma}_{ss}$ given $n_p$

**Input:** A collection $\{\boldsymbol{x}_i \equiv (\boldsymbol{s}_i, \boldsymbol{r}_i)\}$, $i \in [\![1, n_s]\!]$ of paired simulation and surrogate statistics; a large number of unpaired surrogate samples $\{\boldsymbol{r}^*_j\}$, $j \in [\![1, n_r]\!]$; a small number $n_s^{test}$ of simulation statistics; a block prior $\mathbf{\Psi}$; a set $\mathcal{N}_p$ of 'prior weights' $n_p^k$, $k \leq card(\mathcal{N}_p)$.

```
/* Here we compute the 'loss' on a single
   test simulations set for
   simplification, but K-fold
   cross-validation is also an option.   */
```

**for** $n_p^k \in \mathcal{N}_p$ **do**
  Compute $\widehat{\mathbf{\Sigma}}^{\rm MAP}_{ss}(n_p^k)$ using equations (14) to (17).
  Compute the MVN likelihood $\mathcal{L}\left(\{\boldsymbol{s}\}_{test}|\mathbf{\Sigma}_{ss}(n_p)\right)$.
**end**
Determine $n_p^\star = \text{argmax}_{\mathcal{N}_p}\,\mathcal{L}\left(\{\boldsymbol{s}\}_{test}|\mathbf{\Sigma}_{ss}(n_p)\right)$
**return** $\widehat{\mathbf{\Sigma}}^{\rm MAP}_{ss}(n_p^\star)$; $\widehat{\boldsymbol{\mu}^{\rm MAP}_{s|r}}(n_p^\star)$

## 3.6 Correction factor for the precision

To compute confidence bounds of the cosmological parameters in the context of a likelihood-analysis, we need to invert the covariance matrix estimate. We briefly explain the correction used for the standard sample covariance.

### 3.6.1 *Classical result for the sample covariance*

It is well-known that taking the inverse of the bias-corrected version of the ML estimator from equation (10), i.e $\gamma\widehat{\mathbf{\Sigma}}_{ss}$ where $\gamma \equiv n_s/(ns - 1)$, results in a biased estimator of the precision matrix (Kaufman 1967; Hartlap, Simon & Schneider 2007). For data sampled from an MVN, an unbiased estimator of the precision is

$$\widehat{\boldsymbol{P}}_{ss} = \frac{n_s - p_s - 2}{n_s - 1}\widehat{\mathbf{\Sigma}}_{ss}^{\ -1} \tag{25}$$

We chose, for this study, to include what in the cosmology literature is referred to as the 'Hartlap factor' to the inverse of the bias-corrected sample covariance of simulations summary statistics (including the truth using 15 000 simulations).

### 3.6.2 *Bias from covariance estimates using simulations and surrogates*

Our MAP estimate derived in Section 3.3.1 is constructed to ensure that the result will be a symmetric positive semidefinite matrix (as well as the other MAP estimate from Section 3.3.2). As a consequence, we lose formal unbiasedness but gain dramatically improved estimates according to multiple criteria, as discussed in Section 4. If unbiasedness of the covariance estimate is important the method in CW21 can be used. Moreover, we should stress that if the 'Hartlap' factor makes the inverse of an unbiased covariance

[5]We compared this to using $K$-fold cross-validation but found no significant impact on the determination of the optimal $n_p$ comparatively to just evaluating the likelihood once without splitting the data.

estimate unbiased (for data sampled from an MVN distribution), there is no reason the same factor would make the inverses of our *biased* point MAP estimates unbiased with respect to the precision.

## 4 NUMERICAL EXPERIMENTS ON ΛCDM SIMULATION STATISTICS

### 4.1 Simulation and surrogate data

The simulation and surrogate solvers we use are identical to those CW21 and CWAV20. We recall the main points here for convenience. For more details please refer to CWAV20. The solvers evolve $\mathcal{N}_\mathrm{p} = 512^3$ cold dark matter (CDM) particles in a box volume of $(1000\ h^{-1}\mathrm{Mpc})^3$. The simulation-surrogate sample pairs take the same second-order Lagrangian perturbation theory (2LPT) initial conditions at starting redshift $z_i = 127.0$.

#### *4.1.1* N-*body solver*

We downloaded the *N*-body snapshots clustering statistics from the publicly available *Quijote* simulation suite[6] (Villaescusa-Navarro et al. 2020). The solver for all the simulations is the TreePM code GADGET-III built upon the previous version GADGET-II by Springel (2005). The force mesh grid size to solve the comoving Poisson equation at each time-step is $\mathcal{N}_\mathrm{m} = 1024$. In the following, we will use the sample covariance of all 15 000 available realizations of the fiducial cosmology as the simulation 'truth', or more precisely the best covariance estimate we have access to.

#### *4.1.2 Surrogate solver*

We generate the fast surrogate samples with The COmoving Lagrangian Acceleration (COLA) method from Tassev et al. (2013) (see also Leclercq et al. 2020), which allows generating approximate gravitational *N*-body outputs using a smaller number of time-steps than our simulation code. The principle of COLA is to add residual displacements, computed with a PM *N*-body solver, to the trajectory given by analytical LPT approximations (usually first- or second-order). Izard, Crocce & Fosalba (2016) proposed tests of the accuracy and computational cost of COLA against *N*-body simulations at different redshifts and with different time-stepping parameters. Like in CWAV20 and CW21, we used the parallel MPI implementation L-PICOLA developed by Howlett, Manera & Percival (2015), with a coarser force mesh grid size of $N_m^\mathrm{cola} = 512$.

#### *4.1.3 Post-processing of snapshots*

To extract the summary statistics from our `L-PICOLA` snapshots, we used the exact same code modules and parameters used to compute the clustering statistics available in the *Quijote* data outputs. Therefore, the simulation and surrogate summary statistics have the same dimension $p_s = p_r$. We transform the snapshots in density contrast fields with the Cloud-In-Cell (CiC) algorithm. For the matter power spectra and the correlation functions, we used the PYTHON3 module PYLIANS3,[7] For the bispectra, the results of which appearing in Appendix C3, the post-processing code is PYSPECTRUM.[8] More details can be found in CWAV20.

[6] https://quijote-simulations.readthedocs.io/en/latest/
[7] https://github.com/franciscovillaescusa/Pylians3
[8] Available at https://github.com/changhoonhahn/pySpectrum

### 4.2 The CARPool Bayes estimator and results on clustering statistics

The following tests of the Bayesian covariance estimation approach in this paper use the sample covariance matrix with all the simulations we have ($n_s^{truth} = 15\,000$) as the 'truth' to compare with other estimates. Within the main part of this paper we only present a subset of the estimators that gave the best match in terms of parameters constraints with respect to the truth.

In particular, we use the MAP estimator from Section 3.3.1 with the 'diagonal' empirical Bayes prior $\mathbf{\Psi}_\mathrm{emp}$, equation (24), estimated on the paired set of $n_s^{cov}$ simulations and surrogates. All our MAP covariance estimates with simulations and surrogates use the optimal $n_p$ determined through the process described in Algorithm 1 with a small number of test simulations. We display the total number of simulations used for each covariance matrix estimate as $n_s = n_s^\mathrm{cov} + n_s^\mathrm{test}$.

In the following, we will refer to this approach as the 'CARPool Bayes' estimator.

We find that the alternative estimator written in terms of the regression parameters, Section 3.3.2, performs comparatively poorer than the CARPool Bayes estimator. We show an example on the power spectrum covariance and discuss the reasons for this in Appendix C2. Briefly summarized, this estimator requires using a proper prior and therefore affords us less flexibility in choosing the weight of the prior. It therefore tends to give covariance estimates that are more sensitive to the choice of the prior parameters.

The plan for the remainder of this sections is as follows: we will first present the power spectrum results in more details that were already partially described in Section 2.

Then, we turn to the real space two-point correlation function. This is an interesting case because it illustrates the power of limiting the range of the estimator to the set of all positive definite covariance matrices, a feature of the Bayesian version of CARPool. The unbiased CARPool approach to the covariance matrix in CW21 failed to yield positive-definite covariance estimates for this application in spite of a significant reduction of variance for the covariance matrix individual elements.

For a complete comparison with CW21, we also computed results on the bispectrum covariance matrix. Since these show similar, large improvement over the CW21 approach as for the power spectrum, we relegate details to Appendix C3.

#### *4.2.1 Matter power spectrum covariance*

The matter power spectrum [$\mathrm{Mpc}^3$], at wavenumber $k[h\mathrm{Mpc}^{-1}]$, under the conditions of homogeneity and isotropy (*cosmological principle*), is the average in 3D Fourier space of $|\delta(k)|^2, k \in [k - \Delta k/2, k + \Delta k/2]$, where $\delta(\boldsymbol{x})$ is the matter density contrast in real space. For each snapshot, we compute $\delta(\boldsymbol{x})$ on a square grid of size 1024 with the CIC algorithm. The following analysis is for $k \in [8.900 \times 10^{-3}, 1.0]h\mathrm{Mpc}^{-1}$. We have then $p_s = 158$ linearly space bins. Note that the power spectrum is not compressed unlike in CWAV20 and CW21, making the covariance estimation tasks more difficult.

Fig. 1 shows that using only $n_s = 10 + 5$ simulations with paired surrogates and an additional set of surrogate samples, we get confidence bounds for the cosmological parameters which are very close to the ones given by the 'true' sample covariance using 15,000 simulations (for $p_s = 158$). This result is all the more encouraging that with only 10 simulations we would get a sample covariance of rank at most 9. In other words, we can see the small set of simulations in the 'CARPool Bayes' estimate a correction to the

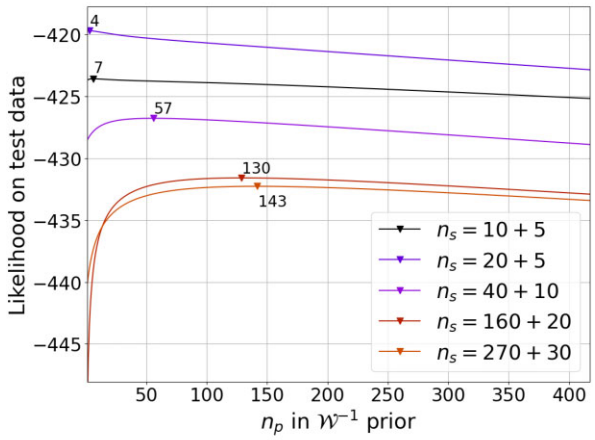


**Figure 2.** Illustration of step 5 of Algorithm 1 for the matter power spectrum example: for fixed $n_s = n_s^{\rm cov} + n_s^{\rm test}$ and fixed prior $\mathbf{\Psi}_{\rm emp}$, we compute the likelihood of $\widehat{\mathbf{\Sigma}}_{ss}^{\mathbf{MAP}}(n_p)$ on test simulation samples.

eigenvalues and eigenvectors of the precision matrix computed from a biased but correlated surrogate. We also show in Appendix C1 the relatively small gain, in terms of closeness of the parameters confidence contours to the truth, of running $n_s = 40 + 10$ simulations for comparison.

Here, we examine the procedure to determine the best $n_p$ for a given $n_s$ and $\mathbf{\Psi}$ in Fig. 2. There are several points to notice here:

(i) For $n_p \approx 1$, especially for $n_s \geq p_s + 1$ (when the sample covariance can be full-rank), the likelihood on test data rapidly increases. It shows for this case that a minimum of 'regularization' brought by the prior greatly improves the estimate of $\mathbf{\Sigma}_{ss}$.

(ii) Around the empirical $n_p^\star$, the likelihood is rather flat and slowly decreases when $n_p > n_p^\star$. In other words, once a certain threshold of 'improvement' is reached with $n_p$, misestimating $n_p$ does not radically worsen the estimate of $\mathbf{\Sigma}_{ss}$.

In Fig. 3, we visualize the estimated covariance matrices (top row) and their inverse (bottom row). For the 'CARPool Bayes' estimate with the prior $\mathbf{\Psi}_{\rm emp}$, i.e. our 'headline' estimate with $n_s = 10 + 5$ that gives the confidence bounds in Fig. 1, we notice some structure in the covariance due to the small number of simulations. The closeness to the truth of the 'CARPool Bayes' covariance with very few simulations is particularly visible for the structure of the precision matrix. It can be seen that at low $k$, where the correlation between surrogates and simulations is particularly high, the CARPool Bayes estimate (and the ML estimate without the prior) is significantly less noisy than the standard estimator even though it uses an order of magnitude less simulations.

In Fig. 4, we compare the covariance estimates to the large-sample 'truth' in the spectral domain, showing the eigenvalues and the co-diagonalization coefficients.[9]

At the top, we show the ordered eigenvalues ratio of each matrix. A ratio far from 1, and especially close to zero for the smallest eigenvalues as for the standard sample covariance, indicates a very poor conditioning of the matrix. At the bottom, we see the co-diagonalization coefficients. A horizontal line at 1 would indicate that the matrices are identical. The CARPool Bayes estimate clearly outperforms the other estimates and avoids the characteristic under-estimation of small eigenvalues for covariance matrices estimated from a small number of samples. In addition, we clearly see that the 'ML w/ surrogates' estimator (from Section 3.2) only improves the underestimated small eigenvalues of the sample covariance by a slight amount. We also found that the parameter contours obtained with the 'ML w/ surrogate' covariance are much thinner than the truth, much like with the sample covariance of simulations in CW21 where we did not improve the Fisher matrix using the 'Hartlap' factor for the precision matrix. We do not show the 'ML w/ surrogates' contours in this paper not to overload the figures.

[9]For $\boldsymbol{A}$ and $\boldsymbol{B}$ two $p \times p$ real symmetric matrices, if $\boldsymbol{A}$ is positive definite, then there exists a matrix $\boldsymbol{M}$ such that $\boldsymbol{M}^T\boldsymbol{A}\boldsymbol{M} = \boldsymbol{I}_p$ and $\boldsymbol{M}^T\boldsymbol{B}\boldsymbol{M} = \boldsymbol{D}$ with $\boldsymbol{D} = \mathrm{diag}(d_1, \ldots, d_p)$. We call the $d_i$ 'co-diagonalization coefficients.' This is a simplified statement from theorem A9.9 in Muirhead (1982). If $D = \mathbb{1}_p$ then $A = B$.

### 4.2.2 *Matter correlation function covariance*

The example of the two-point matter correlation function $\xi(\boldsymbol{r})$ for $\boldsymbol{r} \in [5.0, 160.0]\,h^{-1}$Mpc ($p_s = 159$) is of particular interest in our study. With the variance reduction approach in CW21 for the covariance matrix, we found no improvement over the standard estimator. While unbiased and strongly reducing the errors of all individual elements of the covariance matrix the resulting matrix failed to be positive-definite. This means that no estimate for the precision matrix could be obtained, as would be required to derive Fisher matrices or for a likelihood approximation to derive parameter constraints.

As we can observe in Fig. 5, the structure of the covariance is particular, with a band of high-magnitude covariances around the diagonal of variances. As a result, the precision estimate based on the standard sample covariance estimator is very noisy for $n_s = 200$, which we compare with our estimate including surrogates from Algorithm 1, with $n_s = 160 + 20$. Looking at the precision matrices (bottom row) would indicate a significantly better recovery of the structure of the true precision.

In terms of cosmological parameter forecast constraints, as shown in Fig. 6, we get a slight improvement with respect to the sample covariance matrix (and the precision including the Hartlap factor), but not nearly as large as for the matter power spectrum. The CARPool Bayes estimate with $n_s = 160 + 20$ produces bounds for $\Omega_{\rm m}$, $n_s$ and $\sigma_8$ that are closer to the truth than with the sample covariance with $n_s = 180$. But the confidence regions for $\Omega_b$ and $h$ are not improved.

Similarly to the previous section, in Fig. 7, the 'CARPool Bayes' estimator raises up the smallest eigenvalues – as well as the smallest 'co-diagonalization' coefficient – contrarily to the ML solutions with and without surrogates where they are close to 0.

The wide band of correlations visible in Fig. 5 indicate that our choice of 'diagonal' prior is not optimal for this case. Choosing a prior with a more gradual falloff of correlation from the diagonal would likely produce better results (and we would likely see that the smallest eigenvalues are not as overestimated as in Fig. 7). Fig. 8 indicates that for various number of simulations $n_s$, the CARPool Bayes estimator for the matter correlation function covariance consistently prefers low $n_p$ (i.e. prior weight) values with the 'diagonal' prior from Section 3.4.2.

In summary, the application to the matter correlation function demonstrates that the CARPool Bayes estimator is guaranteed to produce positive definite matrices. It visibly improves the structure of the precision matrix (Fig. 5) and the relative errors of the small eigenvalues (Fig. 7). This translates into some, but not all, parameter confidence bounds being closer to the truth than for the sample covariance based on 180 simulations.

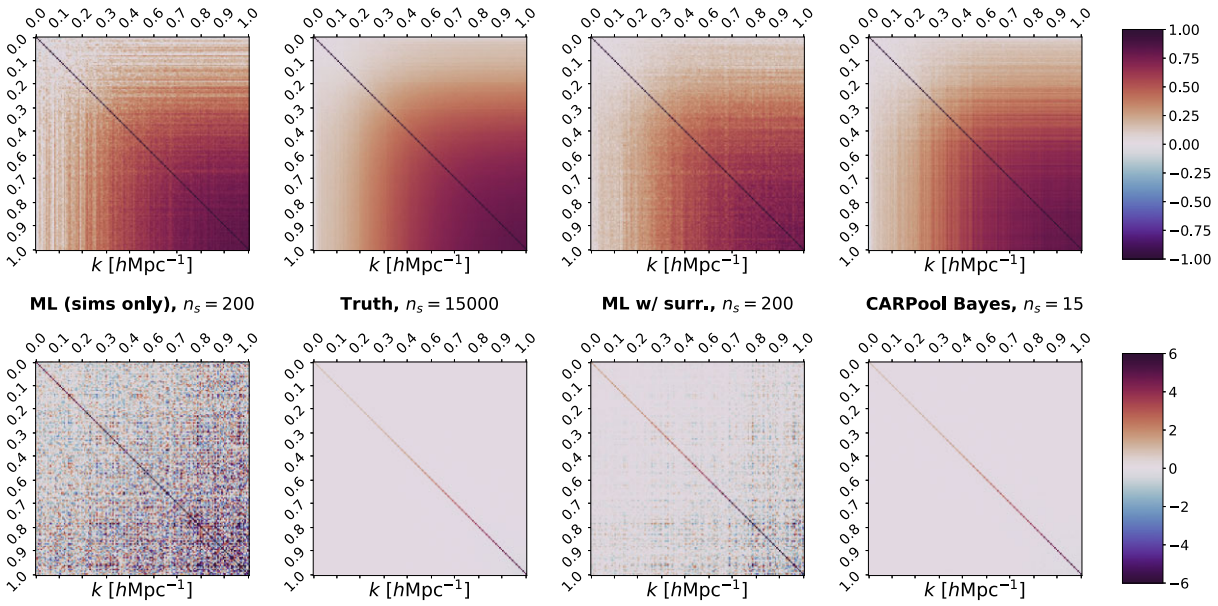


**Figure 3.** Matter power spectrum covariance estimates (top row) and their inverse (bottom row). We show the covariances as correlation matrices with the normalization $\boldsymbol{D}^{-1}\widehat{\boldsymbol{\Sigma}}\boldsymbol{D}^{-1}$ with the diagonal $\boldsymbol{D} = \sqrt{\text{diag}\left(\widehat{\boldsymbol{\Sigma}}\right)}$, and the precision matrices below are the inverses of these correlation matrices. Columns from left to right show the standard sample covariance estimate from 200 simulations; the reference covariance from 15 000 simulations; the ML estimate using the combination of surrogates and 200 simulations (Section 3.2); and the 'CARPool Bayes' estimate with a 'diagonal' prior, combining $n_s = 10 + 5$ GADGET simulations with surrogates.

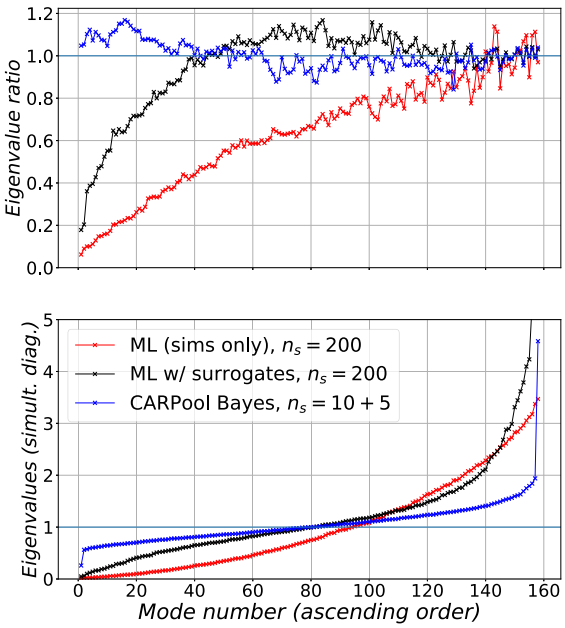


**Figure 4.** Comparison of the CARPool Bayes covariance estimate (Section 3.3.1), the standard ML estimator, and the ML estimator combining simulations and surrogates (Section 3.2) with the large-sample 'truth' in the spectral domain. We show ordered eigenvalue ratio at the top and co-diagonalization coefficients at the bottom. In the lower panel, a perfect method (with respect to the 'true' covariance) would have a value of 1 for every mode. The CARPool Bayes estimator avoids the characteristic underestimation of small eigenvalues for covariance matrices estimated from a small number of samples. See discussion in the text.

## 5 DISCUSSION AND CONCLUSION

### 5.1 A flexible Bayesian framework for mean and covariance estimation

We consider the problem of estimating the covariance matrices of cosmological summary statistics within a Bayesian framework, when paired simulations and surrogates are available.

This study constitutes an extension of the CARPool principle, presented in CWAV20 and applied to covariance matrices in CW21. Our method improves on the latter work by solving a MAP optimization directly in the space of symmetric positive semidefinite matrices and allows introducing priors in analogy to frequentist shrinkage estimators. We prove that our approach, dubbed CARPool Bayes, guarantees positive definite estimates, for the price of abandoning the guarantee of unbiasedness of individual covariance matrix elements provided by the first-order estimator described in CW21. The guarantee of positive (semi)definiteness implies bias, since otherwise a matrix that has one or more zero eigenvalues should give a negative definite estimator some fraction of the time. We note that the literature strongly tends to propose efficient covariance estimators that are biased.

By casting CWAV20 in a Bayesian framework, we provide a new solution to covariance estimation with simulations and surrogates. We demonstrate that this estimator can strongly improve over previous approaches such as the sample covariance or the first-order CARPool approach in CW21 according to multiple criteria. These improvements are particularly noticeable for the inverse covariance or precision matrix required for many applications such as computing Fisher matrices, or for the Gaussian likelihood approximations frequently used for parameter estimation.

Our Bayesian approach combines estimations for the both the mean (through the well-known regression $\boldsymbol{\mu}_{s|r}$, equation 16) and

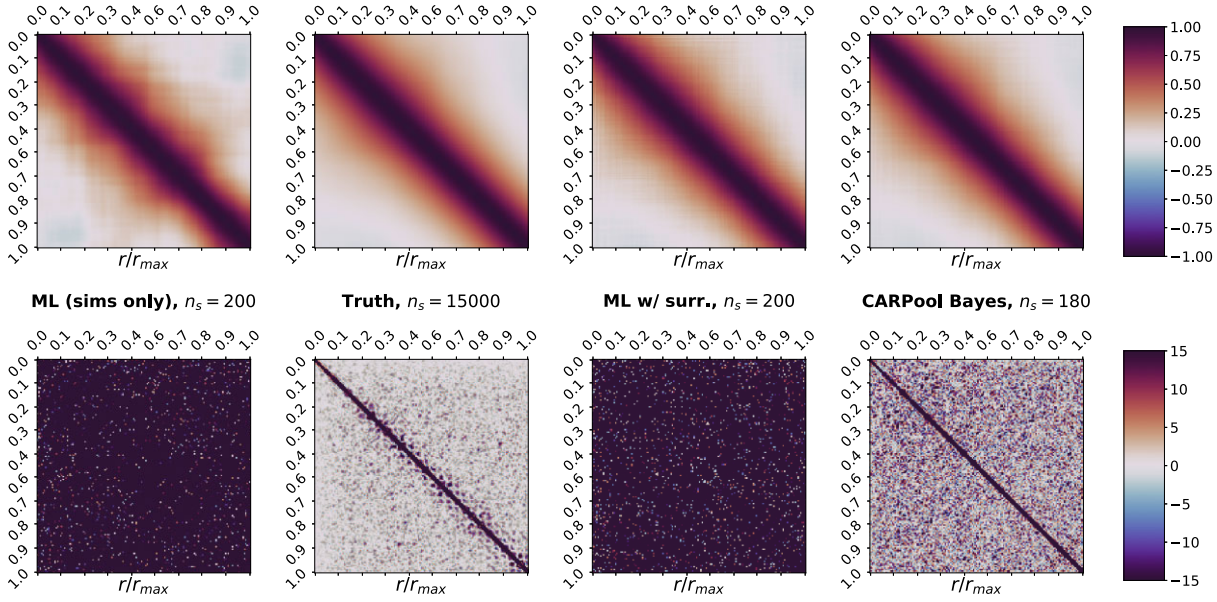


**Figure 5.** Matter correlation function covariance estimates (top) and their inverse (bottom), shown similarly to Fig. 3. The 'CARPool Bayes' estimates uses $n_s = 160 + 20$ GADGET simulations.

the covariance of simulation summary statistics using surrogates. In this paper, we focused on showing the results for the simulation covariance estimates $\mathbf{\Sigma}_{ss}$ since this is the first time that the control variate approach has been cast in a Bayesian framework for covariance estimation.

Our Bayesian approach used a multivariate Gaussian model for the simulations and surrogates and includes a conjugate Inverse–Wishart distribution prior for the covariance matrix. In the generic case we found a 'diagonal' prior on the block covariance of simulation and surrogate summary statistics, whose diagonal elements were evaluated on simulation-surrogate pairs, Section 3.4.2, to give excellent results, especially for the matter power spectrum and bispectrum. We obtain the same confidence bounds as with the true covariance of the power spectrum with $p_s = 158$ bins up to $k \approx 1.0\,h\mathrm{Mpc}^{-1}$ with only $n_s = 10 + 5$ simulations. In this case, we can think of the actual 10 simulations of the covariance estimate as correcting the eigenspectrum of the well-converged covariance of the correlated surrogate that incorporates many samples.

The same outstanding gain appears for the bispectrum, as we show in Appendix C3 for two triangle configurations. This demonstrates the superiority of the CARPool Bayes approach over CW21.

Regarding the two-point matter correlation function in real space, we do get positive-definite estimates by construction – this is not guaranteed in CW21 – and we obtain a slight improvement on the parameter constraints with respect to the sample covariance of simulations when $n_s \gtrsim p_s + 1$ is close to the dimension of the summary statistics. But in a case where running a high enough number (we tested $n_s = 300$) of simulations is possible, the gain over the sample covariance diminishes as $n_s$ increases, at least regarding the impact of the matter correlation function covariance on the parameter constraints.

Throughout this study, we applied the 'diagonal' empirical Bayes prior through taking the diagonal of each block of the summary statistics as a concatenation of the simulation and surrogate output. Using the former was sufficient to demonstrate the capability of the method for the case where we consider the problem of estimating the whole block covariance $\mathbf{\Sigma}$ to then extract the simulation block $\mathbf{\Sigma}_{ss}$. In summary, we claim that our method should always be be used over the standard unbiased sample covariance in cases where you have access to surrogates, since it guarantees positive-definiteness and allows computing robust, well-conditioned covariances of simulation summary statistics, even with 'naïve' prior structures which lead to Bayesian shrinkage in the MAP closed-form solutions. Having a small test set serves to set the prior hyperparameters, as we did in Algorithm 1 which is the main recipe in this study.

Regarding the inclusion of priors (but the use of surrogates), the motivations of our approach are similar to Hall & Taylor (2019), where theoretical covariance matrices improve the estimator based on simulations. This reference proposes a closed-form solution for a covariance estimate that appears as a particular case of linear shrinkage estimator, but extracted from a marginal likelihood derived in a Bayesian framework. The authors succeeded in computing a robust estimate with a much lower number of simulations than the sample covariance estimate. In addition, the authors investigate the influence of the chosen $\mathcal{W}^{-1}$ distribution hyperparameter on the resulting covariance in a similar spirit to our study. In our method, we can use improper priors with low values of $n_p$ since we do not marginalize the posterior distribution.

We derived new MAP estimators in Section 3.3.2 where we directly estimate the regression parameters allowing to compute the simulation block of the covariance using $\mathbf{\Sigma}_{ss} \equiv (\mathbf{\Sigma}/\mathbf{\Sigma}_{rr}) + \mathbf{\Sigma}_{sr}\mathbf{\Sigma}_{rr}^{-1}\mathbf{\Sigma}_{rs}$ from the hypothesis of data sampled from a MVN distribution. This new estimator did not provide improvement in our tests over the sample covariance, which we attribute to the strong prior dependence inherent in it. We leave for future studies the question whether the this different parametrization can turn beneficial for cosmological survey forecasts when theoretically motivated covariances for the prior are available.

### 5.2 Generating samples from the posterior

As an alternative to focusing on closed-form point estimates of the covariance by taking the MAP of the posterior distribution we could

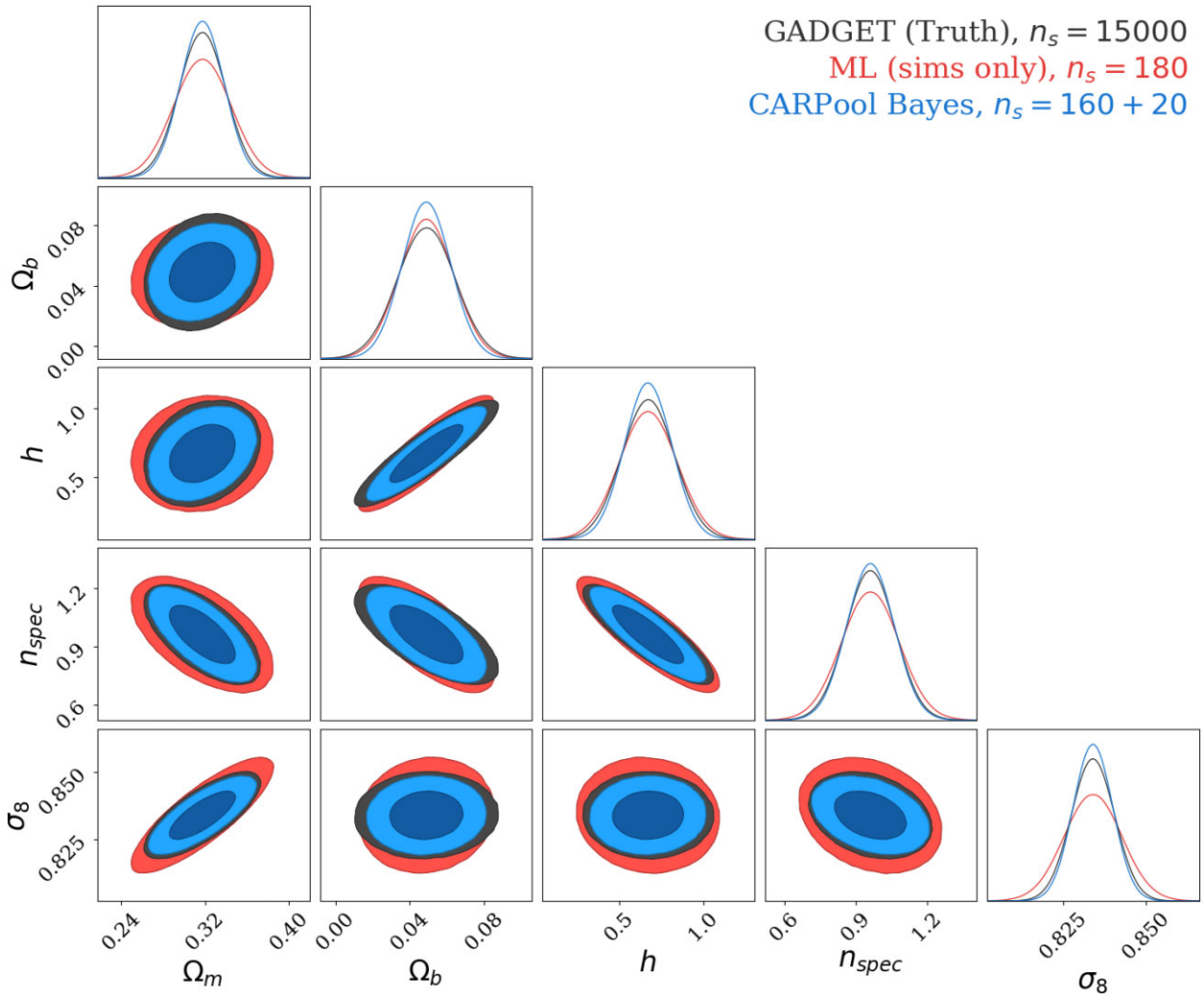


**Figure 6.** Confidence contours of the cosmological parameters computed using the Fisher matrix based on the estimated matter correlation function covariance matrix. The estimators which we compare are the same as in Fig. 1.

have considered generating samples of the simulation covariance matrix from the posterior. This is possible using a Gibbs sampling approach where we explicitly include the missing simulations $\boldsymbol{s}^*$ as latent variables. We briefly sketch the approach here: first draw $\boldsymbol{\Sigma}$ from a conditional Inverse–Wishart for positive (semi)definite covariance matrices given the data augmented by the latest $\boldsymbol{s}^*$ sample. Since the augmented data are a complete set of simulation-surrogate pairs, the $\boldsymbol{\Sigma}$ sample would therefore be guaranteed to be positive (semi)definite. Simply extracting the simulation autocorrelation block from $\boldsymbol{\Sigma}$ would produce samples from the marginal posterior for $\boldsymbol{\Sigma}_{ss}$.

While samples from the marginal posterior would potentially be useful to propagate the uncertainty in the estimates, or to study other posterior summaries such as the posterior mean, we do not explore this approach further, for two reasons: one is computational cost though that is perhaps tolerable for summaries with moderate dimension (i.e. up to (100)); the other is that we would like to obtain a point estimate for the covariance that we can use in other contexts, without worrying if the Monte Carlo estimate, e.g. of the posterior mean of the signal covariance, has converged to sufficient accuracy.

### 5.3 Potential for future applications in cosmology and beyond

Our numerical experiments demonstrate the capability of running fewer intensive simulations in order to get theoretical predictions of the means and covariances of observables for next-generation surveys. Many additional applications of these techniques remain to be explored. The free choice of what to use as surrogates makes our methods very broadly applicable.

Some surrogates might be useful because they are nearly free computationally. A case in point would be Eulerian linear perturbation theory for the power spectrum applied to the initial conditions of an $N$-body simulation. In this case each simulation comes with the paired surrogate for free (since the initial conditions are necessary to run the simulation in the first place) and its expectation and covariance can be computed analytically nearly for free. It could be argued that such automatic surrogates ought to be exploited systematically when predicting commonly used clustering statistics from simulations. A very similar application of this idea to a non-perturbative statistic would be to the computation of halo number functions: apply the Press–Schechter approach to the initial conditions as a surrogate for the mass function for a given realization. In this example, the classical Press–Schechter formula provides the expectation of the surrogate

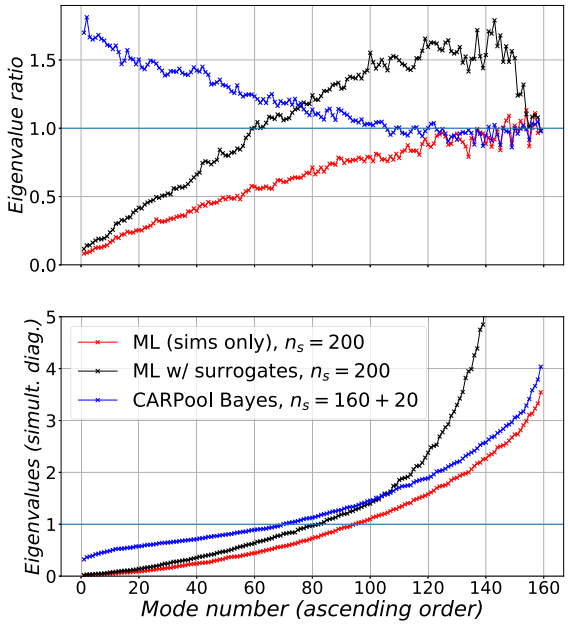


**Figure 7.** Same as Fig. 4 for the matter correlation function.

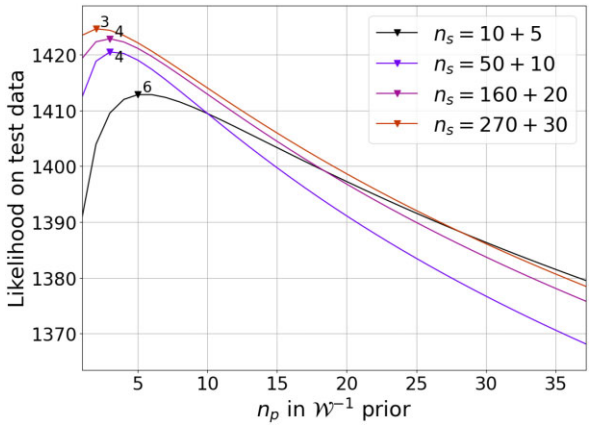


**Figure 8.** Same plot as in Fig. 2 for the matter correlation function, still with $\Psi_{\rm emp}$.

and would reduce the variance in the number function for the largest (and rarest) clusters in the simulations, thus increasing the effective volume of the simulations.

In other cases, the surrogates may consist of costly simulations that have already been run at a different set of parameters. In this case it may be possible to 'update' the means and covariances from the previous simulation set to a new set of parameters by pairing a small number of the existing old simulations (now surrogates) with the same number of new simulations.

The availability of perturbative results and analytical estimates, the increasing need for accurate simulations to analyse current and upcoming data sets in all subfields of cosmology, and the vast parameter space to explore with cosmological simulations make it likely that the concepts described here will continue to find powerful applications. We look forward to seeing the cosmological advances that CARPool will enable.

## ACKNOWLEDGEMENTS

We warmly thank Ethan Anderes and Francisco Villaescusa-Navarro for stimulating discussion and feedback. NC acknowledges funding from LabEx ENS-ICFP (PSL). BDW acknowledges support by the ANR BIG4 project, grant ANR-16-CE23-0002 of the French Agence Nationale de la Recherche; and the Labex ILP (reference ANR-10-LABX-63) part of the Idex SUPER, and received financial state aid managed by the Agence Nationale de la Recherche, as part of the programme Investissements d'avenir under the reference ANR-11-IDEX-0004-02. The Flatiron Institute is supported by the Simons Foundation. This work has made use of the Infinity Cluster hosted by Institut d'Astrophysique de Paris.

## DATA AVAILABILITY

The data samples underlying this article are available through *globus.org*, and instructions to reproduce the summary statistics from snapshots can be found at https://github.com/franciscovillaescusa/Quijote-simulations. Additionally, a PYTHON3 package with code examples and documentation is provided at https://github.com/CompiledAtBirth/pyCARPool to experiment with CARPool.

## REFERENCES

Alsing J., Wandelt B., 2018, MNRAS, 476, L60
Alsing J. et al., 2020, ApJS, 249, 5
Alves de Oliveira R., Li Y., Villaescusa-Navarro F., Ho S., Spergel D. N., 2020, preprint (arXiv:2012.00240)
Anderson T. W., 1957, J. Am. Stat. Assoc., 52, 200
Angulo R. E., Pontzen A., 2016, MNRAS, 462, L1
Angulo R. E., Zennaro M., Contreras S., Aricò G., Pellejero-Ibañez M., Stücker J., 2021, MNRAS, 507, 5869
Bai Z. D., Yin Y. Q., 1993, Ann. Probab., 21, 1275
Bernardeau F., Colombi S., Gaztanaga E., Scoccimarro R., 2002, Phys. Rep., 367, 1
Blot L., Corasaniti P. S., Alimi J.-M., Reverdy V., Rasera Y., 2014, MNRAS, 446, 1756
Blot L., Corasaniti P. S., Amendola L., Kitching T. D., 2016, MNRAS, 458, 4462
Blot L. et al., 2019, MNRAS, 485, 2806
Chartier N., Wandelt B. D., 2021, MNRAS, 509, 2220
Chartier N., Wandelt B., Akrami Y., Villaescusa-Navarro F., 2021, MNRAS, 503, 1897
Cheng S., Yu H.-R., Inman D., Liao Q., Wu Q., Lin J., 2020, 20th IEEE/ACM International Symposium on Cluster, Cloud and Internet Computing (CCGRID). p. 685
Chuang C.-H., Kitaura F.-S., Prada F., Zhao C., Yepes G., 2015, MNRAS, 446, 2621
Colavincenzo M. et al., 2019, MNRAS, 482, 4883
Dai B., Seljak U., 2020, Proc. Natl. Acad. Sci., 118, e2020324118
Dempster A. P., Laird N. M., Rubin D. B., 1977, J. R. Stat. Soc. Ser. B (Meth.), 39, 1
DeRose J. et al., 2019, ApJ, 875, 69
Desjacques V., Jeong D., Schmidt F., 2018, Phys. Rep., 733, 1
Ding Z. et al., 2022, MNRAS, 514, 3308
Dodelson S., Schneider M. D., 2013, Phys. Rev. D, 88, 063537
Eifler T., Schneider P., Hartlap J., 2009, A&A, 502, 721
Escoffier S. et al., 2016, preprint (arXiv:1606.00233)
Favole G., Granett B. R., Silva Lafaurie J., Sapone D., 2021, MNRAS, 505, 5833
Feng Y., Chu M.-Y., Seljak U., McDonald P., 2016, MNRAS, 463, 2273
Friedrich O., Eifler T., 2018, MNRAS, 473, 4150
Friedrich O. et al., 2021, MNRAS, 508, 3125
Gallier J., 2011, Geometric Methods and Applications. Springer, New York

Garrison L., 2019, PhD thesis, University Of Washington
Garrison L. H., Eisenstein D. J., Ferrer D., Tinker J. L., Pinto P. A., Weinberg D. H., 2018, ApJS, 236, 43
Giocoli C. et al., 2021, A&A, 653, A19
Habib S. et al., 2016, New Astron., 42, 49
Hall A., Taylor A., 2019, MNRAS, 483, 189
Harnois-Déraps J., Pen U.-L., 2013, MNRAS, 431, 3349
Harnois-Déraps J., Vafaei S., Van Waerbeke L., 2012, MNRAS, 426, 1262
Harnois-Déraps J., Pen U.-L., Iliev I. T., Merz H., Emberson J. D., Desjacques V., 2013, MNRAS, 436, 540
Harnois-Déraps J., Giblin B., Joachimi B., 2019, A&A, 631, A160
Hartlap J., Simon P., Schneider P., 2007, A&A, 464, 399
Hassan S. et al., 2021, preprint (arXiv:2110.02983)
He S., Li Y., Feng Y., Ho S., Ravanbakhsh S., Chen W., Póczos B., 2019, Proc. Natl. Acad. Sci., 116, 13825
Heavens A. F., Jimenez R., Lahav O., 2000, MNRAS, 317, 965
Hikage C., Takahashi R., Koyama K., 2020, Phys. Rev. D, 102, 083514
Howlett C., Manera M., Percival W., 2015, Astron. Comput., 12, 109
Ishiyama T., Fukushige T., Makino J., 2009, PASJ, 61, 1319
Izard A., Crocce M., Fosalba P., 2016, MNRAS, 459, 2327
Joachimi B., 2017, MNRAS, 466, L83
Joachimi B., Taylor A., 2014, Proc. IAU Symp. 306, Statistical Challenges in 21st Century Cosmology. Kluwer, Dordrecht, p. 99
Kasim M. F. et al., 2022, Mach. Learn.: Sci. Technol., 3, 015013
Kaufman G. M., 1967, Center for Operations Research and Econometrics Report no. 6710. Catholic University of Louvain. Heverlee, Belgium
Kitaura F. S., Yepes G., Prada F., 2014, MNRAS, 439, L21
Kodi Ramanah D., Charnock T., Villaescusa-Navarro F., Wandelt B. D., 2020, MNRAS, 495, 4227
Leclercq F., Faure B., Lavaux G., Wandelt B. D., Jaffe A. H., Heavens A. F., Percival W. J., 2020, A&A, 639, A91
Li Y., Singh S., Yu B., Feng Y., Seljak U., 2019, J. Cosmol. Astropart. Phys., 2019, 016
Lippich M. et al., 2019, MNRAS, 482, 1786
Lucie-Smith L., Peiris H. V., Pontzen A., 2019, MNRAS, 490, 331
Lucie-Smith L., Peiris H. V., Pontzen A., Nord B., Thiyagalingam J., 2020, preprint (arXiv:2011.10577)
McClintock T. et al., 2019a, preprint (arXiv:1907.13167)
McClintock T. et al., 2019b, ApJ, 872, 53
Maksimova N. A., Garrison L. H., Eisenstein D. J., Hadzhiyska B., Bose S., Satterthwaite T. P., 2021, MNRAS, 508, 4017
Modi C., Lanusse F., Seljak U., 2021a, Astron. Comput., 37, 100505
Modi C., Lanusse F., Seljak U., Spergel D. N., Perreault-Levasseur L., 2021b, preprint (arXiv:2104.12864)
Mohammed I., Seljak U., 2014, MNRAS, 445, 3382
Mohammed I., Seljak U., Vlah Z., 2017, MNRAS, 466, 780
Monaco P., Sefusatti E., Borgani S., Crocce M., Fosalba P., Sheth R. K., Theuns T., 2013, MNRAS, 433, 2389
Muirhead R. J., 1982, Aspects of Multivariate Statistical Theory. John Wiley & Sons, Inc, Hoboken, New Jersey
Paz D. J., Sánchez A. G., 2015, MNRAS, 454, 4326
Pearson D. W., Samushia L., 2016, MNRAS, 457, 993
Pedersen C., Font-Ribera A., Rogers K. K., McDonald P., Peiris H. V., Pontzen A., Slosar A., 2021, J. Cosmol. Astropart. Phys., 2021, 033
Percival W. J. et al., 2014, MNRAS, 439, 2531
Percival W. J., Friedrich O., Sellentin E., Heavens A., 2022, MNRAS, 510, 3207
Philcox O. H. E., Eisenstein D. J., 2019, MNRAS, 490, 5931
Philcox O. H. E., Eisenstein D. J., O'Connell R., Wiegand A., 2020, MNRAS, 491, 3290
Philcox O. H. E., Ivanov M. M., Zaldarriaga M., Simonović M., Schmittfull M., 2021, Phys. Rev. D, 103, 043508
Pontzen A., Slosar A., Roth N., Peiris H. V., 2016, Phys. Rev. D, 93, 103519
Pope A. C., Szapudi I., 2008, MNRAS, 389, 766
Potter D., Stadel J., Teyssier R., 2017, Comput. Astrophys. Cosmol., 4, 2
Remy B., Lanusse F., Ramzi Z., Liu J., Jeffrey N., Starck J.-L., 2020, preprint (arXiv:2011.08271)
Rogers K. K., Peiris H. V., 2021, Phys. Rev. D, 103, 043526
Schäfer J., Strimmer K., 2005, Stat. Appl. Genet. Mol. Biol., 4, Article32
Scoccimarro R., Sheth R. K., 2002, MNRAS, 329, 629
Sellentin E., Heavens A. F., 2018, MNRAS, 473, 2355
Smith A., de Mattia A., Burtin E., Chuang C.-H., Zhao C., 2021, MNRAS, 500, 259
Springel V., 2005, MNRAS, 364, 1105
Spurio Mancini A., Piras D., Alsing J., Joachimi B., Hobson M. P., 2022, MNRAS, 511, 1771
Taffoni G., Monaco P., Theuns T., 2002, MNRAS, 333, 623
Takahashi R. et al., 2009, ApJ, 700, 479
Tassev S., Zaldarriaga M., 2012, J. Cosmol. Astropart. Phys., 2012, 013
Tassev S., Zaldarriaga M., Eisenstein D. J., 2013, J. Cosmol. Astropart. Phys., 2013, 036
Tassev S., Eisenstein D. J., Wandelt B. D., Zaldarriaga M., 2015, preprint (arXiv:1502.07751)
Taylor A., Joachimi B., 2014, MNRAS, 442, 2728
Taylor A., Joachimi B., Kitching T., 2013, MNRAS, 432, 1928
Villaescusa-Navarro F. et al., 2018, ApJ, 867, 137
Villaescusa-Navarro F. et al., 2020, ApJS, 250, 2
Villaescusa-Navarro F. et al., 2021a, preprint (arXiv:2109.09747)
Villaescusa-Navarro F. et al., 2021b, ApJ, 915, 71
Villaescusa-Navarro F. et al., 2022, ApJS, 259, 61
Wadekar D., Ivanov M. M., Scoccimarro R., 2020, Phys. Rev. D, 102, 123521
Warren M. S., 2013, SC'13 – International Conference for High Performance Computing, Networking, Storage and Analysis. p. 1
White M., Tinker J. L., McBride C. K., 2014, MNRAS, 437, 2594
Yu H.-R., Pen U.-L., Wang X., 2018, ApJS, 237, 24
Zhai Z. et al., 2019, ApJ, 874, 95

## APPENDIX A: DERIVATION OF ESTIMATORS AND PROOF OF POSITIVE DEFINITENESS USING EXPECTATION MAXIMIZATION

### A1 Expectation maximization

In this section, we aim at showing the equivalence of the results given by the expectation–maximum (EM) algorithm – which naturally comes to mind in the presence of missing samples – and the simple result from the ML and MAP problems formulated in Sections 3.2 and 3.3.1.

#### *A1.1 Iterative algorithm*

The EM algorithm (Dempster et al. 1977) is an iterative technique to maximize the likelihood (or posterior) in the presence of missing data. Briefly, it works by casting the problem as a sequence of simpler optimization problems. Each iteration consists of two steps: the *E-step* which removes the missing data from the log-likelihood by taking its expectation with respect to the missing data assuming the current iterates are the true values of the parameters; and the *M-step* which updates the parameters by finding their values that maximize the expected log-likelihood from the E-step. We focus in this appendix on the covariance estimation; including the solution for the estimators of the mean is straightforward and we give the result in the main text.

We recall equation (7) here for convenience as a starting point

$$-2\ln\left[\mathcal{L}(\{\boldsymbol{x}\},\{\boldsymbol{x}^*\}|\boldsymbol{\Sigma})\right] = (n_s+n_r)\ln\left[\det(2\pi\boldsymbol{\Sigma})\right] + \left(\sum_{i=1}^{n_s}\boldsymbol{x}_i^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}_i\right) + \left(\sum_{j=1}^{n_r}\boldsymbol{x}_j^{*T}\boldsymbol{\Sigma}^{-1}\boldsymbol{x}_j^*\right), \quad \text{(A1)}$$

**E-step.** Consider conditional expectation of the log-likelihood over missing data $\boldsymbol{s}^*$ given the (observed) data and the covariance at

the $k$-th step, $\mathbf{\Sigma}[k]$

$$-2\mathbb{E}_{s^*|r^*}\left[\ln\left[\mathcal{L}(\{\boldsymbol{x}\},\{\boldsymbol{x}^*\}|\mathbf{\Sigma}[k])\right]\right] = (n_s+n_r)\ln\left[\det\left(\mathbf{\Sigma}[k]\right)\right] + \left(\sum_{i=1}^{n_s}\boldsymbol{x}_i^{\boldsymbol{T}}\mathbf{\Sigma}[k]^{-1}\boldsymbol{x}_i\right) + \mathbb{E}_{s^*|r^*}\left[\left(\sum_{j=1}^{n_r}\boldsymbol{x}_j^{*\boldsymbol{T}}\mathbf{\Sigma}[k]^{-1}\boldsymbol{x}_j^*\right)\right] + c$$

Using linearity of expectation we can look at each summand of the last term on the RHS

$$\mathbb{E}\left[\left(\boldsymbol{x}_i^{*\boldsymbol{T}}\mathbf{\Sigma}[k]^{-1}\boldsymbol{x}_i^*\right)\right] = \mathrm{tr}\left(\mathbf{\Sigma}[k]^{-1}\mathbb{E}\left[\boldsymbol{x}_i^*\boldsymbol{x}_i^{*\boldsymbol{T}}\right]\right) \tag{A2}$$

Define

$$\boldsymbol{A}_i \equiv \mathbb{E}\left[\boldsymbol{x}_i^*\boldsymbol{x}_i^{*\boldsymbol{T}}\right] = \begin{pmatrix}\boldsymbol{A}_{i,ss} & \boldsymbol{A}_{i,sr}\\ \boldsymbol{A}_{i,sr}^{\boldsymbol{T}} & \boldsymbol{A}_{i,rr}\end{pmatrix}. \tag{A3}$$

Then

$$\boldsymbol{A}_{i,ss} = (\mathbf{\Sigma}/\mathbf{\Sigma}_{\boldsymbol{rr}}) + \boldsymbol{B}\boldsymbol{r}_i^*\boldsymbol{r}_i^{*\boldsymbol{T}}\boldsymbol{B}^{\boldsymbol{T}} \tag{A4}$$

$$\boldsymbol{A}_{i,sr} = \boldsymbol{B}\boldsymbol{r}_i^*\boldsymbol{r}_i^{*\boldsymbol{T}} \tag{A5}$$

$$\boldsymbol{A}_{i,rr} = \boldsymbol{r}_i^*\boldsymbol{r}_i^{*\boldsymbol{T}} \tag{A6}$$

We stress that equations (A4), (A5), and (A6) depend on $k$ because we use $\mathbf{\Sigma}[k]$ as $\mathbf{\Sigma}$.

Writing

$$n_s\widehat{\mathbf{\Sigma}} = \sum_{i=1}^{n_s}\boldsymbol{x}_i\boldsymbol{x}_i^{\boldsymbol{T}}$$

and

$$n_r\boldsymbol{A}[k] = \sum_{i=1}^{n_r}\boldsymbol{A}_i[k],$$

we find the expected log-likelihood

$$-2\mathbb{E}_{s^*|r^*}\left[\ln\left[\mathcal{L}(\{\boldsymbol{x}\},\{\boldsymbol{x}^*\}|\mathbf{\Sigma}[k])\right]\right] = (n_s+n_r)\ln\left[\det(\mathbf{\Sigma}[k])\right] + \mathrm{tr}\left[\mathbf{\Sigma}^{-1}\left(n_s\widehat{\mathbf{\Sigma}} + n_r\boldsymbol{A}[k]\right)\right] + c \tag{A7}$$

**M-step.** Maximizing the expected log-likelihood, equation (A7) to find the next value of the parameter is now trivial:

$$\mathbf{\Sigma}[k+1] = \frac{1}{n_s+n_r}\left(n_s\widehat{\mathbf{\Sigma}} + n_r\boldsymbol{A}[k]\right) \tag{A8}$$

### *A1.2 Inclusion of an Inverse–Wishart prior for* $\mathbf{\Sigma}$

The generalization to maximizing the posterior for $\mathbf{\Sigma}$ with a conjugate prior taking the Inverse–Wishart form is immediate. Taking $\mathbf{\Psi}$ to be the parameter of the prior, $P = 2\dim(s)$, and $\nu > P-1$ the number of degrees of freedom, then this modifies the EM update, equation (A8) to

$$\mathbf{\Sigma}[k+1] = \frac{1}{n_s+n_r+(\nu+P+1)}\left(n_s\widehat{\mathbf{\Sigma}} + n_r\boldsymbol{A}[k] + \mathbf{\Psi}\right) \tag{A9}$$

When $n_s \approx P$, the MAP estimator is quite different from the ML estimator.

### *A1.3 Proof that EM iterations conserve positive (semi)definiteness of* $\mathbf{\Sigma}$

To prove the positive definiteness of the estimated covariance matrix, we recall the following very useful characterization of *positive semidefinite* (*psd*) matrices using the Schur complement:

**Lemma** (e.g. Gallier 2011): Let $\boldsymbol{M}_{22}$ be positive definite, $\boldsymbol{M}_{22} > 0$. Then

$$\boldsymbol{M} = \begin{pmatrix}\boldsymbol{M}_{11} & \boldsymbol{M}_{12}\\ \boldsymbol{M}_{12}^{\boldsymbol{T}} & \boldsymbol{M}_{22}\end{pmatrix} \geq 0 \tag{A10}$$

if and only if $(\boldsymbol{M}/\boldsymbol{M}_{22}) \geq 0$.

We wish to show that as long as we have enough surrogates such that covariance matrix estimated from them is positive definite, then it is true that if we initialize $\mathbf{\Sigma}[0]$ such that $(\mathbf{\Sigma}[0]/\mathbf{\Sigma}_{\boldsymbol{rr}}[0]) \geq 0$ then $\mathbf{\Sigma}[k] \geq 0$ throughout the EM iteration and therefore also for the fixed point. This follows directly from Lemma 1, as follows.

At step $k$ of the EM iteration assume $\mathbf{\Sigma}[k]$ is such that $(\mathbf{\Sigma}[k]/\mathbf{\Sigma}_{\boldsymbol{rr}}[k]) \geq 0$. By assumption we always have enough surrogates so that $\boldsymbol{A}_{rr} > 0$. Therefore $\boldsymbol{A}_{\boldsymbol{rr}}$ is invertible and we have that

$$(\boldsymbol{A}/\boldsymbol{A}_{\boldsymbol{rr}}) = (\mathbf{\Sigma}[k]/\mathbf{\Sigma}_{\boldsymbol{rr}}[k]) \geq 0 \tag{A11}$$

by assumption. This implies $\boldsymbol{A} \geq 0$ by the Lemma. The sum of two *psd* matrices is itself *psd*, and since $\widehat{\mathbf{\Sigma}}$ is manifestly *psd*, this guarantees that $\mathbf{\Sigma}[k+1] \geq 0$. The 'only if' direction of the Lemma guarantees that $(\mathbf{\Sigma}[k+1]/\mathbf{\Sigma}_{\boldsymbol{rr}}[k+1]) \geq 0$ at the next iteration. Therefore, $\mathbf{\Sigma}[k] \geq 0$ for all $i \geq 0$ by induction.

### *A1.4 The Maximum Likelihood and A Posteriori solutions as Fixed Point of the EM iterations*

While the iterations are computationally very light, since we have closed-form solutions for the iterative updates (Sections 3.2 and 3.3), we can do even better by deriving a closed-form solution directly for the iterative fixed point and thus demonstrate the equivalence with EM. Solving $\mathbf{\Sigma}[k+1] = \mathbf{\Sigma}[k] \equiv \widehat{\mathbf{\Sigma}}^{\mathrm{MAP}}$ by combining equation (A8) with equations (A4), (A5), and (A6) gives

$$\widehat{\mathbf{\Sigma}}_{\boldsymbol{rr}}^{\mathrm{EM}} = \frac{n_r\boldsymbol{A}_{\boldsymbol{rr}} + (n_s+n_p)\widehat{\mathbf{\Sigma}}_{\boldsymbol{rr}}^{\boldsymbol{\Delta}}}{n_s+n_r+n_p} \tag{A12}$$

$$\widehat{\mathbf{\Sigma}_{\boldsymbol{sr}}}^{\mathrm{EM}} = (n_s+n_p)\widehat{\mathbf{\Sigma}}_{\boldsymbol{sr}}^{\boldsymbol{\Delta}} \tag{A13}$$

$$\times\left((n_s+n_r+n_p)\mathbb{1}_{p_r} - [\widehat{\mathbf{\Sigma}_{rr}}^{\mathrm{EM}}]^{-1}n_r\boldsymbol{A}_{\boldsymbol{rr}}\right)^{-1} \tag{A14}$$

$$\widehat{\boldsymbol{B}}^{\mathrm{EM}} = \widehat{\mathbf{\Sigma}_{\boldsymbol{sr}}}^{\mathrm{EM}}[\widehat{\mathbf{\Sigma}_{rr}}^{\mathrm{EM}}]^{-1} \tag{A15}$$

$$\widehat{\mathbf{\Sigma}_{\boldsymbol{ss}}}^{\mathrm{EM}} = \frac{(n_s+n_p)\widehat{\mathbf{\Sigma}}_{\boldsymbol{ss}}^{\boldsymbol{\Delta}} + n_r\widehat{\boldsymbol{B}}^{\mathrm{EM}}(\boldsymbol{A}_{\boldsymbol{rr}} - \widehat{\mathbf{\Sigma}_{\boldsymbol{rr}}}^{\mathrm{EM}})\widehat{\boldsymbol{B}}^{\mathrm{EM}^{\boldsymbol{T}}}}{n_s+n_p} \tag{A16}$$

Equation (A16) is equivalent to equation (17), even though it looks more complicated. This is because solving for the EM introduces the block covariance $\boldsymbol{A}$ of the unpaired surrogates and missing simulations vector $\boldsymbol{x}^*$ (the $\boldsymbol{A}_{\boldsymbol{rr}}$ block is $\widehat{\mathbf{\Sigma}_{rr}}$) and not the covariance of the paired and unpaired surrogates $\widehat{\mathbf{\Sigma}}_{\boldsymbol{rr}}^{\star}$.

## A2 Case when the surrogate covariance is known

We can rewrite the simulation summary statistics covariance from Section 3.3.1 as

$$\widehat{\mathbf{\Sigma}_{\boldsymbol{ss}}^{\mathrm{MAP}}} = \widehat{\mathbf{\Sigma}}_{\boldsymbol{ss}}^{\boldsymbol{\Delta}} + \frac{n_r}{n_r+n_s+n_p}\widehat{B}_{\mathrm{MAP}}\left(\widehat{\mathbf{\Sigma}_{rr}} - \widehat{\mathbf{\Sigma}_{rr}}^{\boldsymbol{\Delta}}\right)\widehat{B}_{\mathrm{MAP}}^{\boldsymbol{T}}. \tag{A17}$$

which is strictly equivalent to equation (17).

The case when a theoretical covariance $\mathbf{\Sigma}_{rr}$ for the surrogates is available directly obtains from the limit of equation (A17) as $n_r \to$

$\infty$

$$\widehat{\boldsymbol{\Sigma}}_{ss}^{\mathbf{MAP},\boldsymbol{\Sigma}_{rr}} = \lim_{n_r \to \infty} \widehat{\boldsymbol{\Sigma}}_{ss}^{\mathbf{MAP}} = \widehat{\boldsymbol{\Sigma}}_{ss}^{\boldsymbol{\Delta}} + \widehat{B}_{\mathrm{MAP}} \left( \boldsymbol{\Sigma}_{rr} - \widehat{\boldsymbol{\Sigma}}_{rr}^{\boldsymbol{\Delta}} \right) \widehat{B}_{\mathrm{MAP}}^{T}. \tag{A18}$$

## APPENDIX B: MAP DERIVATION (REGRESSION PARAMETERS)

This section presents the derivation of the closed-form solutions for the covariance in Section 3.3.2. We can extend Anderson's in Anderson (1957) derivation by including an Inverse-Wishart Prior with parameters $\boldsymbol{\Psi}$ and $\nu$. Under the hypothesis that the block covariance $\Sigma$ of simulation and surrogates summary statistics is drawn from and Inverse–Wishart distribution (equation 11), the following properties hold true:

(i) $\Sigma_{rr} \perp\!\!\!\perp {\Sigma_{rr}}^{-1} \Sigma_{rs} = B^T$.
(ii) $\boldsymbol{\Sigma_{rr}} \perp\!\!\!\perp \boldsymbol{\Sigma_{s|r}}$.
(iii) $\boldsymbol{\Sigma_{rr}} \sim \mathcal{W}^{-1}(\boldsymbol{\Psi_{rr}}, \nu - p_s)$.
(iv) $\boldsymbol{\Sigma_{s|r}} \sim \mathcal{W}^{-1}(\boldsymbol{\Psi_{s|r}}, \nu)$ with $\boldsymbol{\Psi_{s|r}} \equiv (\boldsymbol{\Psi}/\boldsymbol{\Psi_{rr}})$.
(v) $B^T | \Sigma_{s|r} \sim \mathcal{MN}({\Psi_{rr}}^{-1}\Psi_{rs}, \Sigma_{s|r} \otimes {\Psi_{rr}}^{-1})$,

where $\perp\!\!\!\perp$ indicates probabilistic independence, $\otimes$ is the Kronecker product and $\mathcal{MN}$ designates the matrix normal distribution. This is particularly convenient for our problem and we can extend Anderson's result straightforwardly to a MAP estimate. In particular, we can reparametrize the distribution

$$\mathcal{P}(\boldsymbol{\Sigma}) = \mathcal{P}(\boldsymbol{B^T} | \boldsymbol{\Sigma_{s|r}}) \mathcal{P}(\boldsymbol{\Sigma_{s|r}}) \mathcal{P}(\boldsymbol{\Sigma_{rr}}). \tag{B1}$$

Let's index the unpaired surrogate samples as $\boldsymbol{r}_i^*$ with $i = 1, \ldots, n_r$ and the surrogate samples that are part of the pairs $\boldsymbol{x}_i i = 1, \ldots, n_s$ as $\boldsymbol{r}_i$ with $i = n_r + 1, \ldots, n_r + n_s$. We factorize the likelihood as Anderson, that is to say

$$\begin{aligned} \mathcal{L}(\{\boldsymbol{x}\}, \{\boldsymbol{r}^*\} | \boldsymbol{\Sigma}) &= \prod_{i=1}^{n_s} \mathcal{P}(\boldsymbol{x}_i | \boldsymbol{\mu_s}, \boldsymbol{\Sigma_{ss}}) \prod_{j=1}^{n_r} \mathcal{P}(\boldsymbol{r}_j^* | \boldsymbol{\mu_r}, \boldsymbol{\Sigma_{rr}}) \\ &= \prod_{i=1}^{n_s+n_r} \mathcal{P}(\boldsymbol{r}_i | \boldsymbol{\mu_r}, \boldsymbol{\Sigma_{rr}}) \prod_{i=1}^{n_s} \mathcal{P}(\boldsymbol{s}_i | \boldsymbol{\mu}_{s_i|r_i}, \boldsymbol{\Sigma_{s|r}}) \end{aligned} \tag{B2}$$

The right-hand side depends separately on $\boldsymbol{\Sigma_{rr}}$, $\boldsymbol{\Sigma_{s|r}}$ and $\boldsymbol{B^T}$ (through $\boldsymbol{\mu_{s|r}}$) as the prior, so we can solve the MAP problem from equation (12). Note that since in the factorization we regroup the unpaired and paired surrogates, we drop the $\boldsymbol{r}^*$ notation and just use $\boldsymbol{r}$ for the rest of the derivation.

Then the natural logarithm posterior distribution is

$$\begin{aligned} -2 \ln [\mathcal{P}(\boldsymbol{\Sigma} | \{\boldsymbol{s}\}, \{\boldsymbol{r}\}] &= (n_s + n_r + \nu - p_s + p_r + 1) \ln [\det(\boldsymbol{\Sigma_{rr}})] \\ &+ (n_s + \nu + 2p_s + 1) \ln \left[\det(\boldsymbol{\Sigma_{s|r}})\right] \\ &+ \mathrm{Tr}\left( \left[ \sum_{i=1}^{n_s+n_r} (\boldsymbol{r}_i - \boldsymbol{\mu_r})(\boldsymbol{r}_i - \boldsymbol{\mu_r})^T + \boldsymbol{\Psi_{rr}} \right] {\boldsymbol{\Sigma_{rr}}}^{-1} \right) \\ &+ \mathrm{Tr}\left( \left[ \sum_{j=1}^{n_s} (\boldsymbol{s}_j - \boldsymbol{\mu_{s|r}})(\boldsymbol{s}_j - \boldsymbol{\mu_{s|r}})^T + \boldsymbol{\Psi_{s|r}} \right. \right. \\ &\left. \left. + (\boldsymbol{B^T} - \boldsymbol{\Gamma^T})^T \boldsymbol{\Psi_{rr}} (\boldsymbol{B^T} - \boldsymbol{\Gamma^T}) \right] {\boldsymbol{\Sigma_{s|r}}}^{-1} \right) \end{aligned} \tag{B3}$$

We then solve successively $\frac{\partial \ln[\mathcal{P}(\boldsymbol{\Sigma}|\{\boldsymbol{s}\},\{\boldsymbol{r}\}]}{\partial \alpha} = 0$ for $\alpha \in \left\{ \boldsymbol{\Sigma_{rr}}, \boldsymbol{B^T}, \boldsymbol{\Sigma s|r} \right\}$, after a bit of derivation and linear algebra, and we find the solutions from Section 3.3.2, i.e. equations (19), (20), and (21) which allow to compute $\widehat{\boldsymbol{\Sigma}}_{ss}^{\mathbf{MAP}}$ from equation (22).

## APPENDIX C: SOME ADDITIONAL RESULTS

### C1 Relative gain for the power spectrum

We simply show the confidence bounds for the ΛCDM parameters using the power spectrum covariance matrix, this time with more simulations for the CARPool Bayes covariance, i.e $n_s = 40 + 10$ versus $n_s = 10 + 5$ in Section 2. Fig. C1 shows CARPool Bayes marginal bounds even closer to the truth than in Fig. 1 at the price of running 50 simulations in total instead of 15. This demonstrates the relative gain of running more simulations is small for the covariance matrix when the simulation and surrogate summary statistics are well correlated.

### C2 MAP on the regression parameters

We chose to present one particular example of the MAP estimate from Section 3.3.2 on the power spectrum, which showed the most successful results with the 'block' parametrization from Section 3.3.1 with only $n_s = 10 + 5$ simulations. We fix $\nu = 2{*}p_s + 2$ in this case and do not consider it a free parameter, nor do we allow it to define an improper prior, i.e. we do not allow $\nu \leq 2{*}ps - 1$. This corresponds to the lowest integer for which the expectation of the Inverse–Wishart exists. In Fig. C2, the marginal confidence bounds are much wider than the truth for both $n_s = 10$ and $n_s = 160$ for the CARPool Bayes estimator (this time the 'regression' framework from Section 3.3.2). Since the MAP on the regression parameters does not allow for an improper prior, the estimator of the simulation covariance puts too much weight on the naïve diagonal empirical Bayes prior we use (Section 3.4.2). For future studies, we can explore whether having a 'smarter prior', for instance a model covariance computed from theoretical approximations to parametrize the Inverse–Wishart distribution, can significantly improve or not both the CARPool Bayes estimators from Sections 3.3.1 and 3.3.2.

### C3 Results from the bispectrum

Here, we directly present the confidence bounds for the ΛCDM parameter found using various covariance estimators of the bispectrum. The motivation here is to demonstrate the improvement over CW21 for the same summary statistics. The first summary statistics we test is the set of squeezed isosceles triangles, that is to say the bispectra computed for $k_1 = k_2$ and in ascending order of the ratio $k_3/k_1 \leq 0.20$ ($p_s = 98$ in this case).

Fig. C3 demonstrates that we get parameter constraints much more representative of the truth with $n_s = 20 + 10$ simulations that with the sample covariance using $n_s = 110$ simulations. The CARpool Bayes estimator is the one from Section 3.3.1 using the empirical Bayes prior from Section 3.4.2.

Then, we take a look at the reduced bispectrum of equilateral equilateral triangles with $k_1 = k_2 = k_3$ varying up to $k_{\mathrm{max}} = 0.75\,h\mathrm{Mpc}^{-1}$ ($p_s = 40$). In Fig. C4, we observe the CARPool Bayes estimator gives almost identical parameters marginal contours to the truth with only $n_s = 10 + 5$ simulations, while the sample covariance of simulations uses $n_s = 100$ simulations.

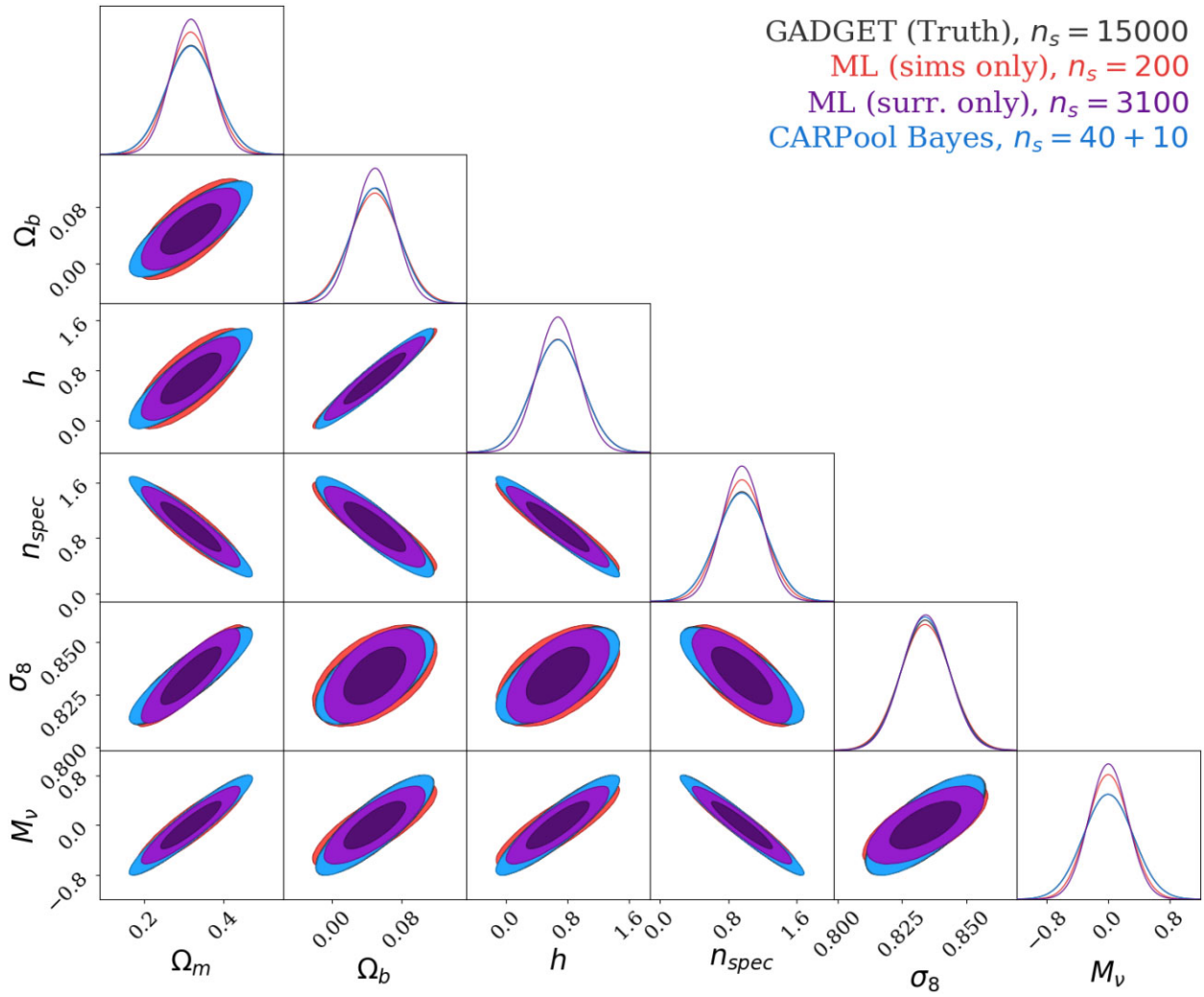


**Figure C1.** Fisher confidence contours of the cosmological parameters based on the estimated covariance matrix of the matter power spectrum. The estimators which we compare are the same as in Fig. 1, except that we have now $n_s = 40 + 10$ simulations for CARPool Bayes (empirical Bayes prior on the block covariance).

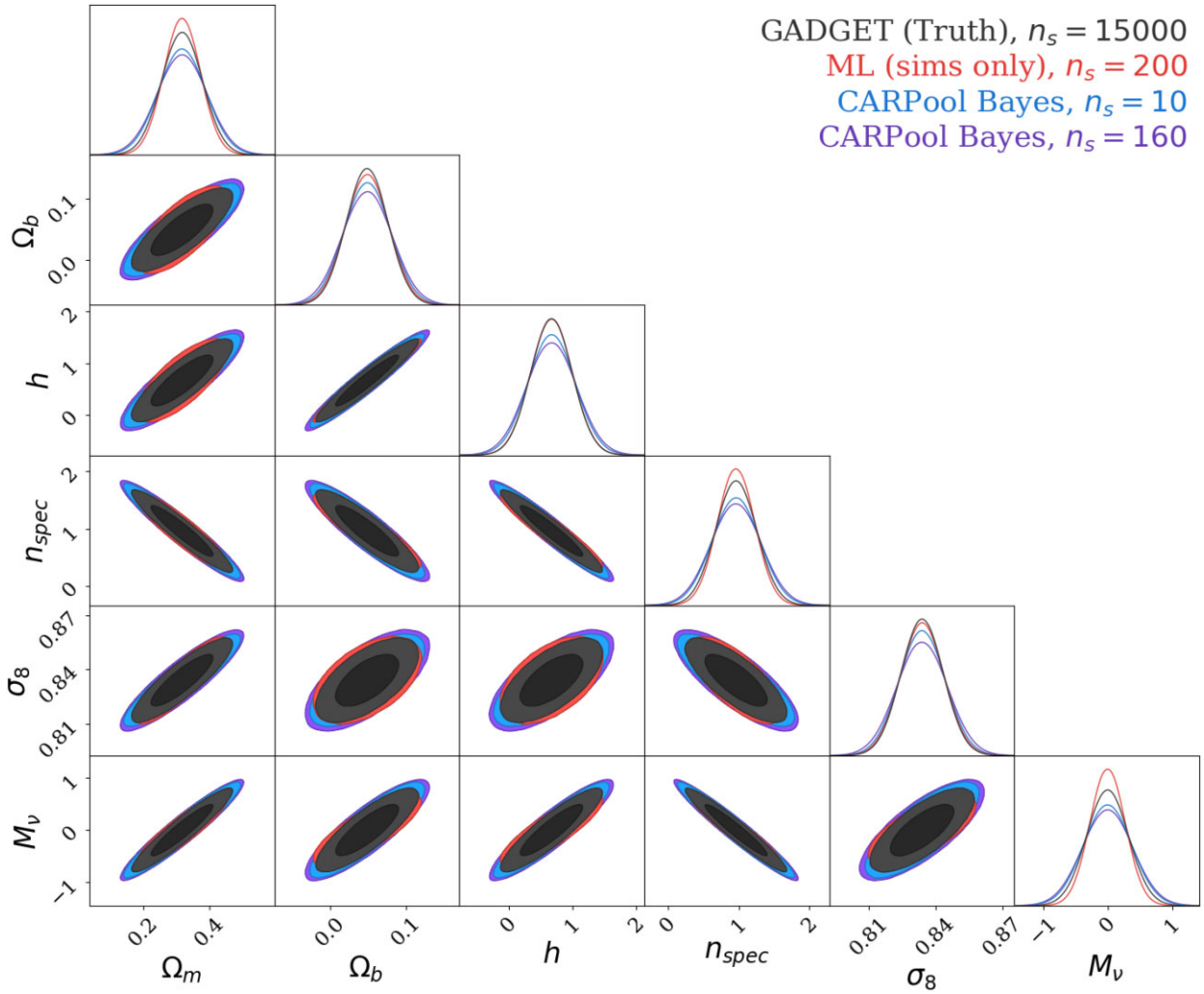


**Figure C2.** Confidence contours of the cosmological parameters computed using the Fisher matrix based on the estimated matter power spectrum covariance The 'CARPool Bayes' estimates follow the computations of Section 3.3.2, where the prior, still the empirical Bayes one from Section 3.4.2, is parametrized given the regression parameters. We stress that this is the only Figure in the paper that shows a computation of the 'regression' MAP from Section 3.3.2.

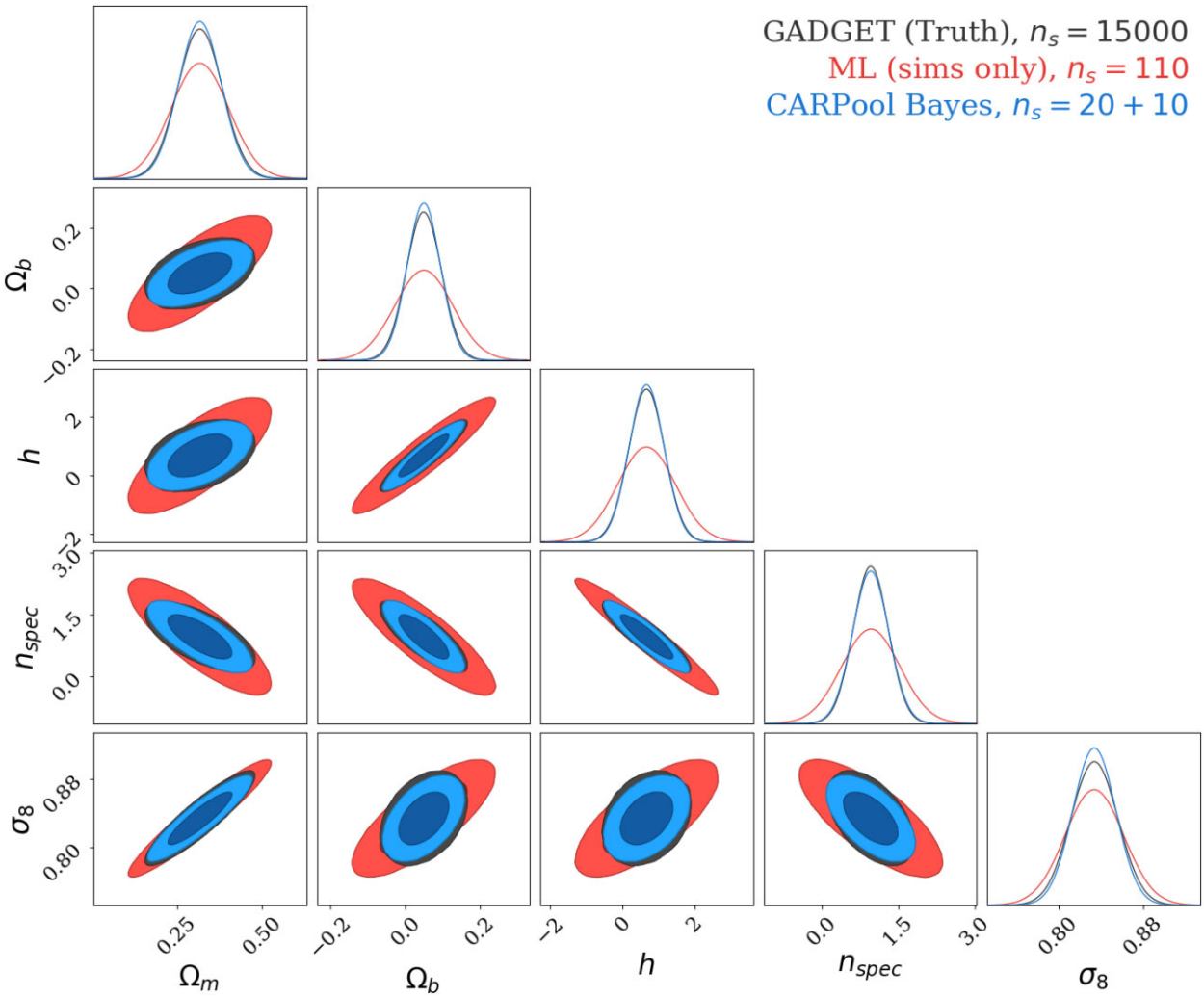


**Figure C3.** Confidence contours of the cosmological parameters computed using the Fisher matrix based on the estimated matter bispectrum covariance matrix, for a set of squeezed isosceles triangles. The estimators result from the same computations as in Fig. 1.

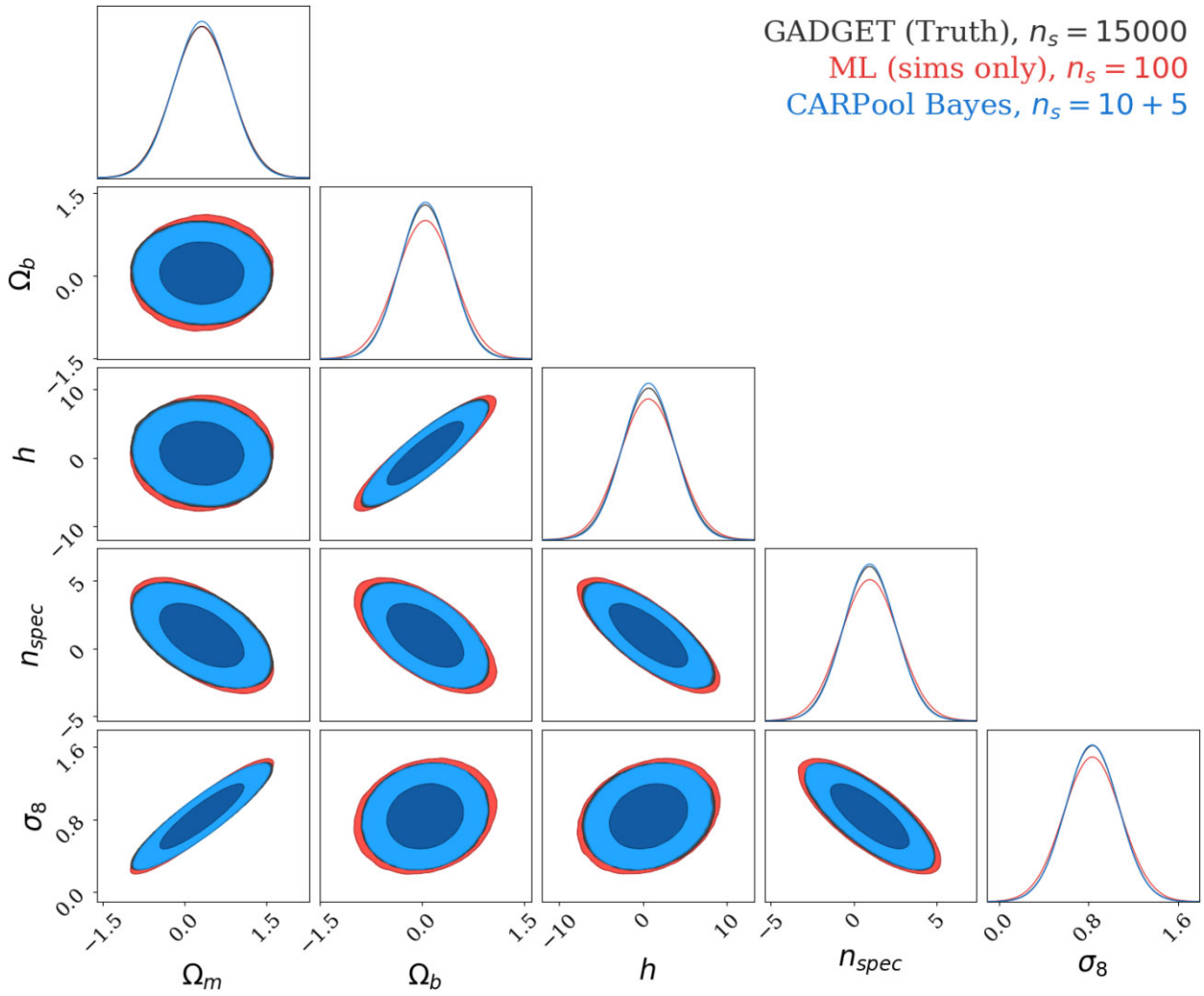


**Figure C4.** Confidence contours of the cosmological parameters computed using the Fisher matrix based on the estimated matter bispectrum covariance matrix, for a set of equilateral triangles. The estimators we compare are the same as in Fig. 1.

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