Fisher Score Matching for Simulation-Based Forecasting and Inference

Ce Sui [1 2] **Shivam Pandey** [2] **Benjamin D. Wandelt** [2 3]

Abstract

We propose a method for estimating the Fisher score—the gradient of the log-likelihood with respect to model parameters—using score matching. By introducing a latent parameter model, we show that the Fisher score can be learned by training a neural network to predict latent scores via a mean squared error loss. We validate our approach on a toy linear Gaussian model and a cosmological example using a differentiable simulator. In both cases, the learned scores closely match ground truth for plausible data-parameter pairs. This method extends the ability to perform Fisher forecasts, and gradient-based Bayesian inference to simulation models, even when they are not differentiable; it therefore has broad potential for advancing cosmological analyses.

1. Introduction

Statistical inference is a core component of modern astronomical and cosmological research. A typical inference pipeline involves several key stages: identifying informative observables, constructing summary statistics, and performing parameter inference. A wide array of tools has been developed to address each of these tasks.

Decades of research on cosmological structure have centered on defining informative summaries and observables, principally amongst them the power spectrum. New surveys, such as those focusing on galaxy clustering, 21cm intensity mapping, and weak lensing (DES Collaboration et al., 2025; Euclid Collaboration et al., 2024; Square Kilometre Array Cosmology Science Working Group et al., 2020) motivate more sophisticated summaries (Beyond-2pt Collaboration et al., 2024; Makinen et al., 2024; Charnock et al., 2018; Cheng et al., 2020). These include both hand-crafted features and machine learning-based methods, aimed at extracting the maximal amount of information from increasingly complex data. To forecast constraints based on these observables and summaries, the Fisher information matrix has been widely used after being introduced to cosmology by Tegmark et al. (1997).

Given observational data summary statistics and a physical model, the next step is parameter inference. Most current cosmological analyses use Bayesian inference (*e.g.*, DES Collaboration et al. (2025); Planck Collaboration et al. (2020)). Traditionally, this involves assuming an explicit form for the likelihood and using MCMC algorithms for posterior sampling. Numerous MCMC packages have been developed to support this (Torrado & Lewis, 2021; Brinckmann & Lesgourgues, 2018). More recently, simulation-based inference (SBI) (Cranmer et al., 2020; Alsing et al., 2019) has allowed approximating the likelihood implicitly using simulations. SBI methods can be more flexible and efficient than standard MCMC, and many recent tools implement SBI techniques to address a wide range of cosmological problems (Ho et al., 2024; Boelts et al., 2025; Zhao et al., 2022; Lemos et al., 2023).

All of these steps—summary construction, Fisher forecasting, and inference—can be greatly simplified if we have access to the Fisher score function: the gradient of the log-likelihood with respect to the model parameters. The Fisher score function itself is a sufficient statistic, since it encodes the likelihood surface up to a constant factor. When evaluated at a fixed, fiducial parameter value, it serves as a locally sufficient statistic (Heavens et al., 2017; Alsing & Wandelt, 2018) and enables straightforward computation of the Fisher matrix. It serves to construct unbiased, minimum-variance estimators that saturate the Cramér-Rao information inequality, and to compute the maximum likelihood estimator. Furthermore, it opens the door to using more efficient gradient-based approaches such as Hamiltonian Monte Carlo (HMC) samplers; or to frequentist analysis avoiding prior choice, such as maximum likelihood estimators.

This surprising utility of the Fisher score has motivated recent efforts to re-implement cosmological simulations using differentiable programming frameworks like JAX[1]

[1]Department of Astronomy, Tsinghua University, Beijing 100084, China. [2]Department of Physics and Astronomy, Johns Hopkins University, Baltimore, MD 21218, USA [3]Department of Applied Mathematics and Statistics, Johns Hopkins University, Baltimore, MD 21218, USA. Correspondence to: Ce Sui <suic20@mails.tsinghua.edu.cn>.

ML4Astro worshop, Vancouver, Canada, 2025.

[1]http://github.com/jax-ml/jax

(Bradbury et al., 2018) that enable automatic differentiation throughout the simulation pipeline and make Fisher score computation feasible (Li et al., 2022; Campagne et al., 2023). In parallel, score matching—a technique used to train neural networks to approximate not the Fisher score, but the probability score $\nabla_x \log P(x)$—has seen widespread success in generative modeling, most notably in diffusion models (Song & Ermon, 2019; Song et al., 2020). Similar techniques have also been adapted to the context of SBI and cosmological inference tasks (Sharrock et al., 2022; Cuesta-Lazaro & Mishra-Sharma, 2024; Mudur et al., 2025).

Here, we propose a new score matching approach for learning the Fisher score directly from simulations. We demonstrate that the learned scores are accurate for data-parameter pairs relevant during inference and can be used for both Fisher analysis and Bayesian inference. Section 2 describes our method, Section 3 presents experimental results, and Section 4 concludes with a discussion of future directions.

2. Method

Given a probabilistic model $P(x \mid \theta)$ describing a physical system, a key quantity of interest is the Fisher score, defined as:

$$s(x, \theta) = \nabla_\theta \log P(x \mid \theta). \tag{1}$$

However, in many cases the likelihood $P(x \mid \theta)$ is only implicitly defined through a simulator, without an analytical form or differentiable implementation, making direct computation of the Fisher score intractable. We propose a method to estimate the Fisher score using score matching.

2.1. Fisher Score Matching

We assume the existence of a latent variable θ^* such that the following Markov chain holds: $\theta \to \theta^* \to x$. That is,

$$P(x \mid \theta) = \int P(x \mid \theta^*) P(\theta^* \mid \theta)\, d\theta^*,$$

where $P(\theta^* \mid \theta)$ defines the latent model. Then we can re-write the Fisher score of the full model as an expectation of the latent model's score as follows:

$$\begin{aligned}
\nabla_\theta \log P(x|\theta) &= \frac{\nabla_\theta P(x|\theta)}{P(x|\theta)} \\
&= \frac{\nabla_\theta \int P(x|\theta^*, \theta) P(\theta^*|\theta) d\theta^*}{P(x|\theta)} \\
&= \frac{\int P(x|\theta^*) \nabla_\theta P(\theta^*|\theta) d\theta^*}{P(x|\theta)} \\
&= \int \frac{P(x|\theta^*) P(\theta^*|\theta)}{P(x|\theta)} \nabla_\theta \log P(\theta^*|\theta) d\theta^* \\
&= \int P(\theta^*|x, \theta) \nabla_\theta \log P(\theta^*|\theta) d\theta^* \\
&= E_{P(\theta^*|x,\theta)}[\nabla_\theta \log P(\theta^*|\theta)]. \qquad (2)
\end{aligned}$$

Since posterior means can be estimated by minimizing mean squared error (see Appendix A), we can train a score estimator $s(x, \theta)$ by minimizing the loss

$$L = E_{p(x,\theta,\theta^*)}[(s(x, \theta) - \nabla_\theta \log P(\theta^*|\theta))^2] \tag{3}$$

Training proceeds as follows: we sample from the joint distribution $p(x, \theta, \theta^*)$ using forward simulations. For each triplet (x, θ, θ^*), the model is asked to predict the latent score $\nabla_\theta \log p(\theta^* \mid \theta)$ given the input (x, θ). Once training converges, the estimator $s(x, \theta)$ provides a deterministic approximation of the true Fisher score $\nabla_\theta \log p(x \mid \theta)$. Since the model is essentially performing a regression task, we expect its scalability with input dimension to follow typical patterns observed in regression problems, depending on the architecture used to learn the score.

2.2. Decomposable and Indecomposable Models

If we can identify a latent variable such that the Markov chain $\theta \to \theta^* \to x$ holds and the gradient $\nabla_\theta \log P(\theta^* \mid \theta)$ is tractable, this approach can be applied directly, a fact also exploited by Brehmer et al. (2020).

A simple example is the linear Gaussian model $x \mid \theta \sim \mathcal{N}(\theta, \Sigma_n)$. We can decompose this into two conditional models

$$\theta^* \mid \theta \sim \mathcal{N}(\theta, \Sigma_{\theta^*}), \quad x \mid \theta^* \sim \mathcal{N}(\theta^*, \Sigma_n - \Sigma_{\theta^*}), \tag{4}$$

where both covariance matrices are positive definite. The Fisher score of the full model can then be learned using only the score from the latent model.

In this case, the choice of decomposition does not affect the result—the training objective will still converge to the true Fisher score of the full model.

To address more complex settings where no natural latent variable is available, we introduce an auxiliary variable $\tilde{\theta}$ and define a latent model $P(\theta \mid \tilde{\theta})$, forming the Markov chain $\tilde{\theta} \to \theta \to x$. We simulate samples using

$$P(\theta, x \mid \tilde{\theta}) = P(x \mid \theta) P(\theta \mid \tilde{\theta}),$$

and apply the same score matching technique to estimate $\nabla_{\tilde{\theta}} \log P(x \mid \tilde{\theta})$. This approximation becomes accurate when $P(\theta \mid \tilde{\theta})$ is sharply peaked (i.e., close to a delta function). A natural choice for the latent model is an additive noise model where $\theta = \tilde{\theta} + w$, with w drawn from a known noise distribution. Under this model, it can be shown (see Appendix B) that

$$\nabla_{\tilde{\theta}} \log P(x \mid \tilde{\theta}) = \int \nabla_\theta \log P(x \mid \theta)\, P(\theta \mid \tilde{\theta}, x)\, d\theta, \tag{5}$$

i.e., the estimated score is a convolution of the true score with a convolution kernel $P(\theta \mid \tilde{\theta}, x)$ that is even more sharply peaked than $P(\theta \mid \tilde{\theta})$.

Pseudocode for both the decomposable and indecomposable cases is provided in Appendix C.

3. Experiments

To demonstrate the effectiveness of our approach, we present experiments aimed at verifying that the trained Fisher score can be used for Fisher forecasting and Bayesian inference. Details of the network architecture and other relevant parameters can be found in Appendix D. The code used for these experiments is available at https://github.com/suicee/FisherScoreMatching.

3.1. Toy example

To demonstrate the application of our method to decomposable models, we first consider a simple linear Gaussian model. The full model is given by $x \mid \theta \sim \mathcal{N}(\theta, \Sigma_n)$, where $\theta \in \mathbb{R}^2$ is the parameter of interest and $\Sigma_n = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$.

We decompose the model according to Equation 4, and choose the latent model covariance to be $\Sigma_{\theta^*} = 0.4 \cdot \mathbf{I}$. We then sample θ from a uniform prior over $[-3, 3]^2$, and generate θ^* and x using the conditional distributions defined by the decomposition. We generate 100,000 sample triplets for training.

For evaluation, we sample a random observation x_{obs}, and compute the Fisher score across the entire parameter space. As shown in Figure 1, the learned score field closely matches the analytical Fisher score. This verifies that our method can recover the correct Fisher score using only the score of the latent model.

3.2. Weak Lensing Example

As a more realistic demonstration, we apply our method to `jax-cosmo`[2] , a library for automatically differentiable cosmological theory computations using JAX (Campagne et al., 2023). It provides access to gradients of theoretical predictions of various cross-correlation statistics of weak lensing and galaxy observations with respect to cosmological parameters, making it well-suited for validating our approach.

In this experiment, we vary two key cosmological parameters in the standard model of cosmology: the cold dark matter density Ω_c and the matter fluctuation amplitude σ_8. For each pair (Ω_c, σ_8), we use `jax-cosmo` to simulate projected two-point galaxy lensing angular power spectra. The angular power spectra C_ℓ are computed at five logarithmically spaced multipole between $\ell = 10$ and $\ell = 1000$. All other cosmological parameters and configuration settings are fixed to their default values as defined in the package.

[2] https://github.com/DifferentiableUniverseInitiative/jax_cosmo

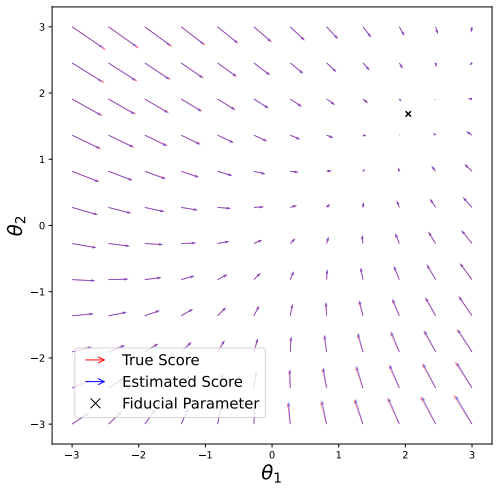

Figure 1. Comparison of the learned Fisher score (blue arrows) and the analytical Fisher score (red arrows) across the parameter space for a fixed observation. The black dot denotes the fiducial parameter value corresponding to the observation. Arrows originate at parameter locations and point in the direction of the Fisher score. The close alignment of the two sets of arrows indicates accurate recovery of the Fisher score.

We assume the data likelihood is Gaussian:

$$x \mid \theta \sim \mathcal{N}(\mu(\theta), \Sigma_{\text{fid}}), \tag{6}$$

where $x = C_\ell$, $\theta = (\Omega_c, \sigma_8)$, $\mu(\theta)$ is the mean spectrum computed by `jax-cosmo`, and Σ_{fid} is a fixed covariance matrix evaluated at a fiducial cosmology.

Under this Gaussian model and using the differentiability of `jax-cosmo`, we can compute the ground truth Fisher score and Fisher information matrix as:

$$\begin{aligned} s^*(x, \theta) &= (\nabla_\theta \mu(\theta))^\top \Sigma_{fid}^{-1}(x - \mu(\theta)), \\ \mathcal{I}^*(\theta) &= (\nabla_\theta \mu(\theta))^\top \Sigma_{fid}^{-1} (\nabla_\theta \mu(\theta)) . \end{aligned} \tag{7}$$

This model is not obviously decomposable, so we introduce an auxiliary latent variable $\tilde{\theta}$. We sample $\tilde{\theta}$ from the ranges $[0.2, 0.4]$ for Ω_c and $[0.70, 0.90]$ for σ_8. Then, we generate θ by adding Gaussian noise: $\theta \sim \mathcal{N}(\tilde{\theta}, \sigma^2 \cdot \mathbf{I})$, with $\sigma = 10^{-3}$. Once θ is sampled, we compute C_ℓ using `jax-cosmo`. We generate 100,000 sample triplets $(\tilde{\theta}, \theta, x)$ for training.

As in the toy example, we first evaluate the learned Fisher score across the parameter space for a fixed mock observation. Figure 2 shows the learned score field accurately

matches the true score directions (obtained via autodifferentiation), which are approximately orthogonal to the degeneracy direction between parameters expected from weak lensing correlation analysis (Jain & Seljak, 1997).

The training set will contain few relevant examples for parameter values that have small probability to have generated the mock observation. This leads to significant errors in the magnitude of the estimated score vectors for parameters that are implausible given the data. In spite of that, the model stably extrapolates across the entire training range to within a factor of 2. This will not affect applications such as forecasting and inference where the model encounters plausible combinations of data and parameters, as evidenced in the following experiments.

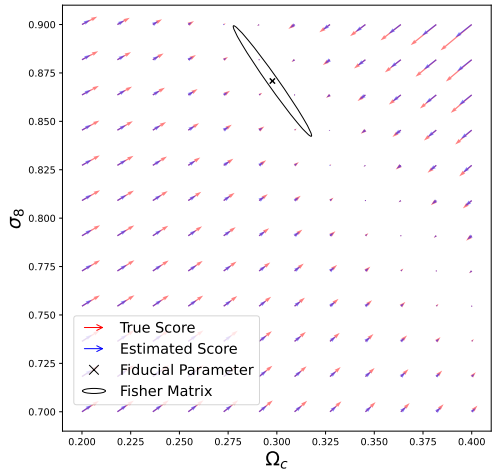

Figure 2. Extrapolation of the learned Fisher score (blue arrows) *vs* the true Fisher score (red arrows) in the weak lensing example, for a fixed observation. The black cross indicates the fiducial parameters and the Fisher matrix contours are shown for scale. For implausible parameters far away from the fiducial, the score is underestimated by $\sim 40\%$, but the direction is still accurately captured. This does not affect sampling of high likelihood regions or maximum likelihood estimation.

By supplying a prior (uniform in the ranges $[0.2, 0.4]$ for Ω_c and $[0.70, 0.90]$ for σ_8) we use HMC to sample the posterior based on the estimated score vector field. As shown in Figure 3, the posterior contours from the estimated scores (blue) align closely with those from the true scores (red), confirming that the learned scores are suitable for Bayesian inference.

As a second example, we can estimate the Fisher information matrix at any point in parameter space θ_0 by computing the covariance of the score model

$$\mathcal{I}(\theta_0) = E_{p(x|\theta_0)}\left[s(x,\theta_0)s(x,\theta_0)^\top\right], \qquad (8)$$

where the expectation is approximated using samples drawn from $p(x \mid \theta_0)$. We evaluate the Fisher matrix on a grid of

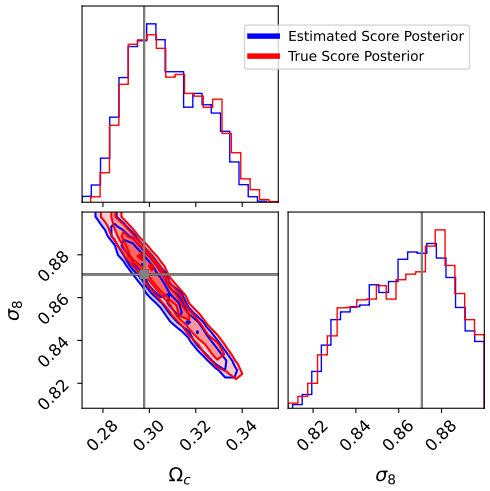

Figure 3. Posterior distribution obtained via HMC using the learned Fisher scores (blue contours) versus the true Fisher scores (red contours). The gray cross marks the fiducial parameter used to generate the mock observation.

parameter values and compare the results to those obtained from Equation 7. As shown in Figure 4, the estimated contours align closely with the true Fisher matrices.

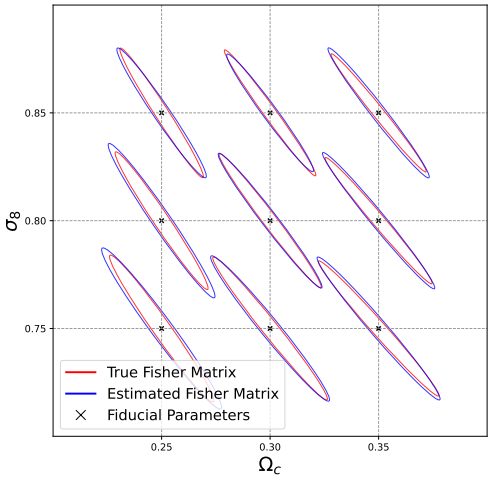

Figure 4. Comparison of estimated Fisher matrices (blue contours) and true Fisher matrices (red contours). Each contour corresponds to a Fisher matrix estimated at its center (indicated by a cross).

4. Conclusion

We propose a score-matching-based method to estimate the Fisher score only from simulations by introducing a latent model to generate the parameters of the target model.

Through a toy example and a cosmological application, we show that this approach performs well, both in scenarios

where a natural latent is suggested by the model structure (decomposable), or when it is added as a perturbation (non-decomposable). Once trained, the model can estimate the Fisher score at any given parameter-data pair. This enables gradient-based Bayesian inference (e.g., via HMC) and fast Fisher forecasting, as demonstrated in our experiments.

Modeling the Fisher score function differs conceptually and in practice from training a conditional diffusion model by separating the act of choosing the prior from the construction of the training set. When training a diffusion model the choice of training set determines the prior, whereas the Fisher score approximation that minimizes our loss function, Equation 3, is independent of the prior. In practice, successful training, of course, requires a training set of sufficient variety and size.

Several directions remain for future work. On the methodological side, large training datasets are typically required for an accurate score estimates in low-probability regions; finding strategies to reduce this would be valuable. On the application side, it would be interesting to apply this framework to high-dimensional problems, such as field-level Fisher analysis or simulation-based inference, where existing techniques face significant challenges.

References

Alsing, J. and Wandelt, B. Generalized massive optimal data compression. *MNRAS*, 476(1):L60–L64, May 2018. doi: 10.1093/mnrasl/sly029.

Alsing, J., Charnock, T., Feeney, S., and Wandelt, B. Fast likelihood-free cosmology with neural density estimators and active learning. *MNRAS*, 488(3):4440–4458, September 2019. doi: 10.1093/mnras/stz1960.

Beyond-2pt Collaboration, :, Krause, E., Kobayashi, Y., Salcedo, A. N., Ivanov, M. M., Abel, T., Akitsu, K., Angulo, R. E., Cabass, G., Contarini, S., Cuesta-Lazaro, C., Hahn, C., Hamaus, N., Jeong, D., Modi, C., Nguyen, N.-M., Nishimichi, T., Paillas, E., Pellejero Ibañez, M., Philcox, O. H. E., Pisani, A., Schmidt, F., Tanaka, S., Verza, G., Yuan, S., and Zennaro, M. A Parameter-Masked Mock Data Challenge for Beyond-Two-Point Galaxy Clustering Statistics. *arXiv e-prints*, art. arXiv:2405.02252, May 2024. doi: 10.48550/arXiv.2405.02252.

Boelts, J., Deistler, M., Gloeckler, M., Álvaro Tejero-Cantero, Lueckmann, J.-M., Moss, G., Steinbach, P., Moreau, T., Muratore, F., Linhart, J., Durkan, C., Vetter, J., Miller, B. K., Herold, M., Ziaeemehr, A., Pals, M., Gruner, T., Bischoff, S., Krouglova, N., Gao, R., Lappalainen, J. K., Mucsányi, B., Pei, F., Schulz, A., Stefanidi, Z., Rodrigues, P., Schröder, C., Zaid, F. A., Beck, J., Kapoor, J., Greenberg, D. S., Gonçalves, P. J., and Macke, J. H. sbi reloaded: a toolkit for simulation-based inference workflows. *Journal of Open Source Software*, 10(108):7754, 2025. doi: 10.21105/joss.07754. URL https://doi.org/10.21105/joss.07754.

Bradbury, J., Frostig, R., Hawkins, P., Johnson, M. J., Leary, C., Maclaurin, D., Necula, G., Paszke, A., VanderPlas, J., Wanderman-Milne, S., and Zhang, Q. JAX: composable transformations of Python+NumPy programs, 2018. URL http://github.com/jax-ml/jax.

Brehmer, J., Louppe, G., Pavez, J., and Cranmer, K. Mining gold from implicit models to improve likelihood-free inference. *Proceedings of the National Academy of Sciences*, 117(10):5242–5249, February 2020. ISSN 1091-6490. doi: 10.1073/pnas.1915980117. URL http://dx.doi.org/10.1073/pnas.1915980117.

Brinckmann, T. and Lesgourgues, J. MontePython 3: boosted MCMC sampler and other features. 2018.

Campagne, J.-E., Lanusse, F., Zuntz, J., Boucaud, A., Casas, S., Karamanis, M., Kirkby, D., Lanzieri, D., Peel, A., and Li, Y. JAX-COSMO: An End-to-End Differentiable and GPU Accelerated Cosmology Library. *The Open Journal of Astrophysics*, 6:15, April 2023. doi: 10.21105/astro.2302.05163.

Charnock, T., Lavaux, G., and Wandelt, B. D. Automatic physical inference with information maximizing neural networks. *Phys. Rev. D*, 97(8):083004, April 2018. doi: 10.1103/PhysRevD.97.083004.

Cheng, S., Ting, Y.-S., Ménard, B., and Bruna, J. A new approach to observational cosmology using the scattering transform. *MNRAS*, 499(4):5902–5914, December 2020. doi: 10.1093/mnras/staa3165.

Cranmer, K., Brehmer, J., and Louppe, G. The frontier of simulation-based inference. *Proceedings of the National Academy of Science*, 117(48):30055–30062, December 2020. doi: 10.1073/pnas.1912789117.

Cuesta-Lazaro, C. and Mishra-Sharma, S. Point cloud approach to generative modeling for galaxy surveys at the field level. *Phys. Rev. D*, 109(12):123531, June 2024. doi: 10.1103/PhysRevD.109.123531.

DES Collaboration, Abbott, T. M. C., Aguena, M., Alarcon, A., Anbajagane, D., Andrade-Oliveira, F., Avila, S., Bacon, D., Becker, M. R., Bhargava, S., Blazek, J., Bocquet, S., Brooks, D., Carnero Rosell, A., Carretero, J., Castander, F. J., Chang, C., Choi, A., Conselice, C., Costanzi, M., Crocce, M., da Costa, L. N., Pereira, M. E. S., Davis, T. M., Desai, S., Diehl, H. T., Dodelson, S., Doel, P., Elvin-Poole, J., Esteves, J., Everett, S., Farahi, A., Ferté, A., Flaugher, B., García-Bellido, J., Gatti, M., Giannini,

G., Giles, P., Grandis, S., Gruen, D., Gruendl, R. A., Gutierrez, G., Harrison, I., Hinton, S. R., Hollowood, D. L., Honscheid, K., Jeffrey, N., Jeltema, T., Krause, E., Lahav, O., Lee, S., Lidman, C., Lima, M., Lin, H., Mohr, J. J., Marshall, J. L., McCullough, J., Mena-Fern, J., Miquel, R., Muir, J., Myles, J., Ogando, R. L. C., Palmese, A., Paterno, M., Plazas Malagón, A. A., Porredon, A., Prat, J., Romer, A. K., Roodman, A., Rozo, E., Rykoff, E. S., Sanchez, E., Sanchez Cid, D., Sevilla-Noarbe, I., Smith, M., Suchyta, E., Tarle, G., Thomas, D., To, C.-H., Troxel, M. A., Vikram, V., Walker, A. R., Weinberg, D. H., Weaverdyck, N., Wechsler, R. H., Weller, J., Wu, H. Y., Yamamoto, M., Yanny, B., Zhang, Y., and Zhou, C. Dark Energy Survey Year 3 Results: Cosmological Constraints from Cluster Abundances, Weak Lensing, and Galaxy Clustering. *arXiv e-prints*, art. arXiv:2503.13632, March 2025. doi: 10.48550/arXiv.2503.13632.

Euclid Collaboration, Congedo, G., Miller, L., Taylor, A. N., Cross, N., Duncan, C. A. J., Kitching, T., Martinet, N., Matthew, S., Schrabback, T., Tewes, M., Welikala, N., Aghanim, N., Amara, A., Andreon, S., Auricchio, N., Baldi, M., Bardelli, S., Bender, R., Bodendorf, C., Bonino, D., Branchini, E., Brescia, M., Brinchmann, J., Camera, S., Capobianco, V., Carbone, C., Cardone, V. F., Carretero, J., Casas, S., Castander, F. J., Castellano, M., Cavuoti, S., Cimatti, A., Conselice, C. J., Conversi, L., Copin, Y., Courbin, F., Courtois, H. M., Cropper, M., Da Silva, A., Degaudenzi, H., Di Giorgio, A. M., Dinis, J., Dubath, F., Dupac, X., Farina, M., Farrens, S., Ferriol, S., Fosalba, P., Frailis, M., Franceschi, E., Galeotta, S., Garilli, B., Gillis, B., Giocoli, C., Grazian, A., Grupp, F., Haugan, S. V. H., Holliman, M. S., Holmes, W., Hormuth, F., Hornstrup, A., Hudelot, P., Jahnke, K., Keihänen, E., Kermiche, S., Kiessling, A., Kilbinger, M., Kubik, B., Kuijken, K., Kümmel, M., Kunz, M., Kurki-Suonio, H., Ligori, S., Lilje, P. B., Lindholm, V., Lloro, I., Maino, D., Maiorano, E., Mansutti, O., Marggraf, O., Markovic, K., Marulli, F., Massey, R., Maurogordato, S., McCracken, H. J., Medinaceli, E., Mei, S., Melchior, M., Meneghetti, M., Merlin, E., Meylan, G., Moresco, M., Morin, B., Moscardini, L., Munari, E., Niemi, S. M., Nightingale, J. W., Padilla, C., Paltani, S., Pasian, F., Pedersen, K., Percival, W. J., Pettorino, V., Pires, S., Polenta, G., Poncet, M., Popa, L. A., Pozzetti, L., Raison, F., Rebolo, R., Renzi, A., Rhodes, J., Riccio, G., Romelli, E., Roncarelli, M., Rossetti, E., Saglia, R., Sapone, D., Sartoris, B., Schneider, P., Secroun, A., Seidel, G., Serrano, S., Sirignano, C., Sirri, G., Stanco, L., Tallada-Crespí, P., Tavagnacco, D., Tereno, I., Toledo-Moreo, R., Torradeflot, F., Tutusaus, I., Valentijn, E. A., Valenziano, L., Vassallo, T., Veropalumbo, A., Wang, Y., Weller, J., Zamorani, G., Zoubian, J., Zucca, E., Biviano, A., Bolzonella, M., Boucaud, A., Bozzo, E., Burigana, C., Colodro-Conde, C., Di Ferdinando, D., Graciá-Carpio, J., Mauri, N., Neissner, C., Nucita, A. A., Sakr, Z., Scottez, V., Tenti, M., Viel, M., Wiesmann, M., Akrami, Y., Allevato, V., Anselmi, S., Baccigalupi, C., Ballardini, M., Borgani, S., Borlaff, A. S., Bruton, S., Cabanac, R., Cappi, A., Carvalho, C. S., Castignani, G., Castro, T., Cañas-Herrera, G., Chambers, K. C., Cooray, A. R., Coupon, J., Davini, S., De Lucia, G., Desprez, G., Di Domizio, S., Dole, H., Díaz-Sánchez, A., Escartin Vigo, J. A., Escoffier, S., Ferrero, I., Finelli, F., Gabarra, L., García-Bellido, J., Gaztanaga, E., Giacomini, F., Gozaliasl, G., Guinet, D., Hall, A., Hildebrandt, H., Ilić, S., Jimenez Muñoz, A., Joudaki, S., Kajava, J. J. E., Kansal, V., and Karagiannis, D. Euclid preparation: LIII. LensMC, weak lensing cosmic shear measurement with forward modelling and Markov Chain Monte Carlo sampling. *A&A*, 691:A319, November 2024. doi: 10.1051/0004-6361/202450617.

Heavens, A. F., Sellentin, E., de Mijolla, D., and Vianello, A. Massive data compression for parameter-dependent covariance matrices. *MNRAS*, 472(4):4244–4250, December 2017. doi: 10.1093/mnras/stx2326.

Ho, M., Bartlett, D. J., Chartier, N., Cuesta-Lazaro, C., Ding, S., Lapel, A., Lemos, P., Lovell, C. C., Makinen, T. L., Modi, C., Pandya, V., Pandey, S., Perez, L. A., Wandelt, B., and Bryan, G. L. LtU-ILI: An All-in-One Framework for Implicit Inference in Astrophysics and Cosmology. *The Open Journal of Astrophysics*, 7:54, July 2024. doi: 10.33232/001c.120559.

Jain, B. and Seljak, U. Cosmological Model Predictions for Weak Lensing: Linear and Nonlinear Regimes. *ApJ*, 484 (2):560–573, July 1997. doi: 10.1086/304372.

Lemos, P., Parker, L. H., Hahn, C., Ho, S., Eickenberg, M., Hou, J., Massara, E., Modi, C., Moradinezhad Dizgah, A., Régaldo-Saint Blancard, B., and Spergel, D. SimBIG: Field-level Simulation-based Inference of Large-scale Structure. In *Machine Learning for Astrophysics*, pp. 18, July 2023. doi: 10.48550/arXiv.2310.15256.

Li, Y., Lu, L., Modi, C., Jamieson, D., Zhang, Y., Feng, Y., Zhou, W., Pok Kwan, N., Lanusse, F., and Greengard, L. pmwd: A Differentiable Cosmological Particle-Mesh N-body Library. *arXiv e-prints*, art. arXiv:2211.09958, November 2022. doi: 10.48550/arXiv.2211.09958.

Makinen, T. L., Sui, C., Wandelt, B. D., Porqueres, N., and Heavens, A. Hybrid Summary Statistics. *arXiv e-prints*, art. arXiv:2410.07548, October 2024. doi: 10.48550/arXiv.2410.07548.

Mudur, N., Cuesta-Lazaro, C., and Finkbeiner, D. P. Diffusion-HMC: Parameter Inference with Diffusion-model-driven Hamiltonian Monte Carlo. *ApJ*, 978(1): 64, January 2025. doi: 10.3847/1538-4357/ad8bc3.

Planck Collaboration, Aghanim, N., Akrami, Y., Ashdown, M., Aumont, J., Baccigalupi, C., Ballardini, M., Banday, A. J., Barreiro, R. B., Bartolo, N., Basak, S., Battye, R., Benabed, K., Bernard, J. P., Bersanelli, M., Bielewicz, P., Bock, J. J., Bond, J. R., Borrill, J., Bouchet, F. R., Boulanger, F., Bucher, M., Burigana, C., Butler, R. C., Calabrese, E., Cardoso, J. F., Carron, J., Challinor, A., Chiang, H. C., Chluba, J., Colombo, L. P. L., Combet, C., Contreras, D., Crill, B. P., Cuttaia, F., de Bernardis, P., de Zotti, G., Delabrouille, J., Delouis, J. M., Di Valentino, E., Diego, J. M., Doré, O., Douspis, M., Ducout, A., Dupac, X., Dusini, S., Efstathiou, G., Elsner, F., Enßlin, T. A., Eriksen, H. K., Fantaye, Y., Farhang, M., Fergusson, J., Fernandez-Cobos, R., Finelli, F., Forastieri, F., Frailis, M., Fraisse, A. A., Franceschi, E., Frolov, A., Galeotta, S., Galli, S., Ganga, K., Génova-Santos, R. T., Gerbino, M., Ghosh, T., González-Nuevo, J., Górski, K. M., Gratton, S., Gruppuso, A., Gudmundsson, J. E., Hamann, J., Handley, W., Hansen, F. K., Herranz, D., Hildebrandt, S. R., Hivon, E., Huang, Z., Jaffe, A. H., Jones, W. C., Karakci, A., Keihänen, E., Keskitalo, R., Kiiveri, K., Kim, J., Kisner, T. S., Knox, L., Krachmalnicoff, N., Kunz, M., Kurki-Suonio, H., Lagache, G., Lamarre, J. M., Lasenby, A., Lattanzi, M., Lawrence, C. R., Le Jeune, M., Lemos, P., Lesgourgues, J., Levrier, F., Lewis, A., Liguori, M., Lilje, P. B., Lilley, M., Lindholm, V., López-Caniego, M., Lubin, P. M., Ma, Y. Z., Macías-Pérez, J. F., Maggio, G., Maino, D., Mandolesi, N., Mangilli, A., Marcos-Caballero, A., Maris, M., Martin, P. G., Martinelli, M., Martínez-González, E., Matarrese, S., Mauri, N., McEwen, J. D., Meinhold, P. R., Melchiorri, A., Mennella, A., Migliaccio, M., Millea, M., Mitra, S., Miville-Deschênes, M. A., Molinari, D., Montier, L., Morgante, G., Moss, A., Natoli, P., Nørgaard-Nielsen, H. U., Pagano, L., Paoletti, D., Partridge, B., Patanchon, G., Peiris, H. V., Perrotta, F., Pettorino, V., Piacentini, F., Polastri, L., Polenta, G., Puget, J. L., Rachen, J. P., Reinecke, M., Remazeilles, M., Renzi, A., Rocha, G., Rosset, C., Roudier, G., Rubiño-Martín, J. A., Ruiz-Granados, B., Salvati, L., Sandri, M., Savelainen, M., Scott, D., Shellard, E. P. S., Sirignano, C., Sirri, G., Spencer, L. D., Sunyaev, R., Suur-Uski, A. S., Tauber, J. A., Tavagnacco, D., Tenti, M., Toffolatti, L., Tomasi, M., Trombetti, T., Valenziano, L., Valiviita, J., Van Tent, B., Vibert, L., Vielva, P., Villa, F., Vittorio, N., Wandelt, B. D., Wehus, I. K., White, M., White, S. D. M., Zacchei, A., and Zonca, A. Planck 2018 results. VI. Cosmological parameters. *A&A*, 641:A6, September 2020. doi: 10.1051/0004-6361/201833910.

Sharrock, L., Simons, J., Liu, S., and Beaumont, M. Sequential Neural Score Estimation: Likelihood-Free Inference with Conditional Score Based Diffusion Models. *arXiv e-prints*, art. arXiv:2210.04872, October 2022. doi: 10.48550/arXiv.2210.04872.

Song, Y. and Ermon, S. Generative Modeling by Estimating Gradients of the Data Distribution. *arXiv e-prints*, art. arXiv:1907.05600, July 2019. doi: 10.48550/arXiv.1907.05600.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-Based Generative Modeling through Stochastic Differential Equations. *arXiv e-prints*, art. arXiv:2011.13456, November 2020. doi: 10.48550/arXiv.2011.13456.

Square Kilometre Array Cosmology Science Working Group, Bacon, D. J., Battye, R. A., Bull, P., Camera, S., Ferreira, P. G., Harrison, I., Parkinson, D., Pourtsidou, A., Santos, M. G., Wolz, L., Abdalla, F., Akrami, Y., Alonso, D., Andrianomena, S., Ballardini, M., Bernal, J. L., Bertacca, D., Bengaly, C. A. P., Bonaldi, A., Bonvin, C., Brown, M. L., Chapman, E., Chen, S., Chen, X., Cunnington, S., Davis, T. M., Dickinson, C., Fonseca, J., Grainge, K., Harper, S., Jarvis, M. J., Maartens, R., Maddox, N., Padmanabhan, H., Pritchard, J. R., Raccanelli, A., Rivi, M., Roychowdhury, S., Sahlén, M., Schwarz, D. J., Siewert, T. M., Viel, M., Villaescusa-Navarro, F., Xu, Y., Yamauchi, D., and Zuntz, J. Cosmology with Phase 1 of the Square Kilometre Array Red Book 2018: Technical specifications and performance forecasts. *PASA*, 37:e007, March 2020. doi: 10.1017/pasa.2019.51.

Tegmark, M., Taylor, A. N., and Heavens, A. F. Karhunen-Loève Eigenvalue Problems in Cosmology: How Should We Tackle Large Data Sets? *ApJ*, 480(1):22–35, May 1997. doi: 10.1086/303939.

Torrado, J. and Lewis, A. Cobaya: code for Bayesian analysis of hierarchical physical models. *J. Cosmology Astropart. Phys.*, 2021(5):057, May 2021. doi: 10.1088/1475-7516/2021/05/057.

Zhao, X., Mao, Y., Cheng, C., and Wandelt, B. D. Simulation-based Inference of Reionization Parameters from 3D Tomographic 21 cm Light-cone Images. *ApJ*, 926(2):151, February 2022. doi: 10.3847/1538-4357/ac457d.

A. Bayesian Least Squares Estimation

Assume $(x, y) \sim P(x, y)$, and let $g(x)$ be a target function. To estimate $g(x)$ from observations y, we consider minimizing the mean squared error:

$$\mathcal{L}(f) = \mathbb{E}_{P(x,y)} \left[(f(y) - g(x))^2 \right].$$

The minimizer of this loss is the Bayes least squares estimator:

$$f^*(y) = \mathbb{E}_{P(x|y)}[g(x)].$$

This follows from the decomposition:

$$\begin{aligned} \mathcal{L}(f) &= \mathbb{E}_{P(y)} \left[\mathbb{E}_{P(x|y)} \left[(f(y) - g(x))^2 \right] \right] \\ &= \mathbb{E}_{P(y)} \left[\left(f(y) - \mathbb{E}_{P(x|y)}[g(x)] \right)^2 + \mathrm{Var}_{P(x|y)}[g(x)] \right], \end{aligned}$$

where $\mathrm{Var}_{P(x|y)}[g(x)] = \mathbb{E}_{P(x|y)} \left[(g(x) - \mathbb{E}_{P(x|y)}[g(x)])^2 \right]$ is independent of f. Thus, the optimal estimator matches the conditional mean.

B. Error Analysis for Indecomposable Models with Additive Noise

When a natural latent variable is unavailable, we introduce an auxiliary variable $\tilde{\theta}$ and define a latent model $P(\theta \mid \tilde{\theta})$. In this case, our method estimates the Fisher score of the extended model, $\nabla_{\tilde{\theta}} \log P(x \mid \tilde{\theta})$, rather than the original model.

Assume an additive noise model:

$$\theta = \tilde{\theta} + w, \quad \text{where } w \sim p(w).$$

Then the marginal likelihood under the extended model is:

$$P(x \mid \tilde{\theta}) = \int P(x \mid \theta)\, p(\theta \mid \tilde{\theta})\, d\theta = \int P(x \mid \tilde{\theta} + w)\, p(w)\, dw.$$

Taking the gradient with respect to $\tilde{\theta}$, we have:

$$\begin{aligned} \nabla_{\tilde{\theta}} P(x \mid \tilde{\theta}) &= \int \nabla_{\tilde{\theta}} P(x \mid \tilde{\theta} + w)\, p(w)\, dw \\ &= \int \nabla_{\tilde{\theta}} \log P(x \mid \tilde{\theta} + w)\, P(x \mid \tilde{\theta} + w)\, p(w)\, dw. \end{aligned}$$

Using the definition of the Fisher score, we obtain:

$$\nabla_{\tilde{\theta}} \log P(x \mid \tilde{\theta}) = \frac{\nabla_{\tilde{\theta}} P(x \mid \tilde{\theta})}{P(x \mid \tilde{\theta})} = \int \nabla_{\tilde{\theta}} \log P(x \mid \tilde{\theta} + w)\, \frac{P(x \mid \tilde{\theta} + w)\, p(w)}{P(x \mid \tilde{\theta})}\, dw.$$

Substituting $\theta = \tilde{\theta} + w$, and applying Bayes' theorem, we rewrite the integrand as

$$\nabla_{\tilde{\theta}} \log P(x \mid \tilde{\theta}) = \int \nabla_{\theta} \log P(x \mid \theta)\, P(\theta \mid \tilde{\theta}, x)\, d\theta.$$

This result shows that the estimated score is a smoothed version of the true score $\nabla_\theta \log P(x \mid \theta)$. The accuracy of this approximation depends on the sharpness of $P(\theta \mid \tilde{\theta})$, which bounds the sharpness of $P(\theta \mid \tilde{\theta}, x)$. When it is highly concentrated (e.g., close to a delta function), the smoothed score approaches the true Fisher score.

C. Fisher Score Matching Algorithms

Algorithm 1 FSM for Decomposable Model

Input: Simulator $x \mid \theta \sim \int p(x \mid \theta^*)p(\theta^* \mid \theta)d\theta^*$, model $s(x, \theta)$, proposal $p(\theta)$
while not converged **do**
  Sample batch: $(\theta, \theta^*, x) \sim p(\theta)p(\theta^* \mid \theta)p(x \mid \theta^*)$
  Compute target: $\nabla_\theta \log p(\theta^* \mid \theta)$
  Predict score: $\hat{s} = s(x, \theta)$
  Compute loss: $L = \|\hat{s} - \text{target}\|^2$
  Update model parameters using ∇L
end while
Output: Trained estimator $s(x, \theta) \approx \nabla_\theta \log p(x \mid \theta)$

Algorithm 2 FSM for Indecomposable Model

Input: Simulator $x \mid \theta \sim p(x \mid \theta)$, model $s(x, \tilde{\theta})$, proposal $p(\tilde{\theta})$, conditional $p(\theta \mid \tilde{\theta})$
while not converged **do**
  Sample batch: $(\tilde{\theta}, \theta, x) \sim p(\tilde{\theta})p(\theta \mid \tilde{\theta})p(x \mid \theta)$
  Compute target: $\nabla_{\tilde{\theta}} \log p(\theta \mid \tilde{\theta})$
  Predict score: $\hat{s} = s(x, \tilde{\theta})$
  Compute loss: $L = \|\hat{s} - \text{target}\|^2$
  Update model parameters using ∇L
end while
Output: Trained estimator $s(x, \tilde{\theta}) \approx \nabla_\theta \log p(x \mid \tilde{\theta})$

D. Network Details

For the Gaussian model, we use a multilayer perceptron (MLP) with two hidden layers of 64 units each, followed by ELU activation. The network is trained with a learning rate of 1e-2 for 1000 epochs. For the Weak Lensing example, we use an MLP with hidden layers of sizes 64, 128, and 128, also followed by ELU activation. The network is trained with a learning rate of 1e-3 for 10,000 epochs.