University of Washington
Survey Science Group

Machine Learning for Massive Scale Cosmology

Scott F. Daniel, Andrew J. Connolly (University of Washington), Jeff Schneider (Carnegie Mellon University)

We use Gaussian processes to improve the efficiency with which an MCMC-like code explores a parameter space. Instead of integrating over the interior of the high-likelihood region of a theory, we attempt to sample only points on the boundary of the 95% confidence limit region. We also use the theoretical uncertainty in Gaussian processes to monitor poorly-explored regions of parameter space and guard against the possibility that there are multiple high-likelihood regions allowed by the theory. We test our data on the 7 year data release of the WMAP satellite, which measures the temperature fluctuations in the cosmic microwave background. We compare our method to the results of the traditional Monte Carlo Markov chain code CosmoMC (Lewis and Bridle, 2002). We find that our code produces comparable 95% confidence limits on cosmological parameters while returning greater assurance that there are no points of interest in unexplored regions of the parameter space.

I. Background – MCMC and its shortcomings

Physicists explore large parameter spaces using Monte Carlo Markov chains (MCMCs). MCMCs randomly walk through parameter space. At each new step, they compare the likelihood value $\mathcal{L}$ at the proposed step to the likelihood value at the old step and make the step with probability

$$\min\{1, \mathcal{L}_{new}/\mathcal{L}_{old}\}$$

A histogram of how long the walk spent at each point in parameter space approximates the Bayesian posterior distribution of the parameters' true values. **There are several drawbacks to this approach.**

1) If the model poorly fits the data, MCMC will still return a **posterior distribution centered on the least bad region of parameter space**.
2) **If there are multiple high-likelihood regions, MCMC may not find them all.** Once MCMC has found one high-likelihood region, it is unlikely to leave it.
3) **MCMC is inefficient**. To find the boundary of the 95% confidence limit region, you must spend time integrating the highest likelihood points, which are uninteresting.

II. Our algorithm

Instead of integrating over parameter space, we implement a method initially proposed by Bryan *et al*. (2007) to selectively evaluate points in parameter space which we expect to lie on the 95% confidence limit boundary. Our algorithm:

1) Generate a small random sample of points uniformly distributed through the parameter space. Find $\chi^2 = -2\ln(\mathcal{L})$ at these points.
2) Generate N>>1 candidate points in the parameter space. Use Gaussian processes to predict their χ^2 values. The predicted value is μ. The uncertainty in the prediction is σ.
3) Evaluate the actual χ^2 value for the candidate point with the maximum value of

$$S = 1.96\,\sigma - |\mu - \chi^2_{lim}|$$

χ^2_{lim} is the value of χ^2 corresponding to the 95% confidence limit for the data set.

4) Repeat until you feel that the parameter space has been adequately sampled.

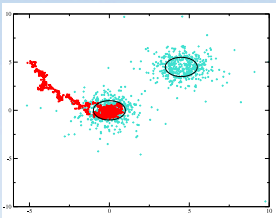
The use of σ in S drives the algorithm towards poorly explored regions of parameter space. The use of $|\mu - \chi^2_{lim}|$ drives the code towards the boundary of the high likelihood region, not its interior. As more points are sampled, the Gaussian process will become more accurate. **The Gaussian process is generally less expensive than evaluating the actual χ^2.**

1,000 points drawn from a toy two-dimensional likelihood function with two high-likelihood regions. The black circles are the boundaries of the high-likelihood regions. Blue crosses are chosen by our algorithm. Red x's are chosen by MCMC.

Our algorithm clusters at the boundary of the likelihood regions, rather than the interior.

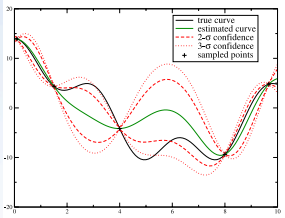
Our algorithm finds both regions.

Our algorithm samples points far from the high-likelihood region.

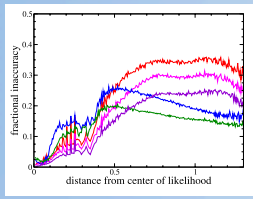


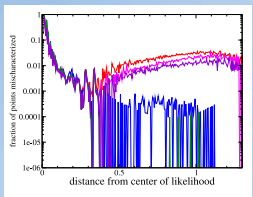
III. Gaussian Processes

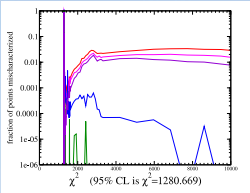
A Gaussian process is a way of taking sparsely sampled data from a non-linear function in N dimensions and turning it into a prediction and confidence estimate for the values of the function in the un-sampled space (see Chapter 2 of Rasmussen and Williams 2006).

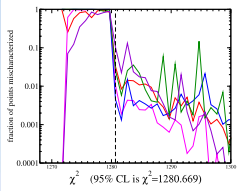


A 1-dimensional example of a Gaussian process









IV. Testing our method: performance

We use the temperature power spectrum from the 7 year release of the WMAP cosmic microwave background experiment (Jarosik *et al*., 2011) and explore the six dimensional parameter space characterizing the spatially flat concordance cosmology. **We simultaneously run four independent MCMCs on the same data**. We find χ^2 for a set of 3 million test points in parameter space. We use the output both from our code and from all four MCMCs combined in a Gaussian process to predict the values of χ^2 at the test points. The plots to the left show both the fractional inaccuracy

$$|(\chi^2 - \mu)/\chi^2|$$

and the fraction of points mischaracterized (points for which $\chi^2 < \chi^2_{lim}$ but the Gaussian process predicts $\mu > \chi^2_{lim}$ and vice versa) both as a function of the true χ^2 and the distance in parameter space from the center of the high-likelihood region.

Green curves represent our code as described in Section II

Blue curves represent a version of our code which has been modified to linger near points within 10% of χ^2_{lim}.

Both iterations of our code are plotted after they have sampled 40,000 points in parameter space.

Red curves represent MCMC after they have sampled a combined 40,000 points.

Magenta curves represent MCMC after they have sampled a combined 160,000 points.

Violet curves represent MCMC after they have sampled a combined 640,000 points.

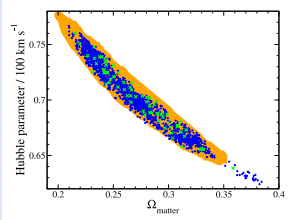
V. Results: performance

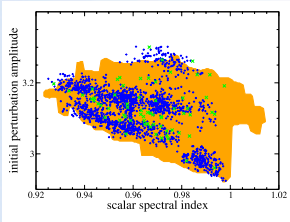
Neither MCMC nor our code yield enough information to accurately predict the values of χ^2 inside the confidence limit.

MCMC micharacterizes a few percent of points outside the confidence limit, even in the region where $\chi^2 > 2\,\chi^2_{lim}$. **It is difficult to be confident that the regions of parameter space ignored by MCMC are actually uninteresting.**

Our code is much better at characterizing points far from the high likelihood region. **We can be more confident that our algorithm has found all of the high likelihood regions in parameter space.**

This confidence is achieved even **though our code has sampled a factor of 16 fewer points than MCMC.** This is important for cases where evaluating the likelihood function is computationally expensive.





VI. Results: parameter constraints

We also consider the parameter constraints derived from our code. In the above plots, we show two instances of the constraints in a reduced, two-dimensional parameter space.

Orange regions are the marginalized Bayesian 95% confidence regions yielded by MCMC after a combined sampling of 640,000 points.

Green points are the good ($\chi^2 < \chi^2_{lim}$) points found by our algorithm.

Blue points are the good points found by our algorithm with the lingering modification described in Section IV.

Our code explores basically the same regions of parameter space as MCMC, but **finds good points spuriously excluded by MCMC**.

The table below shows the single parameter 95% confidence limit constraints derived by our code and MCMC. Recall that **our code has sampled a factor of 16 fewer points** in deriving these constraints.

	our model	our model (modified)	MCMC
Baryon density parameter	0.0208 < x < 0.0235	0.0205 < x < 0.0236	0.021 < x < 0.0235
Dark matter density parameter	0.100 < x < 0.125	0.0982 < x < 0.130	0.0987 < x < 0.123
Hubble parameter	0.639 < x < 0.760	0.624 < x < 0.767	0.649 < x < 0.762
Spectral index	0.927 < x < 0.998	0.925 < x < 0.998	0.934 < x < 1.00
Normalization	2.96 < x < 3.30	2.92 < x < 3.30	2.99 < x < 3.25

References:

Bryan, Brent *et al.*, The Astrophysical Journal **665**, 25 (2007)
Jarosik, N. *et al.*, The Astrophysical Journal Supplement Series **192**, 14 (2011)
Lewis, A. and Bridle, S., Physical Review D **66**, 103511 (2002)
Rasmussen, C.E. and Williams, C. K. I. "Processes for Machine Learning" (2006) www.GaussianProcess.org/gpml

We gratefully acknowledge support from the DOE Applied Mathematics Program through grant DE-FG02-87ER40315 and the NSF through grant IIS-0844580