

# Accelerated Parameter Estimation with DALE $\chi$

Scott F. Daniel<sup>1</sup> and Eric V. Linder<sup>2,3</sup>

<sup>1</sup>Department of Astronomy, University of Washington, Seattle, WA

<sup>2</sup>Berkeley Center for Cosmological Physics & Berkeley Lab, University of California, Berkeley, CA 94720

<sup>3</sup>Energetic Cosmos Laboratory, Nazarbayev University, Astana, Kazakhstan 010000

May 8, 2017

## ABSTRACT

We consider methods for improving the estimation of constraints on a high-dimensional parameter space with a computationally expensive likelihood function. In such cases Markov chain Monte Carlo (MCMC) can take a long time to converge and concentrates on finding the maxima rather than the often-desired confidence contours for accurate error estimation. We employ DALE $\chi$  (Direct Analysis of Limits via the Exterior of  $\chi^2$ ) for determining confidence contours by minimizing a cost function parametrized to incentivize points in parameter space which are both on the confidence limit and far from previously sampled points. We compare DALE $\chi$  to the nested sampling algorithm implemented in MultiNest on a toy likelihood function that is highly non-Gaussian and non-linear in the mapping between parameter values and  $\chi^2$ . We find that in high-dimensional cases DALE $\chi$  finds the same confidence limit as MultiNest using roughly an order of magnitude fewer evaluations of the likelihood function. DALE $\chi$  is open-source and available at <https://github.com/danielsf/Dalex.git>.

## 1. Introduction

When confronting a model with data, determining the best fit parameters is only one element of the result. For informative comparison one must have a robust estimation of the confidence intervals for the parameters, and indeed the full multidimensional confidence contours. This will become increasingly important with future big data experiments where as we strive to achieve precision and accurate cosmology we 1) need to employ numerous foreground, calibration, or other systematics parameters, raising the dimensionality of the fitting space, and 2) combine data from multiple probes and experiments, requiring knowledge of the full multidimensional posterior probability distribution.

Thus, one wants to *optimize* the likelihood function on some parameter space and also *characterize* the behavior of the likelihood function in the parameter space region surrounding that optimum. For a frequentist confidence limit (see the appendix of Daniel *et al.* (2014)), this means we wish to find all combinations of parameter values that result in  $\chi^2 \leq \chi_{\text{lim}}^2$  where  $\chi_{\text{lim}}^2$  is set by

theoretical considerations (e.g. a certain  $\Delta\chi^2$  above  $\chi_{\min}^2$ ). For a Bayesian credible limit, we wish to find the region of parameter space that maximizes the posterior probability distribution and then integrate that distribution over parameter space until we have a contour that contains some percentage of the total posterior probability.

Numerous techniques exist for carrying out this process. The most direct is evaluating the likelihood throughout the parameter space, but a grid-based approach rapidly becomes intractable as the dimension of the space increases. Sampling methods select various points at which to carry out an evaluation, and move through parameter space according to some algorithm. The most common such technique in cosmological applications is Markov chain Monte Carlo (MCMC). This works moderately well (though increasingly slowly) as the dimensionality increases but focuses on finding the maximum of the posterior probability; the confidence contours are a byproduct estimated by integrating the distribution of sampled points over the parameter space.

Often we are most interested in the joint confidence intervals for the parameters, and MCMC can sometimes be slow at converging on these since the algorithm is more interested in the interior of the contour rather than its boundary. Moreover, MCMC slows down significantly when confronted with highly degenerate likelihood surfaces. Other sampling techniques such as nested sampling, with MultiNest (Feroz and Hobson 2008; Feroz *et al.* 2009, 2013) an implementation popular for cosmological applications, can alleviate those issues, but still do not focus on the confidence interval.

One alternative to these sampling algorithms is Active Parameter Search. Active Parameter Search (APS) is designed specifically to go after confidence intervals (Bryan *et al.* 2007; Bryan 2007; Daniel *et al.* 2014). APS uses a combination of function optimization and Gaussian Processes (Rasmussen and Williams 2006) to specifically sample points on the  $\chi^2 = \chi_{\lim}^2$  confidence limit. In this work, we adopt the philosophy of APS (targeted sampling of  $\chi^2 = \chi_{\lim}^2$  points), but increase its robustness and speed of convergence by several improvements and new, focused search strategies. These are implemented using a Nelder-Mead simplex (Nelder and Mead 1965) to minimize a cost function of the form

$$F(\vec{\theta}) = \chi^2(\vec{\theta}) - N(\vec{\theta}) \times E(\chi^2) \times (\chi_{\lim}^2 - \chi_{\min}^2) \quad (1)$$

$$E(\chi^2(\vec{\theta})) = \begin{cases} \exp\left[\left(\chi_{\lim}^2 - \chi^2(\vec{\theta})\right)/\ell\right], & \text{if } \chi^2(\vec{\theta}) > \chi_{\lim}^2 \\ 1, & \text{otherwise} \end{cases}$$

where  $\vec{\theta}$  is the position in parameter space,  $N(\vec{\theta})$  is the neighbor function, and  $E(\chi^2(\vec{\theta}))$  is the exterior function. The parameter  $\ell$  controls how much optimizations of  $F$  are allowed to explore outside of  $\chi^2 \leq \chi_{\lim}^2$ , and  $N(\vec{\theta})$  is given by the harmonic mean of the parameter space distances between  $\vec{\theta}$  and the previously-sampled points such that  $\chi^2 \leq \chi_{\lim}^2$ . Minimizing this function drives the Nelder-Mead simplex to discover points that are both within the confidence contour and far from previously sampled points, increasing the likelihood that our algorithm will discover previously unknown regions tracing the true confidence contour. We refer to this algorithm as DALEX (Direct Analysis of Limits via the Exterior of  $\chi^2$ ). We find that in problems with complicated, high-

dimension likelihood surfaces, DALE $\chi$  achieves convergent, accurate parameter estimation more rapidly than MCMC and MultiNest.

In Sec. 2 we briefly review the methods employed by MCMC, MultiNest, and conventional APS. We introduce and discuss the strategies of DALE $\chi$  in Sec. 3, with their motivations and implementations. Section 4 introduces the cartoon likelihood function, incorporating aspects of high-dimensionality, non-Gaussianity, and degeneracy, on which we test DALE $\chi$ . Section 5 carries out the comparison of parameter estimation between the techniques. We conclude in Sec. 6.

Open-source C++ code implementing DALE $\chi$  is available at the GitHub repository <https://github.com/danielsf/Dalex.git>. This software depends on the publicly available software libraries BLAS, LAPACK (available from [www.netlib.org](http://www.netlib.org)), and ARPACK (available from [www.caam.rice.edu/software/ARPACK](http://www.caam.rice.edu/software/ARPACK)). DALE $\chi$  also relies on the KD-Tree algorithm described by Bentley (1975), which we use to store the history of parameter-space points explored by DALE $\chi$ , and the random number generator described by Marsaglia (2003). Those interested in using DALE $\chi$  for their own research should not hesitate to either contact the authors at [danielsf@astro.washington.edu](mailto:danielsf@astro.washington.edu) or open an issue on the GitHub repository should they have any questions.

## 2. Parameter Estimation Methods

### 2.1. Markov Chain Monte Carlo

MCMC uses chains of samples to evaluate the posterior probability of the data having a good fit to a point in parameter space, calculated through a likelihood function and priors on the parameter distribution. In the commonly used Metropolis-Hastings algorithm, a point is evaluated to check if it has a higher probability than the previous point, and accepted or rejected according to a probability criterion. The Markov chain part of the algorithm indicates that only the preceding point affects the next point, while the Monte Carlo part denotes the random selection of candidate points and the integration of the probability over the parameter space to determine the confidence contours. Note that MCMC per se is only trying to find the maximum; points defining the confidence contours are in some sense incidental.

DALE $\chi$  as discussed below is designed to guide the selection of desirable points by using all the information already collected, rather than only the previous point, taking a more global view, and also to accelerate the robust estimation of the confidence intervals by focusing on points informing this result and tracing contours rather than concentrating points near the probability maximum.

## 2.2. Nested Sampling (MultiNest)

Nested sampling (Skilling 2004) is a Monte Carlo technique for sweeping through the parameter space by comparing a new point with the likelihood of all previous points, and accepting it if its likelihood is higher than some of them (the lowest of which is then abandoned). Thus the sample moves through nested shells of probability. MultiNest breaks the volume of sampling points into ellipsoidal subvolumes and carries out the nested sampling within their union, but samples within one ellipsoid at a time. This concentrates the evaluation in a more likely viable region of parameter space while allowing for multi-modal distributions.

MultiNest exhibits accelerated convergence relative to MCMC, and better treatment of multi-modal probabilities. However, MultiNest is still designed to integrate the entire Bayesian posterior over all of parameter space. Credible limit contours on this posterior are derived as a side-effect, rather than being intentionally mapped out in detail. One result of this, which we present in Section 5.2, is that MultiNest spends considerable amount of time evaluating  $\chi^2$  at points in parameter space well outside of the credible limit contour. Such evaluations are important when trying to find the full Bayesian evidence of a model, but are not very helpful if all one is interested in are the multi-dimensional credible limit contours. We designed DALE $\chi$  for those situations where we want to accurately map out the confidence intervals of parameters, rather than characterize the full Bayesian posterior across all of parameter space.

## 2.3. Active Parameter Search

MCMC and MultiNest are both sampling algorithms. They attempt to find Bayesian credible limits by drawing random samples from parameter space in such a way that the distribution of the samples accurately characterizes the Bayesian posterior probability distribution on the parameter space. This strategy directly carries out the *optimize* step described in the introduction – drawing samples with a frequency determined by their posterior Bayesian probability ensures that the maximum likelihood point in parameter space has a high probability of being found – and only incidentally carries out the *characterize* step.

The APS algorithm (Bryan 2007; Bryan *et al.* 2007; Daniel *et al.* 2014), was designed to separately *optimize* and *characterize* the  $\chi^2$  function on parameter space. Separate searches are implemented to find local minima in  $\chi^2$  – this is the *optimization* – and then to locate points about those minima such that  $\chi^2 = \chi_{\text{lim}}^2$ , where  $\chi_{\text{lim}}^2$  defines the desired frequentist confidence limit (for example in two dimensions the 95% CL is  $\chi_{\text{lim}}^2 = \chi_{\text{min}}^2 + 6.0$ ). Daniel *et al.* (2014) present a way of translating these into Bayesian credible limits (see their Section 2.4). These latter searches represent the *characterization*.

Though APS improved over MCMC in accurately characterizing complex probability surfaces, and its convergence properties were comparable, it remains desirable to speed up parameter estimation. Indeed, as mentioned in the previous subsection, nested sampling techniques such as

MultiNest can also deal well with multimodal distributions and run faster than MCMC. Motivated by the desire for speed, while retaining and improving APS’ ability to efficiently characterize confidence intervals, even for complex probability surfaces, we present an algorithm (DALE $\chi$ ) with accelerated convergence in mapping the likelihood.

### 3. DALE $\chi$ (Direct Analysis of Limits via the Exterior of $\chi^2$ )

Like APS, DALE $\chi$  operates by trying to find points directly on the  $\chi^2 = \chi_{\text{lim}}^2$  contour, rather than by inferring that contour from samples drawn from the posterior. This requires accurate knowledge of the value of  $\chi_{\text{lim}}^2$ , which, as discussed in Section 4.1 below, requires accurate knowledge of the value of  $\chi_{\text{min}}^2$ . Thus, before beginning the *characterization* of the  $\chi^2$  contour, we must *optimize* by finding  $\chi_{\text{min}}^2$ .

In the descriptions below, for each subsection we first give a short qualitative description to provide the key flavor of the subroutine, then a more quantitative description of its implementation, and finally any detailed treatments of computational subtleties. According to the readers’ interest, they can go into whatever depth they want. The summary of the overall algorithm is given in Section 3.2.5.

#### 3.1. Optimization in DALE $\chi$

The *optimization* carried out by DALE $\chi$  is based on the simplex searching algorithm of Nelder and Mead (1965). It is well known that the simplex search is not robust against functions with multiple disconnected local minima. We therefore augment the simplex with a Monte Carlo component designed to allow DALE $\chi$  to dig itself out of false  $\chi^2$  minima.

Qualitatively, the algorithm described below involves exploring the parameter space along diverse directions, detecting local minima, finding absolute minima, and checking for multimodality. In detail, for a parameter space of dimension  $D$ , initialize  $2 \times (D + 1) + D/2$  points on an ellipsoid whose axis in each direction is the span of the full parameter space to be searched. As will be seen below, this initial search culminates by using two independent Nelder-Mead simplexes to try to find the global  $\chi_{\text{min}}^2$ . The number of points initialized is chosen to be enough to initialize these simplexes (the Nelder-Mead simplex is driven by the motion of  $D + 1$  ‘particles’ in parameter space) plus a little extra. Each of these initial  $2 \times (D + 1) + D/2$  particles will take  $100 \times D$  steps driven by a Metropolis-Hastings MCMC algorithm with Gibbs sampling (Casella and George 1992) and simulated annealing (Kirkpatrick *et al.* 1983).

That is to say: at each step for each particle, randomly choose a basis vector (i.e. direction in basis parameter space) along which to step. Select a value  $r$  from a normal distribution with mean zero and standard deviation unity. Step from the particle’s current position along the chosen basis vector a distance  $r \times n_i$  where  $n_i$  is the difference between the maximum and min-

imum particle coordinate value along the chosen basis vector. Accept the step with probability  $\exp[-0.5 \times (\chi_{\text{new}}^2 - \chi_{\text{old}}^2)/T]$  where the “temperature”  $T$  is initialized at unity and adjusted every  $10 \times D$  steps. When adjusted,  $T$  is set to that value which would have caused half of the steps taken since the last  $T$ -adjustment to be accepted. Every  $4 \times D$  steps, select a new set of random basis vectors and adjust the  $n_i$  values accordingly. Keep track of both the absolute  $\chi^2$  minimum and the recent  $\chi^2$  minimum (to be defined momentarily) encountered by each particle. This is not necessarily the particle’s current position, as the  $T$  term in the acceptance probability is designed to cause the particles to jump out of local minima in search of other local minima. If ever a particle goes  $10 \times D$  steps without updating its recent minimum, assume that it has settled into a local minimum from which it will not escape. Move the particle back onto the ellipsoid where the particles were initialized. Set the particle’s recent  $\chi^2$  minimum (but not its absolute) to the value at its new location (which will almost certainly not be a local minimum) and resume walking.

After each particle has taken its  $100 \times D$  steps, choose the  $D + 1$  smallest absolute minima encountered by the particles, and use them to seed a simplex minimization search. After that search, choose the  $D + 1$  smallest absolute minima which are not connected to the absolute minimum with the lowest  $\chi^2$  value and use them to seed another simplex search, in case the first simplex search converged to a false minimum. The connectedness of two absolute minima is judged by evaluating  $\chi^2$  at the point directly between the two minima. If  $\chi^2$  at this midpoint is less than the larger of the two  $\chi^2$  absolute minima, then it is assumed that there is a valley in  $\chi^2$  connecting the two, and they are deemed “connected.” If not, it is assumed that there is a barrier of high  $\chi^2$  cutting the two off from each other, and they are deemed “disconnected.”

This is a fairly weighty algorithm, requiring a number of  $\chi^2$  evaluations that scales as  $D^2$ . However, in the case of highly complex likelihood surfaces, we find that this is necessary for accuracy in the search for the absolute minimum as well as multimodality. In addition it increases efficiency in that it keeps DALEX from wasting time characterizing the  $\chi^2$  contour around a false minimum. We have tested this algorithm on several multi-dimensional toy  $\chi^2$  functions, running the algorithm many times on each function, each time with a different random number seed. We find that the current algorithm with its values tuned ( $100 \times D$  steps per particle,  $4 \times D$  steps before resetting the bases,  $10 \times D$  steps before adjusting  $T$ , etc.) consistently finds the global  $\chi_{\text{min}}^2$  or a point reasonably close to it. In fact we find in Sec. 5 that it frequently finds a lower  $\chi_{\text{min}}^2$  than MultiNest. That being said, we are open to the possibility that a more efficient means of robustly optimizing  $\chi^2$  exists and would welcome its incorporation into DALEX. The algorithm described above is implemented in the files `include/dalex_initializer.h` and `src/dalex/dalex_initializer.cpp` in our codebase, for anyone wishing to inspect it further.

### 3.2. Characterization in DALEX

Once DALEX has found  $\chi_{\text{min}}^2$ , it can begin to characterize the shape of the  $\chi^2 \leq \chi_{\text{lim}}^2$  contour surrounding that minimum. This is achieved via successive minimizations of the cost function

introduced in equation (1). Equation (1) is designed to drive DALE $\chi$  to find points that are both inside the desired contour and far from points previously discovered. The neighbor function  $N(\vec{\theta})$  increases as DALE $\chi$  gets farther from previously known points. The exterior function  $E(\chi^2)$  exponentially decays with  $\chi^2$  once DALE $\chi$  leaves the confines of  $\chi^2 \leq \chi_{\text{lim}}^2$ . The parameter  $\ell$  in  $E(\chi^2)$  can be tuned to allow DALE $\chi$  to take circuitous routes through parameter space, swerving in and out of  $\chi^2 \leq \chi_{\text{lim}}^2$  in search of the minimum of equation (1). We implement equation (1) as a C++ class `cost_fn` defined in the files `include/cost_fn.h` and `src/dalex/cost_fn.cpp` in our codebase. We describe that class below.

### 3.2.1. The cost function

In order to evaluate equation (1), one requires a set of known  $\chi^2 \leq \chi_{\text{lim}}^2$  points from which to calculate the distances  $N$ . Select a set of previously discovered  $\chi^2 \leq \chi_{\text{min}}^2$  points  $\{\vec{A}\}$  to represent the already explored region of the confidence contour.  $\{\vec{A}\}$  will not always necessarily be all of the  $\chi^2 \leq \chi_{\text{lim}}^2$  points previously discovered by DALE $\chi$ . If the full set of  $\chi^2 \leq \chi_{\text{lim}}^2$  points is large, it may be helpful to thin them, taking, for instance, every third point and adding it to  $\{\vec{A}\}$ , to prevent calculations of  $N(\vec{\theta})$  from taking too much time. There may also be geometric considerations involved in what goes into  $\{\vec{A}\}$ , which we will discuss below. Upon instantiation, `cost_fn` loops through all of the points in  $\{\vec{A}\}$ , finding the range of  $\{\vec{A}\}$  in each of the cardinal basis directions. Choose one of these ranges, either the minimum or median as described below, as the normalizing factor  $s$ . When evaluating equation (1) at a point  $\vec{\theta}$ , the value of  $N(\vec{\theta})$  is the harmonic mean of the parameter space distances

$$d_A = \sqrt{\sum_i^D \left( \frac{\theta_i - \theta_{A,i}}{s} \right)^2} \quad (2)$$

for all of the points  $\vec{\theta}_A$  in  $\{\vec{A}\}$ .

The purpose of the  $N$  term in equation (1) is to drive simplex minimizations towards points in  $\chi^2 \leq \chi_{\text{lim}}^2$  that are far from previously explored regions of parameter space. The effectiveness of our search, however, will still depend on where the simplex search is initialized. Do we start the simplex close to  $\chi_{\text{min}}^2$  and hope that the repulsive properties of  $N(\vec{\theta})$  will drive DALE $\chi$  away from already explored points? Or should we start the simplex far outside of  $\chi^2 \leq \chi_{\text{lim}}^2$ , and use the interplay between the  $\chi^2$  and  $N(\vec{\theta})$  dependences of  $F(\vec{\theta})$  to approach unexplored regions of  $\chi^2 \leq \chi_{\text{lim}}^2$  from without. DALE $\chi$  uses both approaches complementarily, seeding the former with the results of the latter as described below.

### 3.2.2. Refining $\chi_{\min}^2$

First though, we add one refinement to the previous section. Since the utility of equation (1) involves correctly specifying  $\chi_{\lim}^2$ , DALEX’s performance depends on efficiently finding the true  $\chi_{\min}^2$  (this will be demonstrated in Section 4.1 below). Unfortunately, we sometimes find that in highly curved likelihood cases even the elaborate function minimization scheme described in Section 3.1 is insufficient to find the true minimum  $\chi^2$  value. We therefore supplement Section 3.1 with the following extra algorithm to refine  $\chi_{\min}^2$ .

Initialize a set of  $2 \times D$  MCMC chains. Take  $4 \times D$  steps in each of these chains, driving the MCMC with  $F(\vec{\theta})$  in place of  $\chi^2$ . Use simulated annealing (Kirkpatrick *et al.* 1983) to keep the acceptance rate near 0.5. After taking the  $4 \times D$  steps, randomly choose  $D$  of these particles. Use the chosen particles and the current  $\chi_{\min}^2$  point to seed a simplex minimization of  $\chi^2$ . If a new  $\chi_{\min}^2$  is found, repeat the process, remembering the positions of the MCMC particles between iterations. Once no new  $\chi_{\min}^2$  is found, the refinement is considered complete. This is found to work very well; as mentioned previously, in difficult cases DALEX routinely finds lower  $\chi_{\min}^2$  than MultiNest does.

### 3.2.3. Exploring $\chi^2 \leq \chi_{\lim}^2$ from without

With the  $\chi_{\min}^2$  found, we are ready to characterize the confidence contour. The exploratory search described below is implemented by the method `find_tendrils_candidates()` defined in the file `src/dalex/dalex.cpp` in our code base.

Our far-reaching simplex minimization of  $F(\vec{\theta})$  works as follows. Fit a  $D$ -dimensional ellipsoid to all of the  $\chi^2 \leq \chi_{\lim}^2$  points discovered thus far, using the algorithm described in Appendix A. Initialize a cost function using all of the  $\chi^2 \leq \chi_{\lim}^2$  points discovered so far as  $\{\vec{A}\}$ . If more than 20,000  $\chi^2 \leq \chi_{\lim}^2$  points have been discovered so far, thin  $\{\vec{A}\}$  so that it contains between 20,000 and 40,000 points; this is to make calculations of  $N(\vec{\theta})$  tractable. Set  $\ell = 0.25 \times (\chi_{\lim}^2 - \chi_{\min}^2)$  (minimum of  $\ell = 2.0$ ) so that simplex minimization of  $F(\vec{\theta})$  is free to use the region of parameter space just outside of  $\chi^2 \leq \chi_{\lim}^2$  to avoid evaluating  $\chi^2$  near already discovered regions of  $\chi^2 \leq \chi_{\lim}^2$ . Set  $s$  equal to the minimum range in the cardinal spans of  $\{\vec{A}\}$  so that  $N(\vec{\theta})$  is dominant and will drive the simplex towards the far edges of  $\chi^2 \leq \chi_{\lim}^2$  as  $F(\vec{\theta})$  is minimized. For each of the  $D$  axes of the ellipsoid above:

- (1E) Select a point  $3 \times r$  from the center of the ellipsoid, where  $r$  is the length of the semi-axis of the ellipsoid in the chosen direction.
- (2E) Select  $D$  other points centered on the point chosen in (1E). These points will be  $0.1 \times r_i$  away from the point chosen in (1E) along the other  $D$  axes of the ellipsoid.
- (3E) Use these  $D + 1$  points to seed a simplex and minimize  $F(\vec{\theta})$  from those points. Add any  $\chi^2 \leq \chi_{\lim}^2$  points discovered to  $\{\vec{A}\}$ .



- (4E) Repeat steps (1E)-(3E) for the point  $-3 \times r$  away from the ellipsoid’s center.

In principle, this search begins by surrounding  $\chi^2 \leq \chi_{\text{lim}}^2$  at a great distance (the initial seed points at  $\pm 3 \times r$  distance away from the center of the ellipsoid) and closing in on the  $\chi^2 \leq \chi_{\text{lim}}^2$  contour, hopefully finding the most distant corners of the contour under the influence of  $N(\vec{\theta})$ .

Because the initialization of the cost function for this search requires  $\chi^2 \leq \chi_{\text{lim}}^2$  points to populate  $\{\vec{A}\}$ , we initialize DALE $\chi$  by running this search once with the initial seeds selected at  $\pm r$  instead of  $\pm 3 \times r$ . This search is then immediately run again with the usual seeds at  $\pm 3 \times r$ .

### 3.2.4. Exploring the cost function from within

Though the algorithm described in Section 3.2.3 is effective at finding the extremities of  $\chi^2 \leq \chi_{\text{lim}}^2$ , it has no mechanism for filling in the complete contour. Indeed, if  $F(\vec{\theta})$  behaves the way it is designed to behave, we should hope that the end result of Section 3.2.3 is small clusters of  $\chi^2 \leq \chi_{\text{lim}}^2$  points widely spread from each other in parameter space with few evaluated points connecting them. Therefore, in addition to the exploratory search of Section 3.2.3, DALE $\chi$  runs the following search to trace out the contour. We refer to this search as the “tendrill search,” since it is designed to behave like a vine creeping across a wall. A tendrill search is actually a series of simplex minimizations concatenated in such a way as to force DALE $\chi$  to fully explore any non-Gaussian wings present in the likelihood function. A single DALE $\chi$  will perform several independent tendrill searches over the course of its lifetime. We refer to the individual simplex minimizations making up a single tendrill search as ‘legs’ below.

A tendrill search is performed as follows. After concluding the exploratory search described in Section 3.2.3, take the  $2 \times D$  end points of the simplex searches and rank them according to their cost function values, using the cost function from Section 3.2.3. Save only the  $D/2$  points with the lowest cost function values. These will be the potential seeds for the tendrill search. When the time comes to perform a tendrill search, initialize a new cost function. Use every  $\chi^2 \leq \chi_{\text{lim}}^2$  point so far discovered, except those discovered in the course of the search described in Section 3.2.3 to populate  $\{\vec{A}\}$ . Set  $\ell = 1.0$  and use the median cardinal span of  $\{\vec{A}\}$  as the normalizing factor  $s$ . Rank the  $D/2$  seed candidates chosen above according to their values in the new cost function. Choose the candidate with the lowest cost function value, so long as that candidate is not inside any of the exclusion ellipsoids  $\{X\}$  (“exclusion ellipsoids” will be defined below; the first time the tendrill search is run,  $\{X\}$  is empty). If no candidates remain that are outside of  $\{X\}$ , re-run the search from Section 3.2.3. In this way, DALE $\chi$  alternates between trying to fill in the  $\chi^2 \leq \chi_{\text{lim}}^2$  with a tendrill search, and trying to expand the boundaries of  $\chi^2 \leq \chi_{\text{lim}}^2$  with the exploratory search of Section 3.2.3.

Every leg of a tendrill search has an origin point  $\vec{\gamma}$ , from which it will start, and a meta-origin point  $\vec{\delta}$ . The meta-origin  $\vec{\delta}$  of the current leg is the origin  $\vec{\gamma}$  of the previous leg. We use the vector  $\vec{\gamma} - \vec{\delta}$  to impose structure on the seeds  $\vec{\gamma}$  of the simplex underlying the leg and thus try to keep the

legs in a tendril search moving in the same general direction and prevent them from turning around before they have thoroughly explored their present region of parameter space, assuming this still keeps them within  $\chi^2 \leq \chi_{\text{lim}}^2$ . In the case of the first leg in a tendril search,  $\vec{o}$  is set to the  $\chi_{\text{min}}^2$  point.

Set the candidate selected from the end points of Section 3.2.3 as  $\vec{\gamma}$  and build a simplex minimizer around it. The seeds for this simplex are found by fitting an ellipsoid to all of the  $\chi^2 \leq \chi_{\text{lim}}^2$  points discovered by every tendril search performed by DALE $\chi$  thus far that are judged to be “connected” to  $\vec{\gamma}$  (if this is the first tendril search, simply use all of the  $\chi^2 \leq \chi_{\text{lim}}^2$  points discovered thus far). Use bisection to find  $\chi^2 = \chi_{\text{lim}}^2$  points centered on  $\vec{\gamma}$  but displaced from it along the directions  $\vec{e} + \vec{b}$  where  $\vec{e}$  are the axial directions of the ellipsoid fit to the  $\chi^2 \leq \chi_{\text{lim}}^2$  points and  $\vec{b}$  is the unit vector pointing parallel to  $\vec{\gamma} - \vec{o}$ . Choose as seeds the  $D$  points midway between  $\vec{\gamma}$  and the  $\chi^2 = \chi_{\text{lim}}^2$  points discovered by bisection. The  $D + 1$ th seed point is  $\vec{\gamma} + \beta\vec{b}$  where  $\beta$  is the average distance from the first  $D$  seed points to  $\vec{\gamma}$ . Use these  $D + 1$  seed points to seed a simplex minimization of the cost function initialized above. Once the simplex minimization has converged, set  $\vec{\gamma} \rightarrow \vec{o}$  and set the endpoint of the simplex search to be  $\vec{\gamma}$  for the next leg of the tendril search. Rebuild the cost function adding any newly discovered  $\chi^2 \leq \chi_{\text{lim}}^2$  points to  $\{\vec{A}\}$ , and repeat the process above. In this way, DALE $\chi$  concatenates successive simplex minimizations of  $F(\vec{\theta})$ , expanding the set of points used to calculate  $N(\vec{\theta})$  as it goes, hopefully filling in the  $\chi^2 \leq \chi_{\text{lim}}^2$  contour as it proceeds. This process is repeated until convergence, which we will define below, at which point a new candidate is chosen from those discovered in Section 3.2.3, and a new tendril search is begun.

There are a few hitherto unmentioned subtleties in the tendril search as described above. The ellipsoid whose axial directions are used to find the seeds for the simplex is fit to all of the  $\chi^2 \leq \chi_{\text{lim}}^2$  points deemed “connected” to  $\vec{\gamma}$ . We defined one use of the word “connected” in Section 3.1: evaluate  $\chi^2$  at the point halfway between  $\vec{\gamma}$  and  $\vec{\theta}_i$ . If  $\chi^2 \leq \chi_{\text{lim}}^2$ ,  $\vec{\theta}_i$  is connected to  $\vec{\gamma}$ . They are disconnected otherwise. If this were really what DALE $\chi$  did to judge the connection between each point evaluated in each tendril search, DALE $\chi$  would quickly blow its budget of  $\chi^2$  evaluations on judging connectivity. Therefore, as the tendril search progresses, we keep a look up table associating each discovered  $\chi^2 \leq \chi_{\text{lim}}^2$  point with a “key point” in parameter space. For each individual leg of a tendril search, the “key points” are its origin, its end, and its midpoint. Each  $\chi^2 \leq \chi_{\text{lim}}^2$  discovered during the search is associated with the key point that is closest to it in parameter space. To determine if a  $\vec{\gamma}$  is connected to a point previously discovered by tendril searching, it is deemed sufficient to determine if  $\vec{\gamma}$  is connected to the associated key point. This limits the number of unique  $\chi^2$  calls used to judge connectivity.

While the tendril search as described thus far does a better job of filling in the  $\chi^2 \leq \chi_{\text{lim}}^2$  contour than does the purely exploratory search of Section 3.2.3, it is still focused on reaching out into uncharted regions of parameter space. As such, it can sometime miss the breadth of the  $\chi^2 \leq \chi_{\text{lim}}^2$  contour perpendicular to the  $\vec{\gamma} - \vec{o}$  direction. We therefore supplement the tendril search with the following search meant to fill in the breadth of the contour regions discovered by pure

tendrils searching.

After minimizing  $F(\vec{\theta})$ , find the vector  $\vec{g}$  pointing from  $\vec{\gamma}$  to the endpoint of the simplex minimization. Let  $G$  be the parameter space L2 norm of  $\vec{g}$ . Select  $D$  random normal vectors  $\vec{p}_i$  perpendicular to  $\vec{g}$ . For each of the vectors  $\vec{p}_i$ , construct a parameter space direction

$$\vec{c} = \vec{g}/G + \epsilon\vec{p}_i \tag{3}$$

where  $\epsilon$  is a random number between 0 and 1. Normalize  $\vec{c}$ . Sample  $\chi^2$  at the points  $\vec{\gamma} + \delta\vec{c}$ , where  $\delta$  steps from  $0.1 \times G$  to  $G$  in  $0.1 \times G$  increments, thus constructing a  $D$ -dimensional cone with its vertex at  $\vec{\gamma}$  and opening towards the end point of the simplex minimization. In this way, DALEX tries to fill in the full volume of the region of  $\chi^2 \leq \chi_{\text{lim}}^2$  discovered by the tendrils search. This search is performed after each leg of each tendrils search (i.e. after each individual simplex minimization for a specific  $[\vec{\gamma}, \vec{\sigma}]$  pair).

We said above that DALEX runs the tendrils search “until convergence”. Convergence is a tricky concept in Bayesian samplers. It is even trickier in DALEX. Because DALEX is not attempting to sample a posterior probability distribution, there is no summary statistic analogous to Gelman and Rubin’s  $R$  metric for MCMC (Gelman and Rubin 1992; Brooks and Gelman 1998) that will tell us that DALEX has found everything it is going to find. We therefore adopt the heuristic convergence metric that as long as DALEX is exploring previously undiscovered parameter space volumes with  $\chi^2 \leq \chi_{\text{lim}}^2$ , DALEX has not converged. The onus is thus placed on us to determine in a way that can be easily encoded which volumes in parameter space have already been discovered.

In Appendix A we describe an algorithm for taking a set of parameter space points  $\{\vec{P}\}$  and finding an approximately minimal  $D$ -dimensional ellipsoid containing those points. Each time DALEX runs a tendrils search as described in the previous paragraphs, we amass the  $\chi^2 \leq \chi_{\text{lim}}^2$  points discovered into a set  $\{\vec{T}\}$  and fit an ellipsoid to that set as described in Appendix A. While running the tendrils search, DALEX keeps track of the end point of each individual leg of the search. If the simplex search ends in a point that is contained in one of the ellipsoids resulting from a previous tendrils search (the set of “exclusion ellipsoids”  $\{X\}$  referenced earlier), the simplex search is considered to be doubling back on a previously discovered region of parameter space and a “strike” (analogous to what happens when a batter in baseball swings at the ball and misses) is recorded. Similarly, if the simplex point ends inside of the ellipsoid constructed from the points  $\{\vec{T}\}$  discovered by the previous legs comprising the current tendrils search (and the volume of that ellipsoid has not expanded due to the current simplex) a “strike” is recorded.

Each time a “strike” is recorded, the tendrils search backs up and starts from the last acceptable simplex end point, adopting the corresponding  $\vec{\gamma}$  and  $\vec{\sigma}$ . If ever three “strikes” in a row are recorded, the tendrils is deemed to have converged (or, to continue the baseball analogy, to have “struck out”). An ellipsoid is fit to the  $\chi^2 \leq \chi_{\text{lim}}^2$  points  $\{\vec{T}\}$  discovered by the entire tendrils search. This ellipsoid is added to the set  $\{X\}$  of previously discovered “exclusion ellipsoids.” By refusing to start a new tendrils search in any of the previously discovered exclusion ellipsoids, we increase the likelihood that each successive sequence of tendrils searches will discover a previously unknown region of  $\chi^2 \leq \chi_{\text{lim}}^2$ .

We find that in practice (see Section 5) the “three strikes and you’re out” rule works well in both computational efficiency and completeness in characterization – often better than MCMC and MultiNest algorithms that have declared convergence.

### 3.2.5. Final algorithm

We have described three search modes in addition to the initial optimization of Section 3.1. Section 3.2.2 describes an algorithm that takes steps which are clustered around  $\chi_{\min}^2$ , trying to ensure that DALE $\chi$  has found the true value of  $\chi_{\min}^2$ . Section 3.2.3 describes an algorithm that searches for  $\chi^2 \leq \chi_{\lim}^2$  regions by closing in on them from the  $\chi^2 > \chi_{\lim}^2$  parameter space. Section 3.2.4 explores the  $\chi^2 \leq \chi_{\lim}^2$  contour from within, focusing on regions of parameter space that may not have been explored yet. We find that the diversity of optimization and characterization steps is well suited, and indeed crucial, to dealing with complicated likelihood surfaces, with degenerate/highly curved/non-Gaussian or multimodal properties. Equally important, in high-dimensional cases, DALE $\chi$  converges in a number of  $\chi^2$  evaluations significantly lower than required by MultiNest while finding regions of parameter space ignored by MultiNest (and indeed often values of  $\chi_{\min}^2$  lower than MultiNest).

We combine these three search modes into the following master algorithm for DALE $\chi$ .

- (1) Run the *optimization* algorithm from Section 3.1.
- (2) Perform the  $\chi_{\min}^2$ -refining search of Section 3.2.2.
- (3) Perform a single sequence of tendrils searches as described in Section 3.2.4 (recall that if no viable candidates for a tendrils search starting point exist, we will perform the search in Section 3.2.3 to find some).
- (4) Alternate steps (2) and (3) until convergence.

As stated before, “convergence” is not a well-defined formal concept for DALE $\chi$ . Presently, we have no global convergence condition. DALE $\chi$  runs until it has made a user-specified number of calls to  $\chi^2$ . Operationally, our local criterion works quite well, as shown in the next sections. Future work will explore development of a global convergence criterion.

## 4. A Cartoon Likelihood Function

Algorithms such as MCMC, MultiNest, and DALE $\chi$  are designed for use in characterizing likelihood functions complex enough that evaluating them the  $\approx 100^D$  times necessary to evaluate them directly on a well-sampled grid in parameter space would take too much time to be feasible. In many cases, such functions are so complex that even evaluating them the (comparatively) few

number of times necessary for the algorithms under consideration to converge can take hours or, in the worst cases, days. To facilitate more rapid testing during the development of DALE $\chi$ , we designed a cartoon likelihood function which mimicked the non-linear, non-Gaussian behavior of realistic multi-dimensional likelihood functions, but which could be evaluated almost instantaneously on personal computers. These functions are made available in our code via the header file `include/exampleLikelihoods.h`. The likelihood functions do depend on other compiled objects from our code base. Users should consult the `Makefile`, specifically the example executables `curved_4d`, `curved_12d`, and `ellipse_12d`, for examples how to compile software using our toy likelihoods.

In order to understand how we have constructed our cartoon likelihood function, it is useful to consider an example from nature. Since the authors are cosmologists, the example we choose is using the cosmic microwave background (CMB) anisotropy spectrum as measured by an experiment such as Planck (Planck Collaboration 2013) to constrain the values of cosmological parameters. The CMB likelihood function is, in the simplest case, defined on a 6-dimensional parameter space  $\{\Omega_m, \Omega_b, h, \tau, n_s, A_s\}$ . (In reality there can be a dozen or more non-cosmological parameters as well.) Each combination of these 6 parameters implies a different spectrum of anisotropy for the CMB. This anisotropy is defined in terms of the  $C_\ell$  power coefficients of the multipole-moment  $\ell$  decomposition of the full CMB temperature distribution on the sky. Typical CMB experiments measure a few thousand of these  $C_\ell$  coefficients (in the temperature-temperature correlation function). To compute a likelihood value for any combination of cosmological parameters, the parameters are converted into their predicted  $C_\ell$  spectrum and this  $C_\ell$  spectrum is compared to the actual measurement. In other words, there is an unknown, non-linear function  $g(\vec{\theta}, \ell)$  such that

$$g(\{\Omega_m, \Omega_b, h, \tau, n_s, A_s\}, \ell) = C_\ell \quad (4)$$

To find the likelihood of any given combination of cosmological parameters, one must first evaluate  $g(\vec{\theta}, \ell)$  to get a test set of  $C_\ell$ 's and then compare this set to the actual measured  $C_\ell$ 's, taking into account the appropriate covariance matrix.

To mimic this behavior, we construct our cartoon likelihood functions in three steps.

- (1) A function  $\tilde{f}(\vec{\theta}) = \vec{\mu}$  maps the cardinal parameters  $\vec{\theta}$  into auxiliary parameters  $\vec{\mu}$ .
- (2) A second function  $\tilde{g}(\vec{\mu}, \vec{x}) = \vec{y}(\vec{x})$  maps the parameters  $\vec{\mu}$  and an independent variable  $\vec{x}$  (analogous to  $\ell$  in the cosmological case (4) above) to a dependent value  $\vec{y}$  (analogous to  $C_\ell$ ).
- (3) A noisy cartoon data set  $(\vec{x}, \vec{y}_0, \sigma_{\vec{y}_0})$  is constructed by evaluating

$$\vec{y}_{\text{mean}} = \tilde{g}\left(\tilde{f}\left(\vec{\theta}_0\right), \vec{x}\right) \quad (5)$$

at 100 values of  $x$ . At each of these values of  $(x, \vec{y}_{\text{mean}})$ , 100 samples are drawn from a Gaussian distribution centered on  $\vec{y}_{\text{mean}}$ . (A Gaussian distribution is chosen as a simplification of experiment noise properties for the purposes of rapid testing.)  $\vec{y}_0$  is the mean of these 100 samples and  $\sigma_{\vec{y}_0}$  is the standard deviation of these samples.

To evaluate  $\chi^2$  at a test value of  $\vec{\theta}$ , one simply evaluates

$$\vec{y}_{\text{test}} = \tilde{g} \left( \tilde{f} \left( \vec{\theta}_{\text{test}} \right), x \right) \quad (6)$$

and compares

$$\chi^2 \left( \vec{\theta}_{\text{test}} \right) = \sum_i^{100} \left[ \left( \vec{y}_{\text{test},i} - \vec{y}_{0,i} \right) / \sigma_{\vec{y},0,i} \right]^2 \quad (7)$$

#### 4.1. Confidence versus Credible Limits: How to Set $\chi_{\text{lim}}^2$

Before proceeding, we have to define exactly what contours we are searching for. (The reader more interested in a direct comparison of results from different algorithms can skip to Section 5.) A significant philosophical difference between the sampling algorithms (MCMC and MultiNest) and the search algorithms (APS and DALE $\chi$ ) is in how each defines the desired boundary in parameter space (the “confidence limit” in frequentist terminology; the “credible limit” for Bayesians). Sampling algorithms learn the definition boundary as they go. They draw samples from the posterior probability distribution and, after convergence has been achieved, declare the limit to be the region of parameter space containing the desired fraction of that total distribution. Search algorithms require the user to specify some  $\chi_{\text{lim}}^2$  defining the desired limit a priori. This can be done either by setting an absolute value of  $\chi_{\text{lim}}^2$  or by demanding that  $\chi^2$  be within some  $\Delta\chi^2$  of the discovered  $\chi_{\text{min}}^2$ , i.e.  $\chi_{\text{lim}}^2 = \chi_{\text{min}}^2 + \Delta\chi^2$ . These two approaches are not wholly irreconcilable. Wilks’ Theorem (Wilks 1938) states that, for approximately Gaussian likelihood functions, the Bayesian credible limit (i.e. the contour bounding  $(1 - \alpha)\%$  of the total posterior probability) in a  $D$ -dimensional parameter space is equivalent to the contour containing all points corresponding to  $\chi^2 \leq \chi_{\text{min}}^2 + \chi_{(1-\alpha)\%}^2$  where  $\chi_{(1-\alpha)\%}^2$  is the value of  $\chi^2$  bounding the desired probability for a  $\chi^2$  probability distribution with  $D$  degrees of freedom. Even given this result, there remains some ambiguity in how one sets  $\chi_{(1-\alpha)\%}^2$ . Namely: one must choose the correct value of  $D$  for the degrees of freedom. Traditionally, credible and confidence limits are plotted in 2-dimensional sub-spaces of the full parameter space. The limits that are plotted are either marginalized or projected down to two dimensions. In the case of marginalization, the Wilks’-theorem equivalence between confidence and credible limits holds for  $D = 2$  degrees of freedom and  $\Delta\chi^2 = 6.0$  for the 95% contour. In the case of projections,  $D$  is equal to the full dimensionality of the parameter space. We argue and demonstrate below that the preferred choice, with the most robust results, is projection with  $D$  equal to the full dimensionality of the parameter space being considered. We illustrate this with a toy example.

In order to compare different definitions of confidence and credible limits, we construct a cartoon likelihood function on 4 parameters. Choosing such a low-dimensional parameter space allows us to grid the full space and directly integrate the Bayesian credible limits without requiring unreasonable computational resources. Figure 1 shows the marginalized likelihood of our 4-dimensional cartoon likelihood function in each of the 2-dimensional parameter sub-spaces. As one can see,

we have constructed our cartoon likelihood function to have a severely curved degeneracy in the  $\{\theta_0, \theta_1\}$  parameter sub-space and significantly non-Gaussian contours in the other five sub-spaces.

In Figures 2 and 3, we show the confidence and credible limits of our 4-dimensional toy likelihood function computed in several ways. The red contours show the *marginalized* Bayesian credible limits. These are the contours enclosing 95% (68%) of the marginalized likelihood shown in Figure 1. The green contours are the projections of the 4-dimensional contours enclosing  $\chi^2 \leq \chi_{\min}^2 + 6.0$  (2.28), i.e. they were generated by finding all of the points in 4-dimensional space satisfying  $\chi^2 \leq \chi_{\min}^2 + 6.0$  (2.28), converting those points into a scatter-plot in each of the 2-dimensional sub-spaces, and then drawing a contour around those scatter plots. Naively, one would expect these contours to be equivalent to the marginalized Bayesian contours, since 6.0 (2.28) is the 95% (68%) limit for a  $\chi^2$  distribution with two degrees of freedom. (Note that the 95.4% and 68.3% limits, i.e.  $2\sigma$  and  $1\sigma$  in Gaussian probability, would have  $\Delta\chi^2 = 6.17$  and 2.30.) This, however, is clearly not the case. The non-Gaussian nature of our toy likelihood prevents Wilks’ theorem from applying in the marginalized case. While you can integrate away  $d$  dimensions of a  $D$ -dimensional multivariate Gaussian and come up with a  $D - d$  multivariate Gaussian, no such guarantee exists for arbitrary, non-Gaussian functions.

The blue contours in Figure 2 and 3 show the *projected* Bayesian credible limits of our cartoon function. These were constructed by assembling all of the points in 4-dimensional space that contained 95% (68%) of the total likelihood, converting those points to a scatter plot in the 2-dimensional sub-spaces, and drawing a contour around the scatter plots. The purple contours show the similarly projected contours containing all of the points for which  $\chi^2 \leq \chi_{\min}^2 + 9.49$  (4.70), with 9.49 (4.70) being the 95% (68%) limit for a  $\chi^2$  distribution with four degrees of freedom. We refer to these contours as the “full-dimensional LRT (likelihood ratio test)” contours. We see remarkably good agreement between the projected Bayesian and full-dimensional LRT contours in both the 95% and the 68% limit, in accordance with Wilks’ theorem (since we have not marginalized away any of the 4 parameters in our parameter space, Wilks’ theorem still appears to hold).

In the rest of this paper, we will examine likelihood functions based on their projected Bayesian and full-dimensional LRT contours. Though it is traditional to consider marginalized contours in 2-dimensional sub-spaces, we note concerns that this may discard important information about the actual extent of parameter space considered reasonable by the likelihood function. In all cases in Figures 2 and 3, the projected Bayesian and full-dimensional LRT contours were larger than their 2-dimensional marginalized counterparts. This should not be surprising. The projected Bayesian and full-dimensional LRT contours contain all of the pixels in the full 4-dimensional space that fall within the desired credible limit. The marginalized counterparts take those full 4-dimensional contours and then weight them by the amount of marginalized parameter space volume they contain. The final marginalized contour thus represents an amalgam of the raw likelihood function and its extent in parameter space. It is not obvious that this is necessarily what one wants when inferring parameter constraints from a data set.

If a point in parameter space falls within the projected Bayesian or full-dimensional LRT

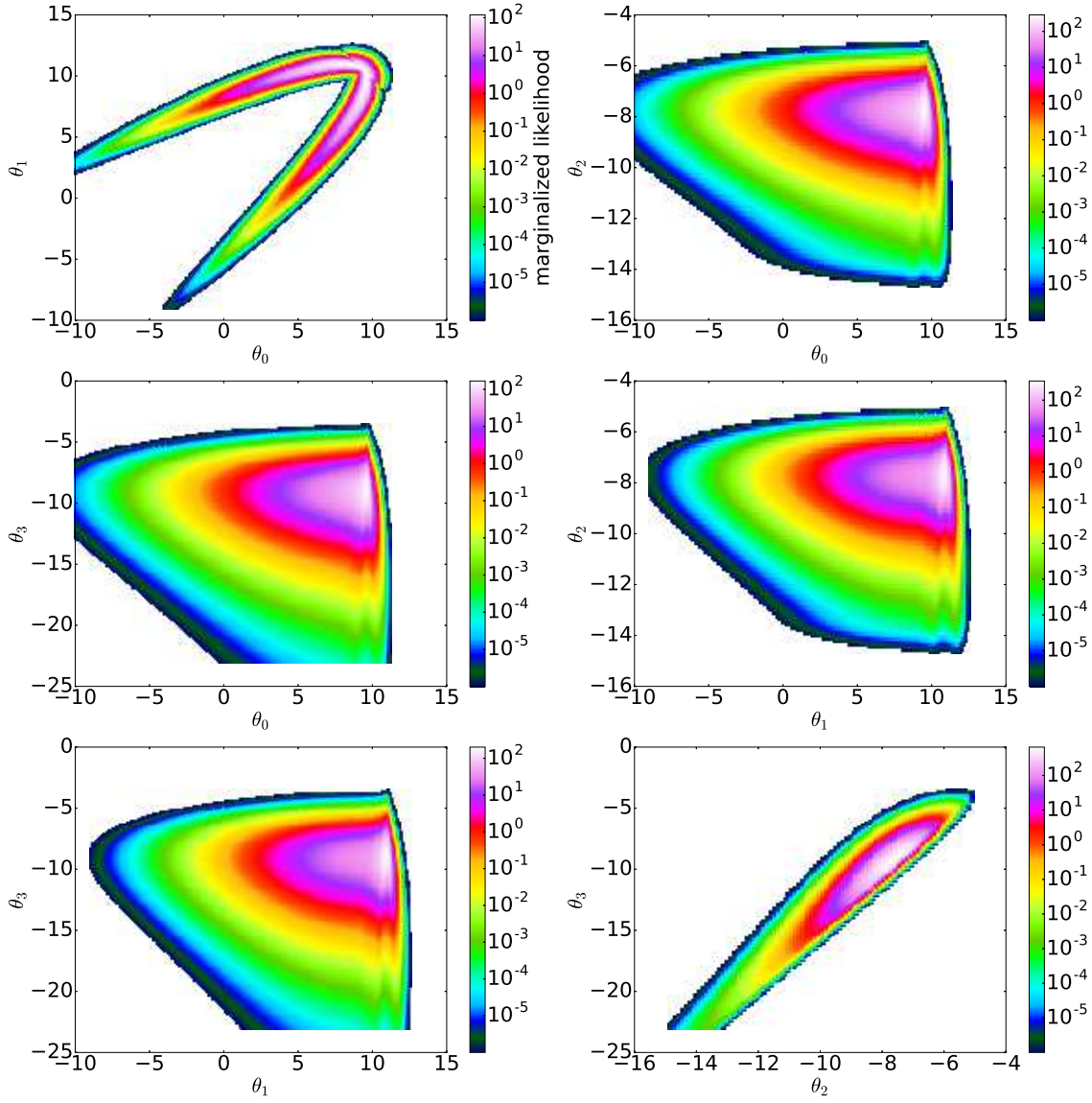


Fig. 1.— The marginalized likelihood in all of the 2-D sub-spaces of our 4-D cartoon likelihood function is shown, with the color bar indicating the marginalized likelihood. Note: the apparent discontinuity in the shape of the confidence contour near the bend in the  $\{\theta_0, \theta_1\}$  sub-space is an actual feature of our likelihood function, resulting from how the function is implemented in C++.



contour, it means that the fit to the data provided by that combination of parameters is, by some metric, reasonable. This is somewhat a matter of taste, but if one is really interested in all of the parameter combinations that give reasonable fits to the data, the conservative approach would be to keep all such combinations, rather than throwing out some, simply because they include rare – but fully accepted by the data – combinations of, say,  $\{\theta_1, \theta_3\}$  (in the case of the gap between the marginalized and projected Bayesian contours in the  $\{\theta_0, \theta_2\}$  plot in Figure 2). For this reason, we proceed<sup>1</sup>, comparing the performance of DALE $\chi$  with MultiNest by examining the projected Bayesian and full-dimensional LRT contours.

We acknowledge that there are some applications for which the approximate equivalence of  $\chi^2 \leq \chi_{\text{lim}}^2$  with a Bayesian credible limit for which we have argued above is insufficient. If a likelihood function is sufficiently non-Gaussian or a data set is sufficiently noisy, the two will not be equivalent. Indeed, in Figures 2 and 3, the blue and purple contours are not identical, but merely approximately the same. In cases where only the Bayesian credible limit is acceptable, we still believe that there is a use for DALE $\chi$ . Bayesian sampling algorithms generally get bogged down during “burn-in”, the period of time at the beginning of the sampling run during which the algorithm is simply learning where the minimum  $\chi^2$  point is and what the degeneracy directions of the contour are without efficiently sampling from the posterior. We show below that DALE $\chi$  converges to its confidence limits, with the same  $\chi_{\text{min}}^2$  and degeneracy directions as the Bayesian credible limit, irrespective of one’s opinion of Bayesian versus frequentist statistical interpretations, after an order of magnitude fewer calls to  $\chi^2$  than required by MultiNest. If one does not feel inclined to trust the full confidence limits returned by DALE $\chi$ , one should be able to achieve a significant speed-up in Bayesian sampling algorithms by using DALE $\chi$  as pre-burner: defining the region of parameter space to be sampled before sampling commences.

## 5. Comparison of Techniques

As a proof of concept, we begin by comparing the performance of MultiNest and DALE $\chi$  on the 4-dimensional likelihood function presented in Figure 1. As said before, four is an extremely low dimensionality and we find that MultiNest converges thoroughly and much more rapidly than DALE $\chi$  (though MultiNest does include some spurious points compared to exact, brute force calculation, unlike DALE $\chi$ , i.e. it has lower purity). Later, we will see that DALE $\chi$  out-performs MultiNest on higher-dimension likelihood functions. The purpose of this current discussion is to demonstrate that DALE $\chi$  does, in fact, find the limits expected from Figures 2 and 3.

Comparing MultiNest and DALE $\chi$  requires defining how long each should run. Because Multi-

---

<sup>1</sup> We do not consider useful the strictly frequentist perspective that, since our cartoon likelihood function “measures”  $\vec{y}(x)$  at 100 values of  $x$ , the confidence limit limit should therefore be set according to the  $\chi^2$  distribution with 100 degrees of freedom. This would set the 95% confidence limit at  $\chi^2 = 124.35$ . Note that since  $\chi_{\text{min}}^2$  for our toy likelihood function ended up being 83.12, such frequentist confidence limit contours would be significantly larger than anything shown in Figures 2 and 3. Such large contours seem unlikely to be useful.

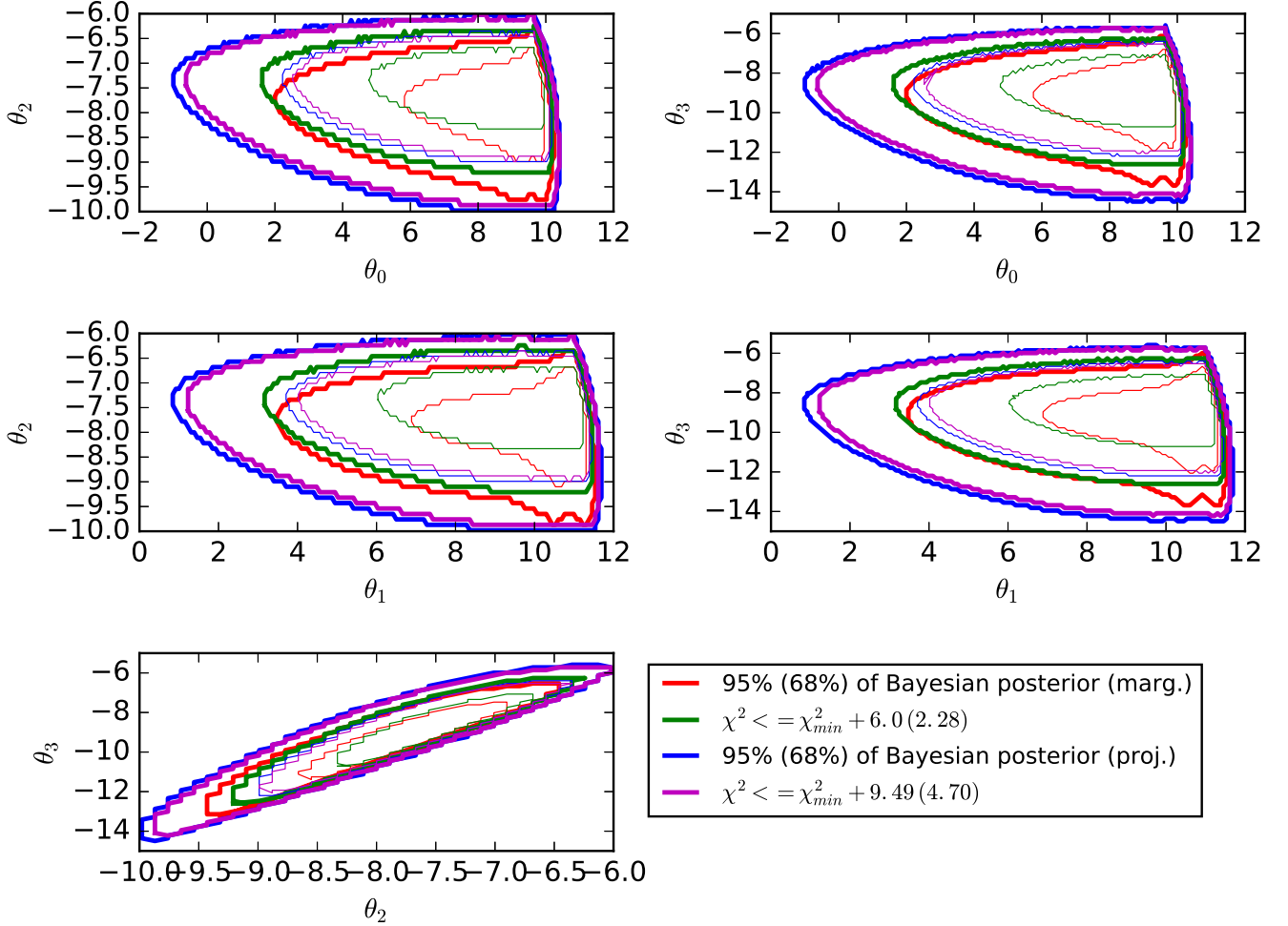


Fig. 2.— Comparison of the different confidence and credible limits are shown for the likelihood function introduced in Figure 1. The *projected* Bayesian credible limit agrees much better with the frequentist full dimensional likelihood ratio test ( $\chi^2$ ) confidence limit than the *marginalized* Bayesian credible limit. Thick contours are the 95% limit; thin contours are the 68% limit. The  $\{\theta_0, \theta_1\}$  parameter sub-space is shown in Figure 3.

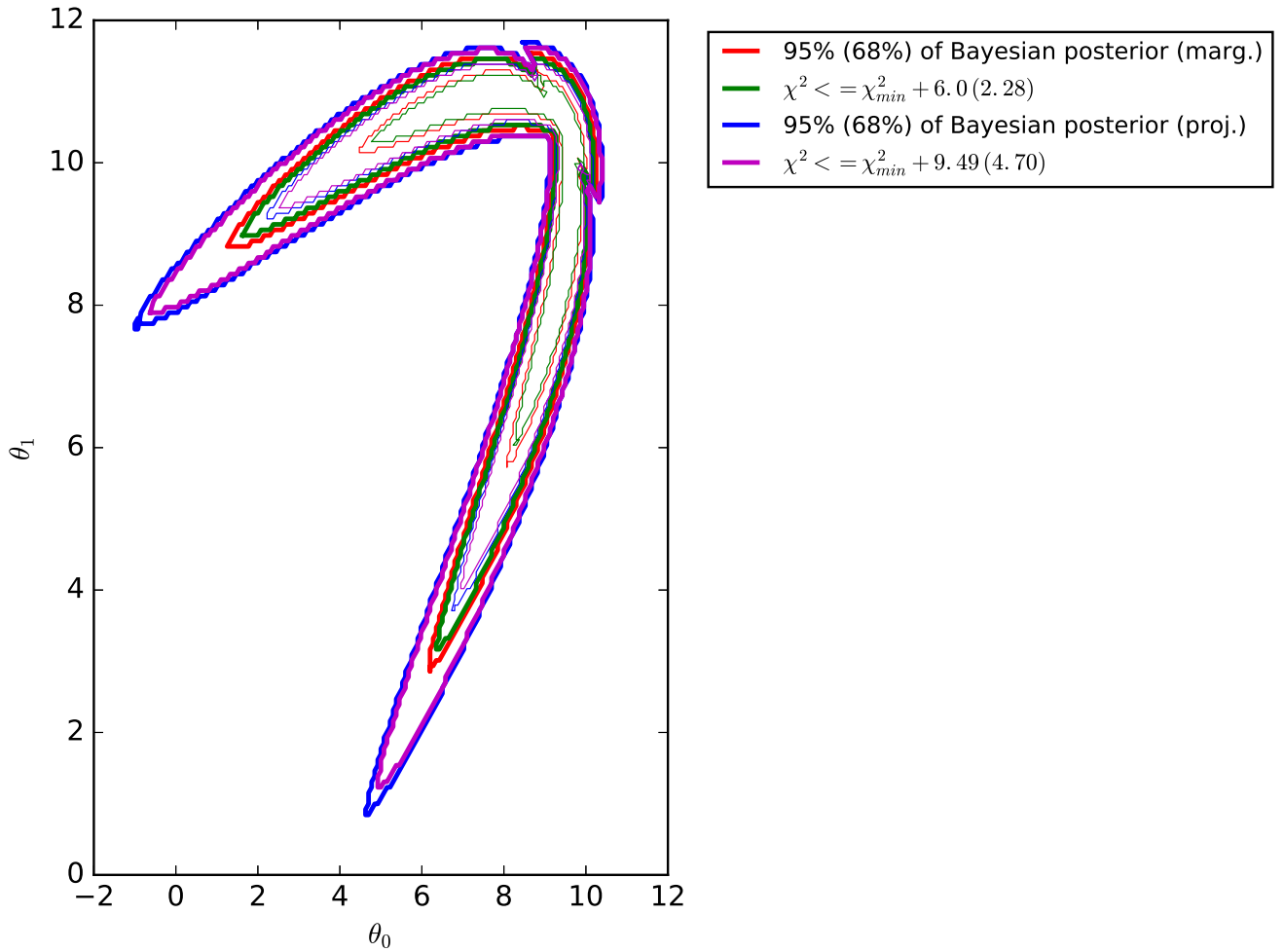


Fig. 3.— As Figure 2 but for the  $\theta_0$ - $\theta_1$  subspace. Due to the better Gaussianity in this subspace, both the projected and marginalized Bayesian credible limits agree with the likelihood ratio test.

Nest is a sampling algorithm designed to integrate the posterior distribution, it has a well defined convergence criterion: when the Bayesian evidence integrated over parameter space ceases to grow, the sampling is done. DALE $\chi$  does not have any analogous global convergence criterion. Individual tendrils searches within DALE $\chi$  can be said to converge when they “strike out” (see Section 3.2.4) but this says nothing about the overall convergence of DALE $\chi$ . It is up to the user to specify how many  $\chi^2$  evaluations they wish DALE $\chi$  to make. In Figure 4 below, we wish to show the convergence of MultiNest and DALE $\chi$  as a function of “time” (time being measured in the number of  $\chi^2$  evaluations being made). To do this, we initialized several MultiNest runs, each with a different number of “live” points. MultiNest instances with more live points took longer to converge (but can be expected to evaluate the posterior distribution more accurately). We then ran a single instance of DALE $\chi$  and asked it to quit after 40,000  $\chi^2$  evaluations. In Figure 4, we compare the MultiNest runs with the progress made by DALE $\chi$  after it had evaluated  $\chi^2$  as many times as it took MultiNest to converge. The red points are the 2-dimensional projections of the points discovered by MultiNest to be in the 95% Bayesian credible limit. The blue points are the  $\chi^2 \leq \chi_{\min}^2 + 9.49$  points discovered by DALE $\chi$  at the equivalent point in time. As stated, DALE $\chi$  does indeed find the correct  $\chi^2 \leq \chi_{\lim}^2$  contour while MultiNest is faster for  $D = 4$  and somewhat impure. We show below that, in higher-dimensional cases, DALE $\chi$  characterizes the contour much more robustly than MultiNest. We used version 3.9 of MultiNest here and throughout this paper. The most up-to-date version of MultiNest can be downloaded from <https://ccpforge.cse.rl.ac.uk/gf/project/multinest/>

### 5.1. Comparison on 12-dimensional case

We showed above that DALE $\chi$  found similar confidence limits, but took much longer to converge than did MultiNest on a 4-dimensional likelihood function. Now we compare DALE $\chi$  to MultiNest on a 12-dimensional cartoon likelihood function. Here we find that DALE $\chi$  converges an order of magnitude faster than MultiNest.

Before we show the direct comparison between DALE $\chi$  and MultiNest, we must be careful to define what we are comparing. MultiNest involves several user-determined parameters that control how the sampling is done. One in particular, the number of ‘live’ points, which is somewhat analogous to the number of independent chains run in an MCMC, has a profound effect both on how fast MultiNest converges and what limit it finds when it does converge. We created two 12-dimensional instantiations of our cartoon likelihood function, one with a very curved parameter space degeneracy, one without, and ran several instantiations of MultiNest with different numbers of ‘live’ points on those functions. Figures 5 and 6 show the projected Bayesian 95% credible limits found by MultiNest in two 2-dimensional sub-spaces for each of those likelihood functions as a function of the number of ‘live’ points. The figures also note how many evaluations of the likelihood function were necessary before MultiNest declared convergence. Not only does increasing the number of ‘live’ points increase the number of likelihood evaluations necessary for MultiNest to converge, it also expands the discovered credible limits. We interpret this to mean that running MultiNest with a small number of ‘live’ points results in convergence to an incomplete credible

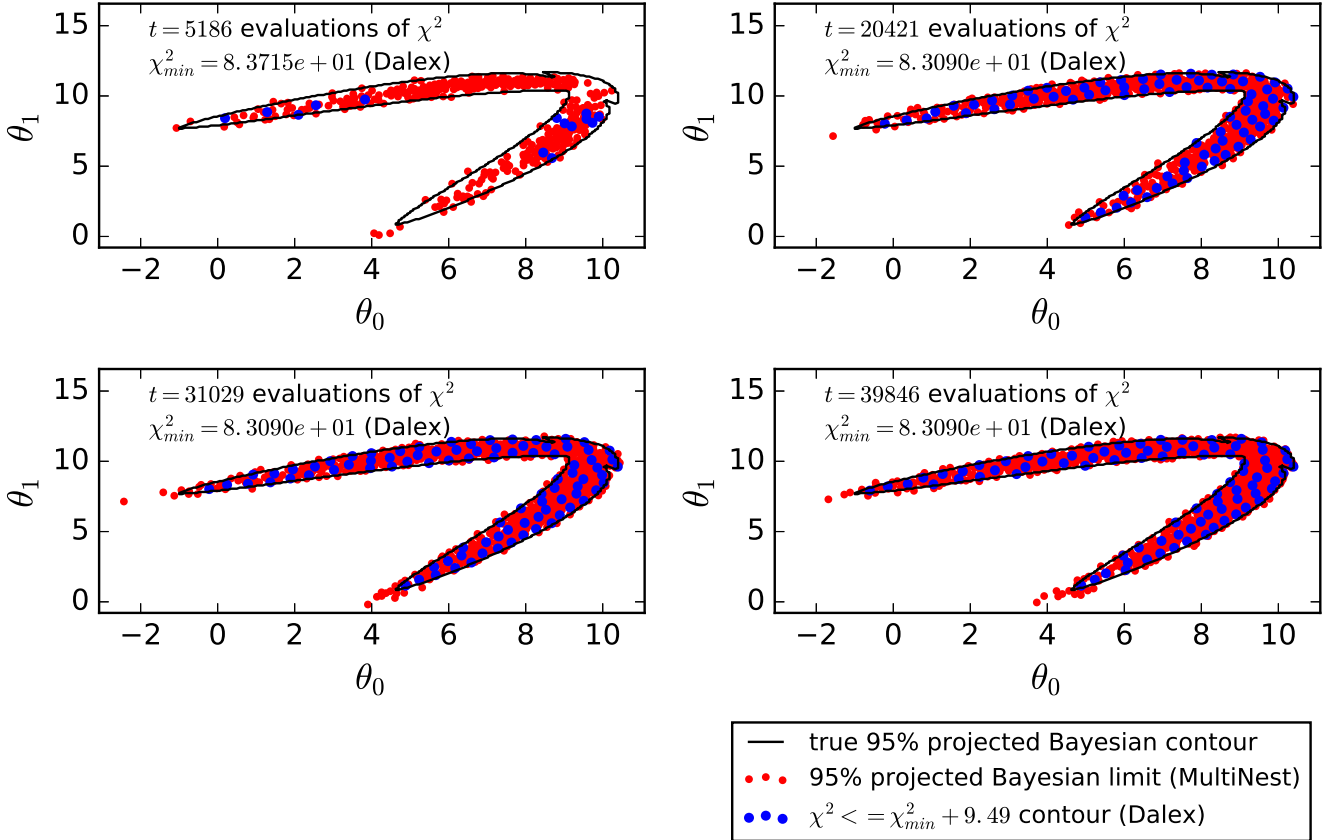


Fig. 4.— We demonstrate the convergence properties of MultiNest and DALEX on the 4-dimensional likelihood function from Figure 1. The black contour represents the true projected 95% Bayesian credible limit. The red points represent the points in the projected 95% Bayesian credible limit discovered by MultiNest. The blue points represent the  $\chi^2 \leq \chi_{min}^2 + 9.49$  points discovered by DALEX. Each panel represents a different MultiNest run, which took  $t$  points to converge. The blue points are all drawn from a single run of DALEX at the point during its history at which it had evaluated  $\chi^2$   $t$  times.

limit. When comparing DALE $\chi$  to MultiNest, we must therefore be careful which instantiation of MultiNest we compare against. Selecting a MultiNest run with too few ‘live’ points could mean that MultiNest appears as fast as DALE $\chi$ , but converges to an incorrect credible limit. Selecting a MultiNest run with too many ‘live’ points ensures that MultiNest has converged to the correct credible limit, but in an artificially large number of likelihood evaluations. In Figures 7 and 8, and Figures 10 and 11 below, we compare DALE $\chi$  both to a MultiNest instantiation that we believe has converged to the correct credible limit and a MultiNest instantiation that converges in a time comparable to DALE $\chi$ . This is to illustrate the relationship between convergence time and accuracy in MultiNest.

In Figures 7 and 8 we compare the convergence of DALE $\chi$  with that of MultiNest as a function of time. The likelihood function considered is a 12-dimensional cartoon constructed as in Section 4, but without the highly curved degeneracy direction. Each panel in the figures shows the state of the 95% confidence limits discovered by DALE $\chi$  after a specified number of evaluations of the likelihood function. The limits are plotted as scatter plots of the  $\chi^2 \leq \chi_{\text{lim}}^2$  points discovered by DALE $\chi$  at that point in time. Because MultiNest does not produce chains in the same way that MCMC does, we cannot display the evolution of the credible limit found by MultiNest as a function of time. Instead, we plot two credible limits: one found by an instantiation of MultiNest that converges after 210,000 evaluations of the likelihood function and one that converges after 2,700,000 evaluations of the likelihood function. As explained above, this difference in convergence time was achieved by tuning the number of ‘live’ points in the MultiNest run. There are two important features to note here.

The first feature is that DALE $\chi$  has converged to its final credible limit after about 125,000-150,000 calls to  $\chi^2$ . The second feature is the difference between the credible limits discovered by the two instantiations of MultiNest. True, one of the MultiNest instantiation converges in roughly (within a factor of two) as many likelihood evaluations as DALE $\chi$ . This instantiation of MultiNest, however, fails to explore the full non-Gaussian wings of the credible limit. See Figure 9 for a visualization of the non-Gaussianity. The slower MultiNest instantiation does find the same limits as DALE $\chi$ . However, it requires more than an order of magnitude more likelihood function evaluations to do so. This is the origin of our claim that DALE $\chi$  converges an order of magnitude faster than MultiNest on high-dimensional likelihood functions. DALE $\chi$  does not do as well as MultiNest characterizing the full width of the contour in the narrow degeneracy direction in Figure 8. However, by the end of the full 250,000  $\chi^2$ -evaluation run, one can see the wings of the contour beginning to trace out the full boundary. (Alternately, if one adopted the strategy suggested at the end of Section 4.1, using DALE $\chi$  to set the proposal distribution for a sampling algorithm, one could imagine filling in these contours more thoroughly than with just MultiNest alone while still only requiring  $\sim 10\%$  more evaluations of the likelihood function.)

Figures 10 and 11 show a similar comparison in the case of a 12-dimensional likelihood function with a highly curved parameter space degeneracy. Here again we see that DALE $\chi$  finds essentially the same limits as MultiNest, but more than an order of magnitude faster. We have done prelimi-

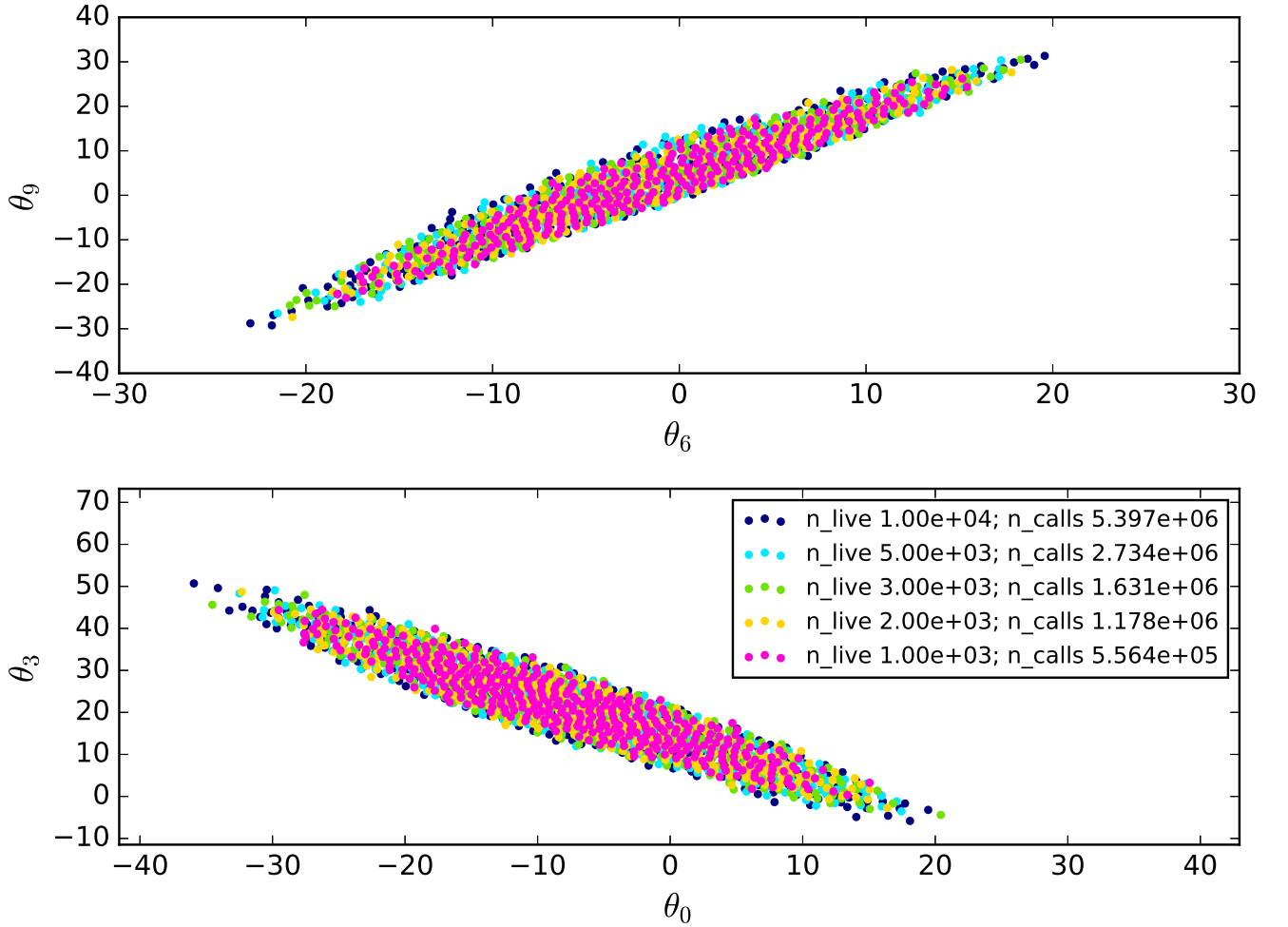


Fig. 5.— The projected 95% Bayesian credible limit found by MultiNest on one of our 12-dimensional cartoon likelihood functions. Each color denotes a different instance of MultiNest run with a different number of ‘live’ points (denoted by `n_live` in the legend). `n_calls` indicates how many calls to the likelihood function were necessary before MultiNest declared convergence.

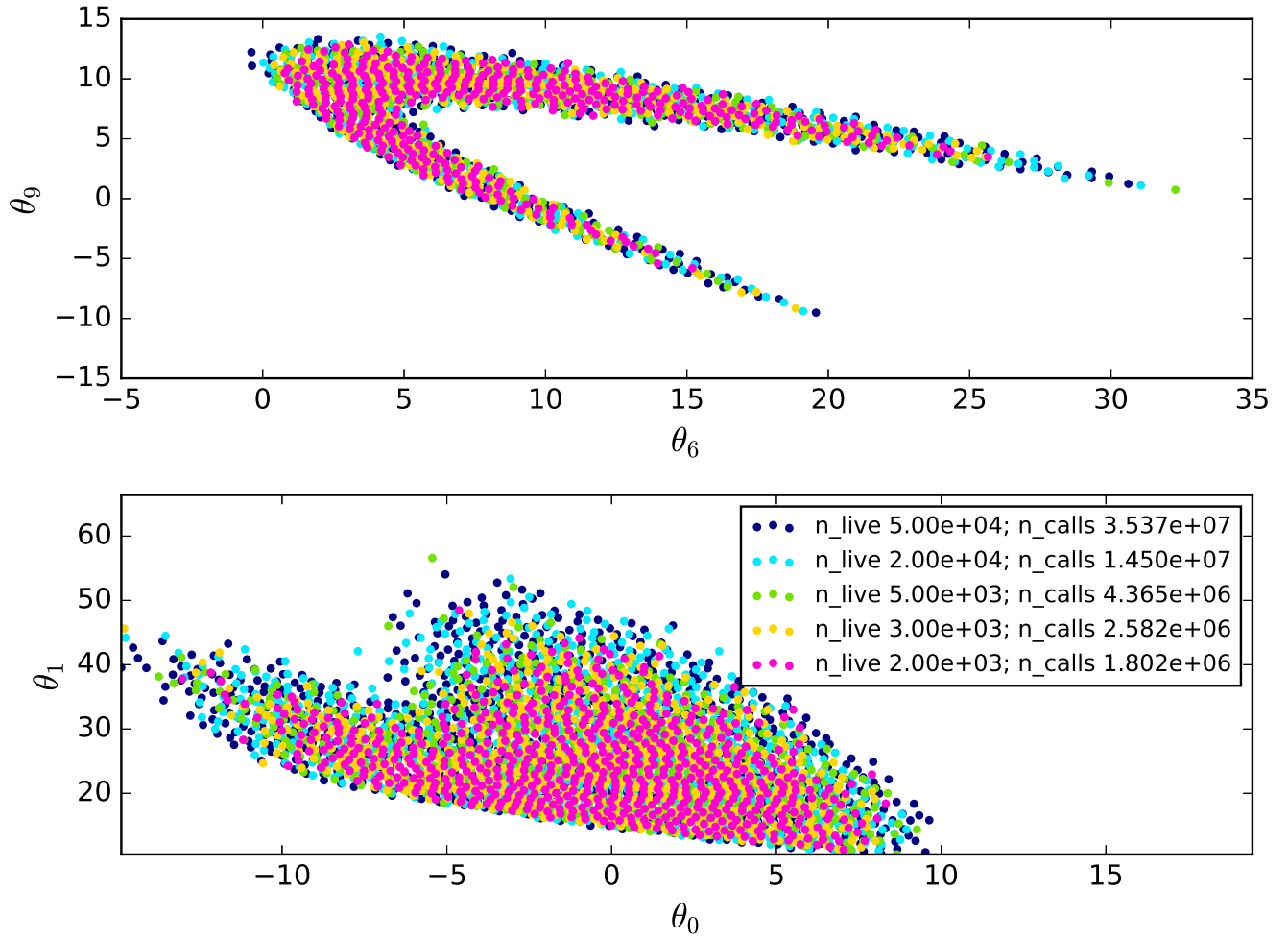


Fig. 6.— As Fig. 5 but for the curved degeneracy likelihood function.



nary tests with a 16-dimensional likelihood function. These show almost two orders of magnitude speed-up of DALE $\chi$  vis-à-vis MultiNest. In the 16-dimensional case, MultiNest requires 75 million calls to  $\chi^2$  to converge. DALE $\chi$  finds the same credible limit after only 1 million calls to  $\chi^2$ .

Regarding some computational specifics, note we have been measuring the speed of DALE $\chi$  and MultiNest in terms of how many calls to  $\chi^2$  are required to reach convergence. This is based on the assumption that, for any real physical use of these algorithms, the evaluation of  $\chi^2(\vec{\theta})$  will be the slowest part of the computation. For MultiNest, this is a fair assumption. As a Markovian process, the present behavior of MultiNest does not depend on its history. The same statement is not true of DALE $\chi$ . Because DALE $\chi$  is attempting to evaluate  $\chi^2$  at parameter-space points as far as possible away from previous  $\chi^2$  evaluations, DALE $\chi$  must always be aware of the full history of its search. Thus, the DALE $\chi$  algorithm implies some computational overhead in addition to the expense of simply evaluating  $\chi^2$ . For our 12-dimensional test cases, this overhead averages out to between 0.005 and 0.015 seconds per  $\chi^2$  evaluation on a personal laptop with a 2.9 GHz processor. For the 16-dimensional test case referenced above, this overhead varied between 0.01 and 0.03 seconds per  $\chi^2$  evaluation on the same machine. It is possible that future work with an eye towards computational optimization could reduce these figures. It is also possible that the steady march of hardware innovation could render such concerns irrelevant going forward. The memory footprint of DALE $\chi$  appears to be well within the capabilities of modern personal computers.

Note that both MultiNest and DALE $\chi$  depend on random number generators: MultiNest for driving sampling of the posterior, DALE $\chi$  for driving the Metropolis-Hastings-like algorithms described in Sections 3.1 and 3.2.2, and for selecting seeds for the simplex minimizations described in Sections 3.2.3 and 3.2.4. We have found that the convergence performance of DALE $\chi$  can depend on how that random number generator is seeded. Figure 12 plots six different runs of DALE $\chi$ , each with a different random number seed. The runs plotted were specifically selected to represent instances of DALE $\chi$  that did not find the full  $(\theta_0, \theta_1)$  confidence limit after 400,000  $\chi^2$  evaluations (all six successfully found the  $(\theta_6, \theta_9)$  confidence limit in 400,000 samples). In each panel, we show the confidence limit as found by DALE $\chi$  after 400,000, 600,000, 800,000  $\chi^2$  evaluations plotted against the true credible limit as discovered by a run of MultiNest with 50,000 live points, taking 35 million evaluations. In each case, the 400,000 evaluation confidence limit traces out the general shape of the true contour while the subsequent evaluations allow DALE $\chi$  to fill in the gaps in that first discovered confidence limit. Referring back to Figure 6, we see that, even after 4.3 million  $\chi^2$  evaluations, MultiNest has not found its final credible limit. Therefore, even these instances of sub-optimal runs of DALE $\chi$  demonstrate a significant speed-up vis-à-vis MultiNest. The difficulty is in knowing how many  $\chi^2$  evaluations is enough for DALE $\chi$ . This is the danger in not having a clear global convergence criterion for DALE $\chi$  and a subject worthy of further study.

For those interested, and in the interest of giving credit where credit is due, DALE $\chi$  uses the random number generator proposed by Marsaglia (2003) with initialization parameters proposed by Press *et al.* (2007). This is implemented in the class `Ran` defined in `include/goto_tools.h` in our code base.

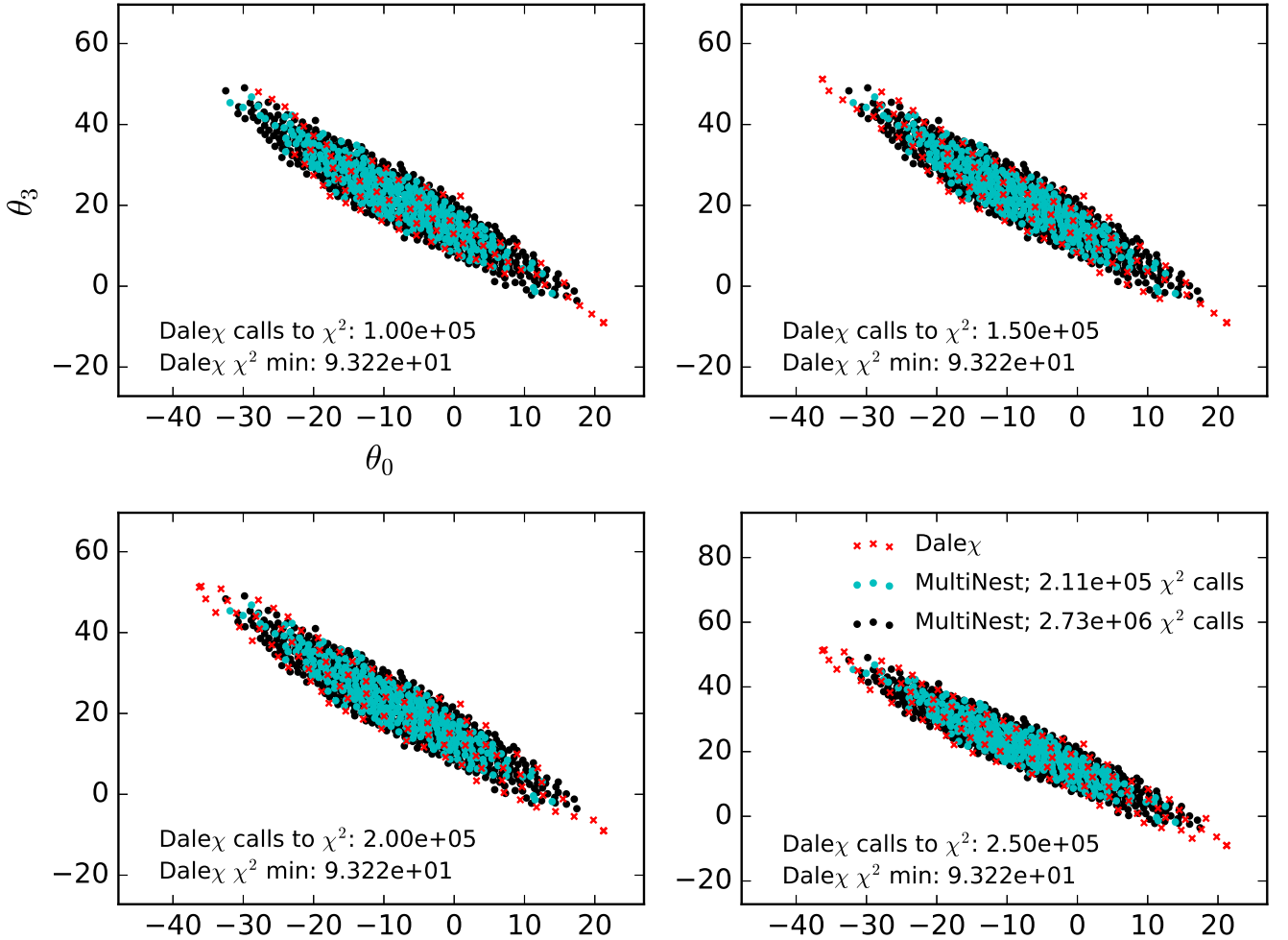


Fig. 7.— We compare the performance of DALE $\chi$  to MultiNest on a cartoon likelihood function constructed as described in Section 4. This cartoon is a 12-dimensional likelihood function in which several of the dimensions have non-Gaussian posteriors (see the marginalized one-dimensional posteriors presented in Figure 9). Each frame shows the progress of DALE $\chi$  after a specified number of calls to  $\chi^2$ . The black and cyan points indicate the projected Bayesian 95% credible limit discovered by MultiNest after a specified number of likelihood evaluations and are the same in all frames. Note the difference between the limits discovered by the rapid convergence MultiNest and the slow convergence MultiNest.

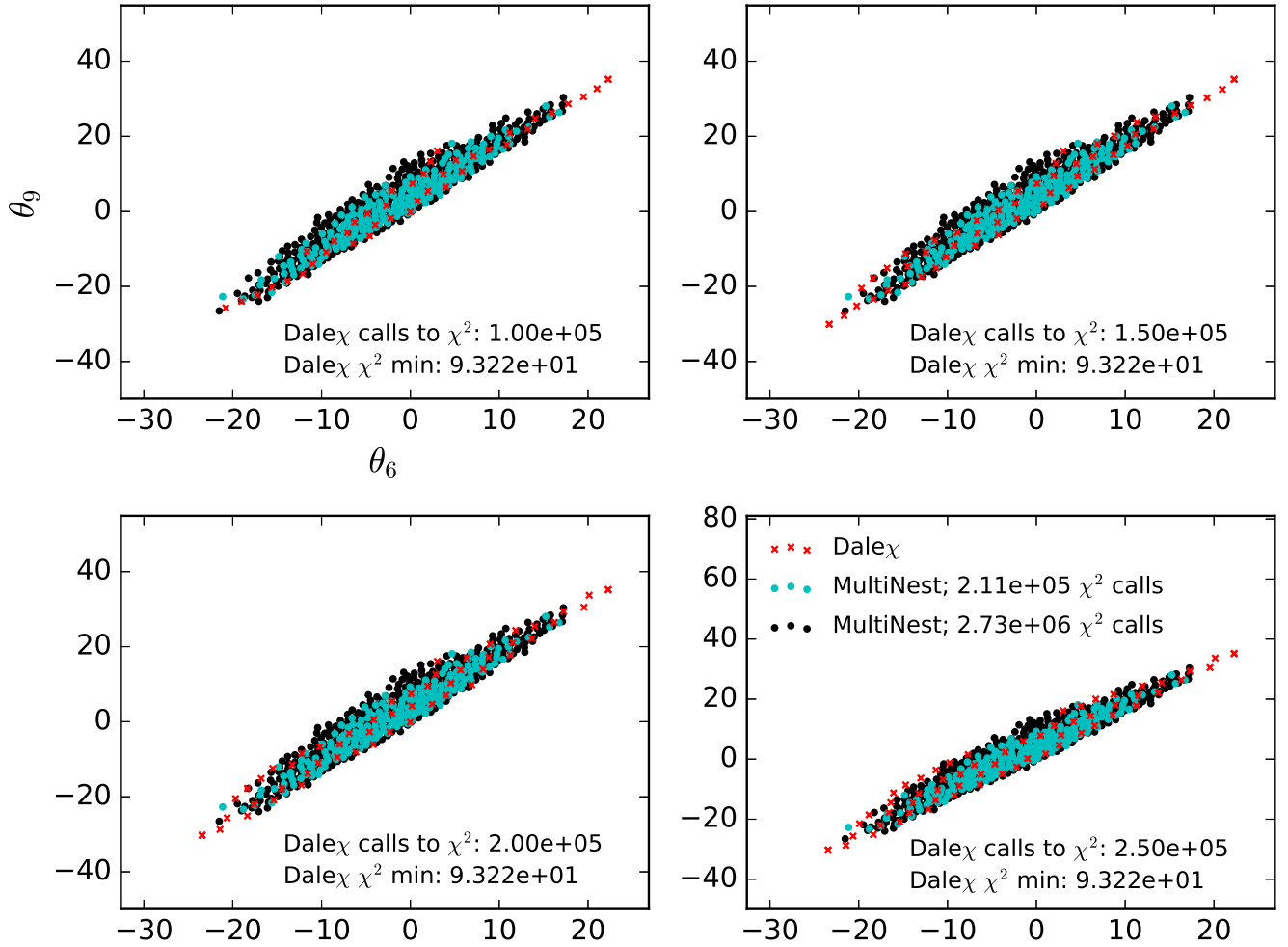


Fig. 8.— As Fig. 7 but for a different two dimensional subspace.

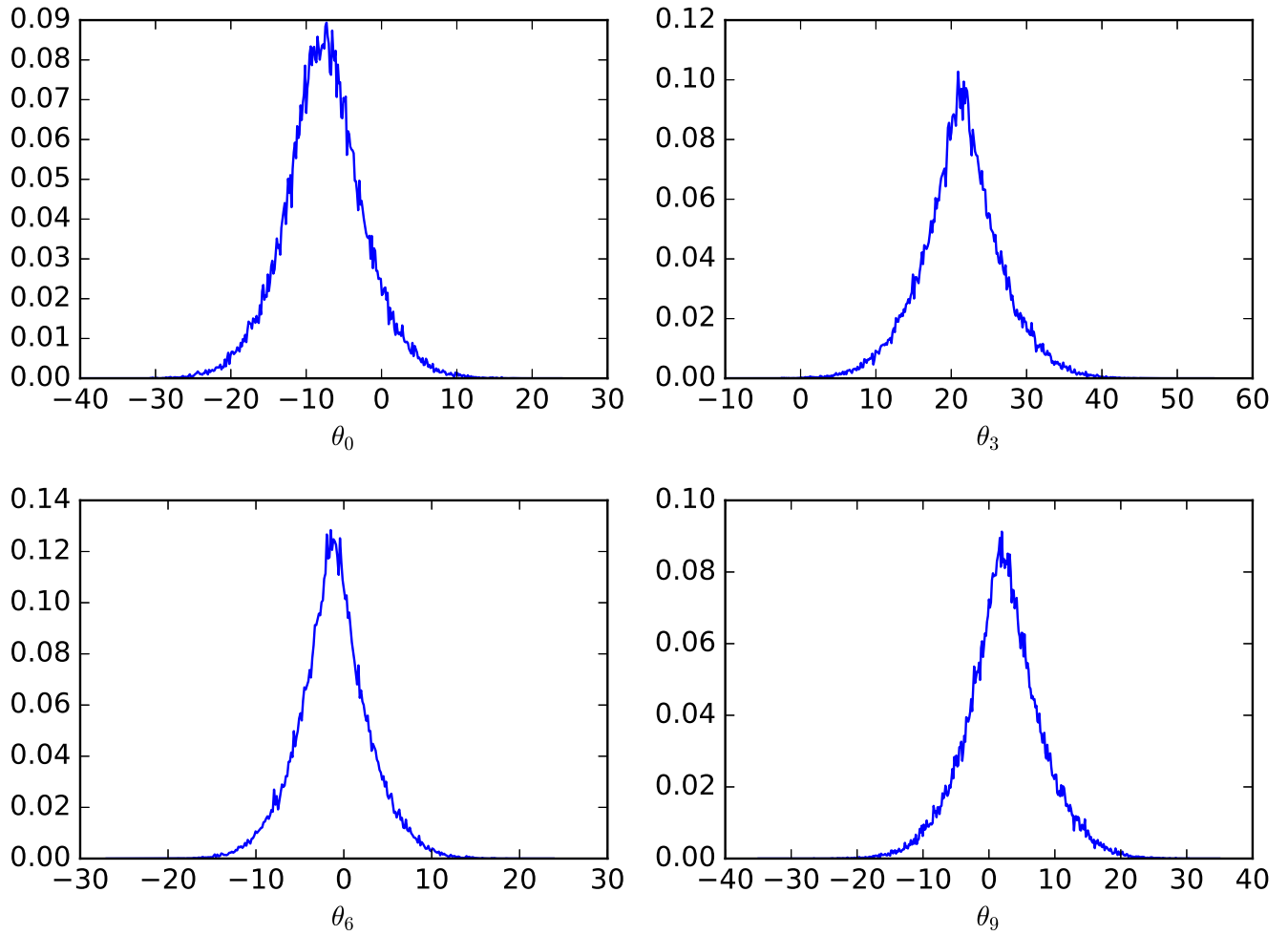


Fig. 9.— Four of the twelve marginalized one-dimensional posteriors of the likelihood function plotted in Figures 5, 7, and 8 as found by MultiNest. Note the non-Gaussianity of the wings of the posteriors.

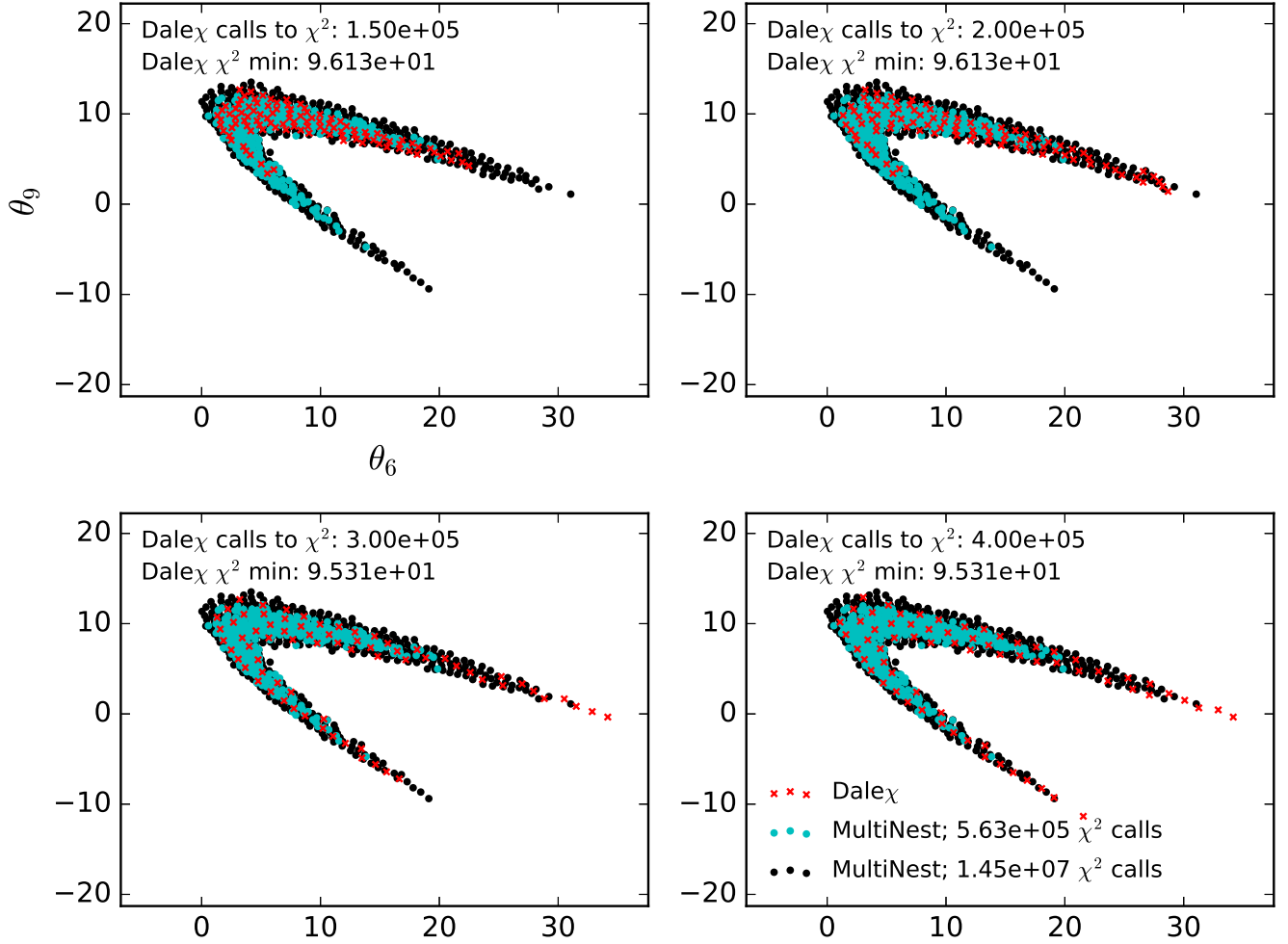


Fig. 10.— As Fig. 7 but for the curved degeneracy likelihood function.

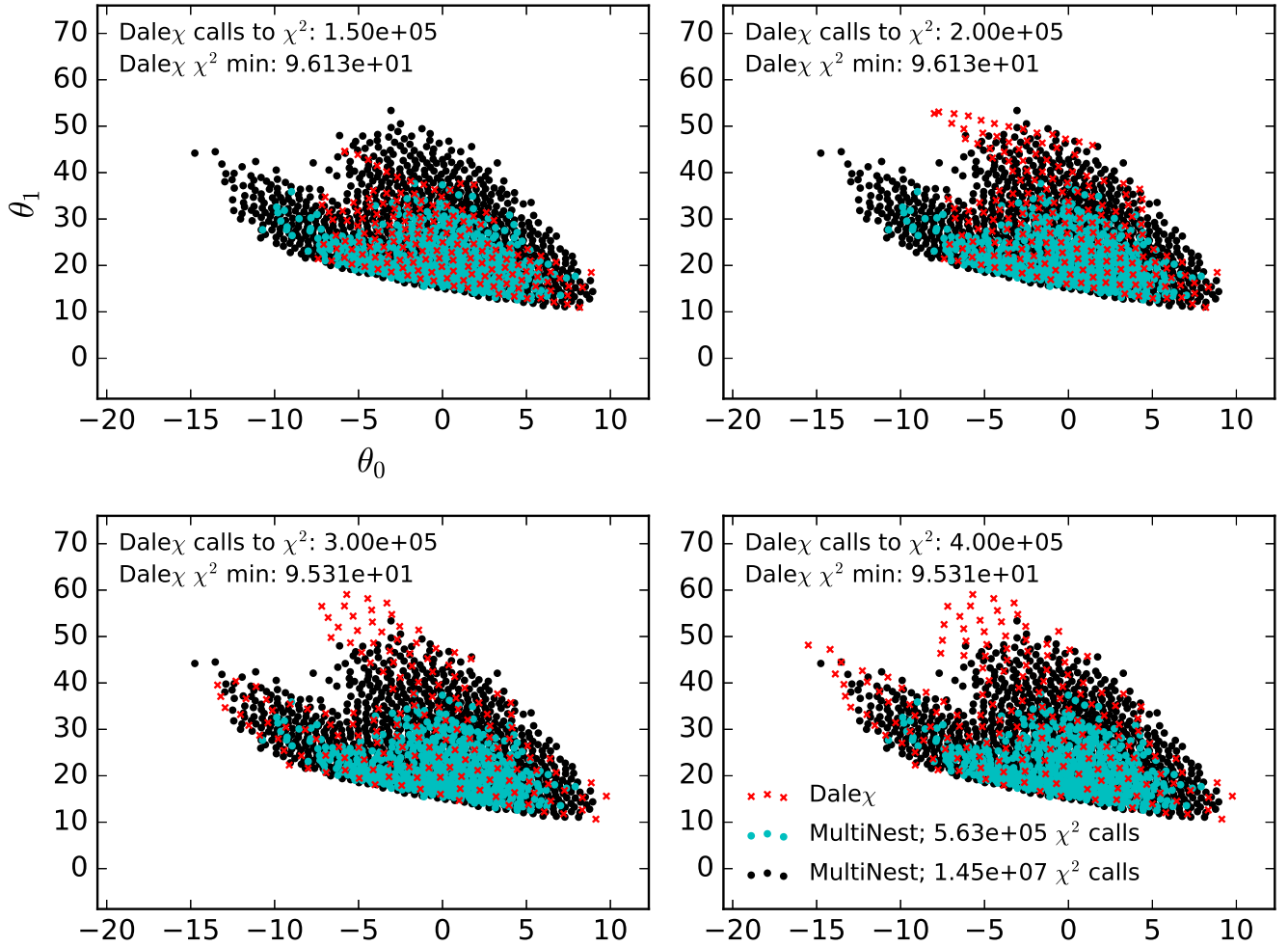


Fig. 11.— As Fig. 10 but for a different two dimensional subspace.

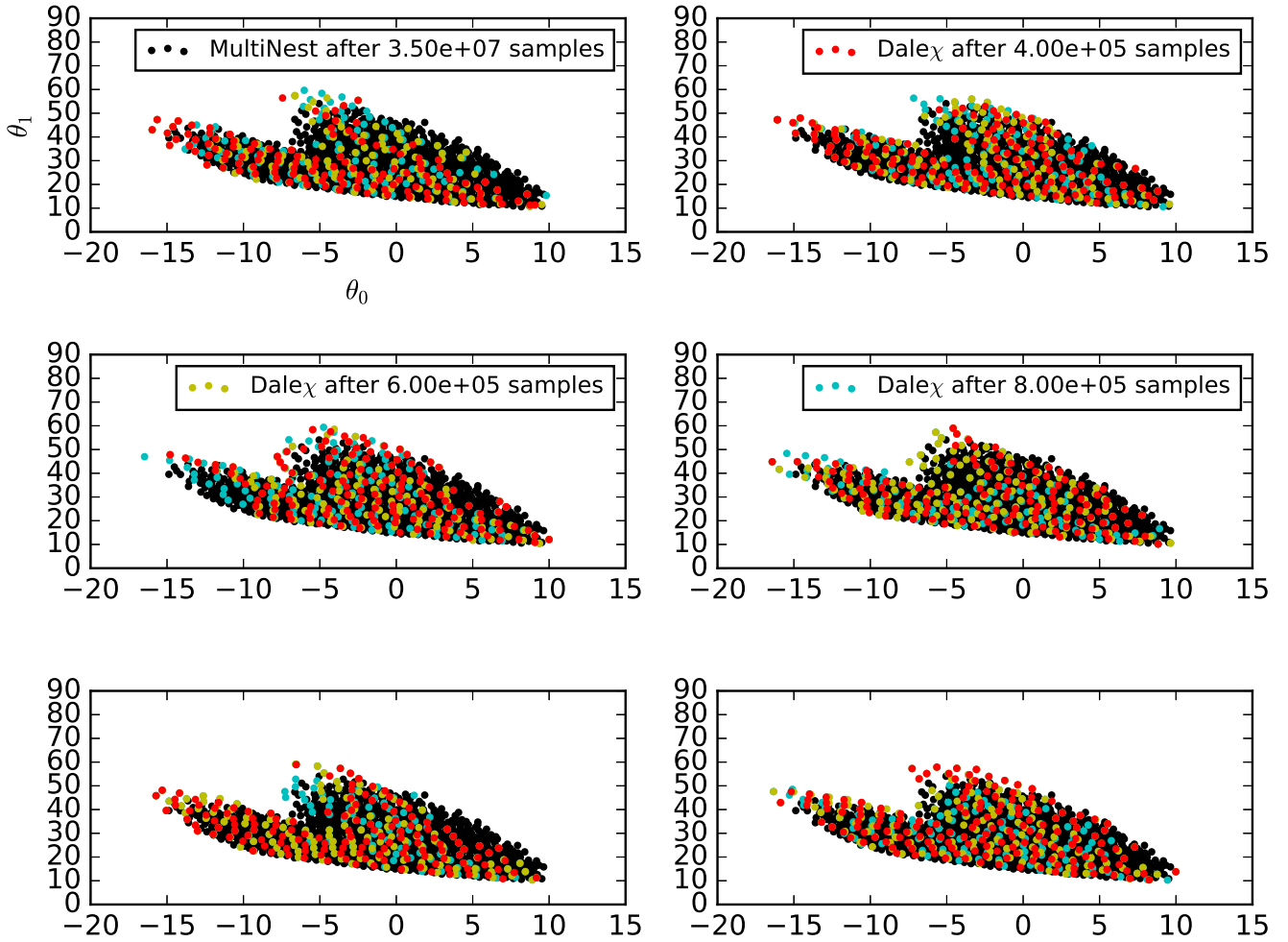


Fig. 12.— Each panel represents a different DALE $\chi$  run with a different seed provided to the random number generator. The colors show the 95% confidence limit as found at different stages during the run’s history. The likelihood function is the same as that tested in Figures 10 and 11.

## 5.2. Efficiency of Sampling

Figure 13 shows the distribution points evaluated by DALE $\chi$  and MultiNest as a function of  $\chi^2$  when run on the highly curved likelihood function from Figures 6, 10, and 11. Note how DALE $\chi$  places a much larger fraction of its points on the  $\chi^2 = \chi_{\text{lim}}^2$  contour than does MultiNest. This should not be surprising as DALE $\chi$  was designed with limit-finding in mind. MultiNest is a code designed to calculate the integrated Bayesian evidence of a model. The fact that one can draw confidence limit contours from the samples produced by MultiNest is basically a happy side-effect of this evidence integration. Because it focuses on finding the contours directly, rather than integrating the posterior over parameter space, DALE $\chi$  is highly efficient at both the optimization and characterization tasks, i.e. the key elements of determining the best fit and its error estimation, as designed. This tradeoff between obtaining the two critical elements for parameter estimation versus mapping the posterior more broadly (and sometimes less accurately) is the root cause of the vastly improved speed demonstrated by DALE $\chi$ .

## 6. Conclusions

As our exploration of the universe becomes increasingly detailed and precise, the range of astrophysical, instrumental, and theoretical systematics that must be taken into account require high dimensional parameter spaces. Accurate estimation of the desired parameters needs fast and robust techniques for exploring high dimensional, frequently nonlinearly degenerate and non-Gaussian joint probability distributions. DALE $\chi$  is designed with two paramount purposes: *optimization* – finding the probability maximum, and *characterization* – mapping out a specified confidence limit  $\chi_{\text{lim}}^2$ .

We have shown that its blend of strategies: to obtain robustly the maximum, search for multiple maxima, focus on the desired confidence limit, and efficiently map it out along multiple dimensions by seeking points distinct from established points and tracing along curved degeneracies – can be quite successful. This diversity is important in cases where the likelihood is not known to be close to Gaussian but may have complicated, degenerate/highly curved/non-Gaussian, or multimodal properties. DALE $\chi$  exhibits excellent completeness and purity: it routinely finds viable regions of the posterior missed by Markov Chain Monte Carlo or MultiNest nested sampling, and it focuses its search in the critical areas rather than accreting points not directly relevant to the maximum or confidence limit.

In the low, four dimensional test case, DALE $\chi$  had better purity (i.e. accuracy of confidence limit estimation) but MultiNest was faster. By the more common case of a 12 dimensional parameter space, DALE $\chi$  achieve a speed advantage of nearly a factor of 10. Note that DALE $\chi$  looks for a user specified confidence limit, e.g. 95%; if the user wants multiple confidence limits, e.g. also 68%, then it is true that DALE $\chi$  must be rerun while MultiNest can generate it from its original results. However, the factor 10 speed advantage (coupled with that the first run of DALE $\chi$  will have



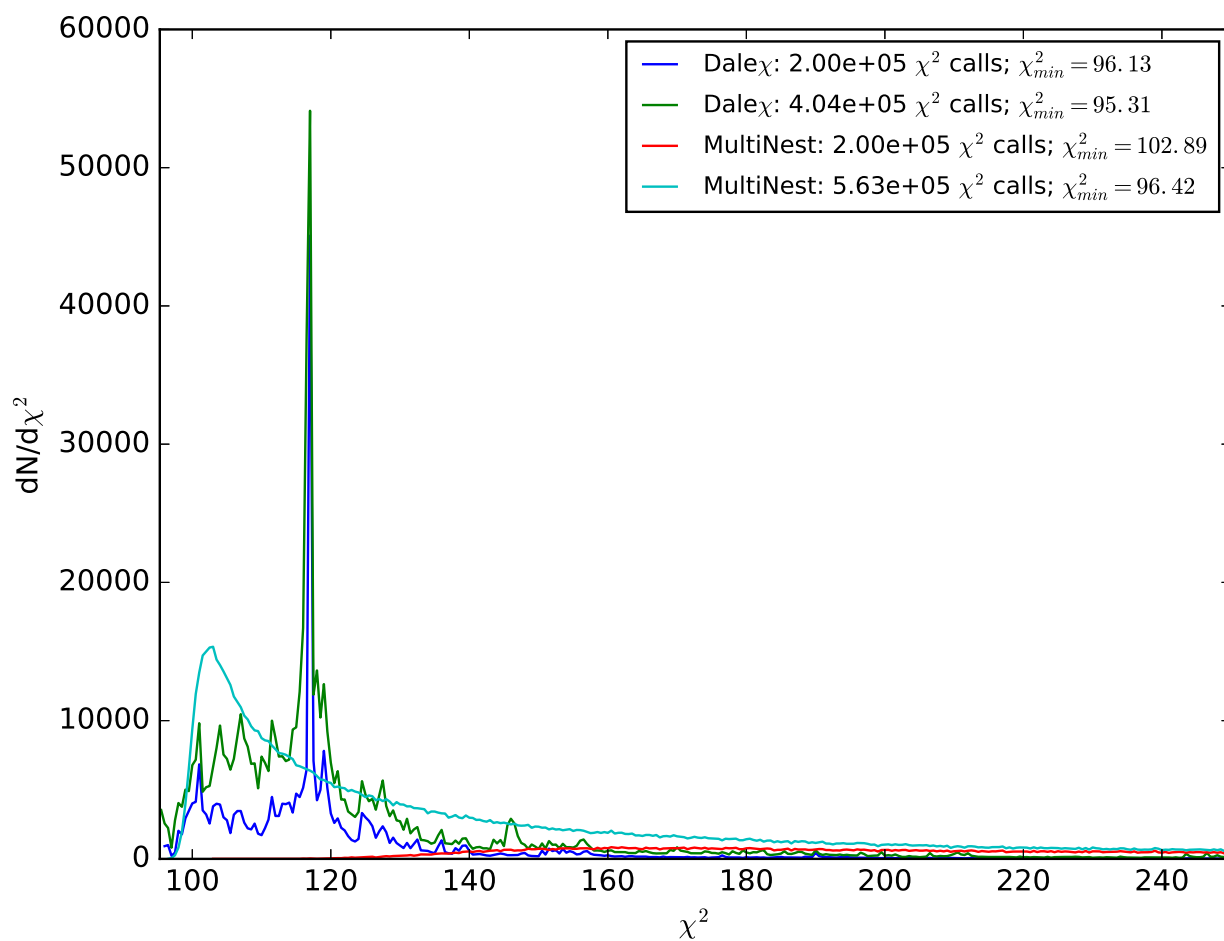


Fig. 13.— The distribution of points sampled by DALE $\chi$  and MultiNest as a function of  $\chi^2$  value at different points during the runs' histories.

already completed the optimization step of finding the probability maximum), means that even for calculation of several confidence limits the speed advantage remains. With higher dimensional parameter spaces we expect the advantage to become even stronger (unless the likelihood function is close to Gaussian in all dimensions). A preliminary test on a 16-dimensional likelihood function showed a factor of 75 speed advantage to DALE $\chi$  over MultiNest. At a minimum, DALE $\chi$  can function as a highly efficient “pre-burner” for sampling methods.

Despite its current success, several areas remain open for further improvement to DALE $\chi$ . It seems likely that the  $\chi^2_{\min}$  refinement procedure could be improved in efficiency. The convergence criterion is local and somewhat heuristic; potentially a more robust condition could further improve performance. Note that DALE $\chi$  is open source so we encourage interested researchers to explore its use and join in to make it even better. It serves as a new tool in our quest to estimate accurately the parameters of our universe.

The Direct Analysis of Limits via the Exterior of  $\chi^2$  (DALE $\chi$ ) has as its mission to look for the terminus of the desired confidence region, the boundary separating it from the unfavored exterior parameter space. One could say that in its quest to keep only the viable models, by terminating the exterior, the watchword for DALE $\chi$  is ex-terminate. Finally, DALE $\chi$  is available at <https://github.com/danielsf/Dalex.git>.

We thank Alex Kim for helpful discussions. SD is supported by the LSST Coropration (<https://www.lsst.org>). EL is supported in part by the Energetic Cosmos Laboratory and by the U.S. Department of Energy, Office of Science, Office of High Energy Physics, under Award DE-SC-0007867 and contract no. DE-AC02-05CH11231.

### A. Fitting ellipsoids to samples

In order to prevent DALE $\chi$  from wasting time exploring regions of  $\chi^2 \leq \chi^2_{\lim}$  parameter space that have already been discovered, we require a construct that represents the region in parameter space bounded by a set of sampled points  $\{\vec{P}\}$  in such a way that we can quickly determine whether a new point  $\vec{\theta}$  is inside (in which case, DALE $\chi$  is doubling back on itself and must be corrected) or outside (in which case, DALE $\chi$  has discovered a new region of  $\chi^2 \leq \chi^2_{\lim}$ ) that region. The simplest way to represent such a region would be to construct a hyperbox whose edges are at the extremal values of each cardinal dimension as described by the points in  $\{\vec{P}\}$ . This, however, would involve erroneously designating too much unexplored volume as “explored”. Consider the case where the samples in  $\{\vec{P}\}$  are tightly clustered around the diagonal of a three-dimensional cube. Using the extremal values of  $\{x, y, z\}$  in  $\{\vec{P}\}$  would entail designating the entire cube as “explored” when, in fact, we know almost nothing about the cube as a whole. We therefore have implemented the following algorithm to take the set of sampled points  $\{\vec{P}\}$  and determine a nearly volumetrically minimal ellipsoid containing  $\{\vec{P}\}$ . This is the region that we designate as “explored” for the purposes of DALE $\chi$ .

The first step in designating our ellipsoid is to find the center of the ellipsoid. To do this, we find the extremal values of  $\{\vec{P}\}$  in the cardinal dimensions ( $\{x, y, z\}$  in a three-dimensional example). We determine the geometric center of  $\{\vec{P}\}$  to be at the point

$$\vec{\gamma}_i = 0.5 \times (P_{i,\max} + P_{i,\min}) , \quad (\text{A1})$$

where  $i$  is an index over the  $D$  dimensions of parameter space. We set the center of our ellipsoid to be the point in  $\{\vec{P}\}$  that is closest to  $\vec{\gamma}$  where, for the purposes of this determination, we calculate the distance  $\delta$  to be

$$\delta^2(\vec{\gamma}, \vec{\theta}) = \sum_i^D \left( \frac{\gamma_i - \theta_i}{P_{i,\max} - P_{i,\min}} \right)^2 . \quad (\text{A2})$$

Note the dimensions in parameter space are normalized by the denominator so each dimension runs from 0 to 1, putting them on an equal footing.

Once we have found the center of our ellipsoid, which we call  $\vec{c}$ , we must find the basis directions along which the axes of our ellipsoid will be reckoned. Initialize an empty set of  $D$ -dimensional vectors  $\{\vec{B}\}$ . Iterate over the points  $\vec{p}$  in  $\{\vec{P}\}$ . For each point, calculate the vector

$$\vec{\mu} = \vec{p} - \vec{c} - \sum_{\vec{\beta}} \vec{\beta} [(\vec{p} - \vec{c}) \cdot \vec{\beta}] \quad (\text{A3})$$

where  $\vec{\beta}$  are the orthogonal unit vectors already in  $\{\vec{B}\}$ . Select the vector  $\vec{\mu}$  with the largest L2 norm. Normalize this vector and add it to  $\{\vec{B}\}$ . Continue until there are  $D$  vectors in  $\{\vec{B}\}$ . In this way, we select a set of orthogonal basis vectors along which the points  $\{\vec{P}\}$  are extremally distributed.

Now that we have determined the center and bases of the coordinate system in which we will draw our ellipsoid, we must find a set of lengths  $\{r\}$  of the ellipsoid's axes such that all of the points in  $\{\vec{P}\}$  are contained in the ellipsoid. The length of the vectors  $\vec{\mu}$  chosen to populate  $\{\vec{P}\}$  provide a good first guess at this set of lengths, but they are only guaranteed to describe the ellipsoid if the points in  $\{\vec{P}\}$  really are only distributed along the vectors  $\{\vec{B}\}$  and centered on  $\vec{c}$ . Since this seems unlikely, we adopt an iterative scheme to adjust the lengths  $\{r\}$  so that our ellipsoid contains  $\{\vec{P}\}$ . For each of the points  $\vec{p}_{\text{bad}}$  that are not contained by the ellipsoid, find the basis vector  $\vec{\beta}_{\text{worst}}$  that maximizes

$$\epsilon = \frac{|(\vec{p}_{\text{bad}} - \vec{c}) \cdot \vec{\beta}|}{r_{\vec{\beta}}} \quad (\text{A4})$$

where  $r_{\vec{\beta}}$  is the length in  $\{r\}$  corresponding to that basis vector. Keep track of the number of points  $\vec{p}_{\text{bad}}$  for which any given basis vector is  $\vec{\beta}_{\text{worst}}$ . Select the basis direction that is  $\vec{\beta}_{\text{worst}}$  for the most points  $\vec{p}_{\text{bad}}$ . Multiply the corresponding  $r$  by 1.1. Continue this process until there are no more points  $\vec{p}_{\text{bad}}$ .

This algorithm is implemented in `include/ellipse.h` and `src/utils/ellipse.cpp` in our code base.

## REFERENCES

- Bentley, J. L. 1975, Communications of the Association for Computing Machinery **18**, 509
- Brooks, S. and Gelman A. 1998, Journal of Computational and Graphical Statistics **7**, 434
- Bryan, B., 2007, Ph.D. thesis, Carnegie Mellon University, <http://reports-archive.adm.cs.cmu.edu/anon/ml2007/abstracts/07-122.html>
- Bryan, B., Schneider, J., Miller, C. J., Nichol, R. C., Genovese, C., and Wasserman, L., 2007, Astrophys. J. **665**, 25
- Casella, G. and George, E. I. 1992, The American Statistician **46**, 167
- Daniel, S.F., Connolly, A.J., and Schneider, J. 2014, ApJ **794** 38 [[arXiv:1205.2708](https://arxiv.org/abs/1205.2708)]
- Feroz, F. and Hobson, M.P. 2008, Monthly Notices of the Royal Astronomical Society **384**, 449
- Feroz, F., and Hobson, M. P., and Bridges, M. 2009, Monthly Notices of the Royal Astronomical Society **398**, 1601
- Feroz, F., Hobson, M. P., Cameron, E., and Pettitt, A. N., [arXiv:1306.2144](https://arxiv.org/abs/1306.2144)
- Gelman, A. and Rubin, D. 1992, Statistical Science **7**, 457
- Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P. 1983, Science **220**, 671
- Marsaglia, G. 2003, Journal of Statistical Software, **8** 14
- Nelder, J. A. and Mead, R. 1965, The Computer Journal, **7** 308
- Planck Collaboration XVI 2014, Astronomy & Astrophysics **571** 16 [[arXiv:1303.5076](https://arxiv.org/abs/1303.5076)]
- Press, W. H., Teukolsky, S. A., Vetterling, W. T. and Flannery, B. P. 2007, “Numerical Recipes” (3d edition), Cambridge University Press, 2007
- Rasmussen, C. E. and Williams, C. K. I., 2006, “Processes for Machine Learning” <http://www.GaussianProcess.org/gpml>
- Skilling J. 2004, in “Bayesian Inference and Maximum Entropy Methods in Science and Engineering”, eds. Fisches R., Preuss R., von Toussaint U., AIP Conf. Proc. 735, 395
- Wilks, S. S. 1938, Annals of Mathematical Statistics, **9** 60