Actively Learning Level-Sets of Composite Functions

Brent Bryan BRENT@GOOGLE.COM
Google Inc., 4720 Forbes Ave., Pittsburgh, PA 15213

Jeff Schneider SCHNEIDE@CS.CMU.EDU
Carnegie Mellon University, 5000 Forbes Ave., Pittsburgh, PA 15213

Abstract

Scientists frequently have multiple types of experiments and data sets on which they can test the validity of their parameterized models and locate plausible regions for the model parameters. By examining multiple data sets, scientists can obtain inferences which typically are much more informative than the deductions derived from each of the data sources independently. Several standard data combination techniques result in target functions which are a weighted sum of the observed data sources. Thus, computing constraints on the plausible regions of the model parameter space can be formulated as finding a level set of a target function which is the sum of observable functions. We propose an active learning algorithm for this problem which selects both a sample (from the parameter space) and an observable function upon which to compute the next sample. Empirical tests on synthetic functions and on real data for an eight parameter cosmological model show that our algorithm significantly reduces the number of samples required to identify the desired level-set.

1. Introduction

Scientists frequently have multiple types of experiments and data sets on which they can test the validity of their parameterized models and the plausible or optimal regions for the model parameters. One task that can be considered is that of computing the parameter setting (from a pre-defined model parameter space) which maximizes the likelihood of all the observations given the models. However, this calculation does not determine whether or not the derived parameter setting is consistent with the data given the models. Instead, a more prudent approach is to compute the set of model parameters (from the parameter space) which cannot be statistically rejected by the combination of the observed data and theoretical models.

When given a single model and data set pair, computation of the feasible regions of parameter space can be done by performing a simple hypothesis test for all points in the space; that is, we are interested in the regions of parameter space where the null hypothesis — that the data was generated by the model — cannot be rejected at some specified confidence level. Extending this to the multiple model and data setting, we are interested in determining regions of parameter space where we cannot reject the hypothesis that each of the data sets was generated by its respective model at a given confidence level.

For example, when determining the spatial location of a disease outbreak, a researcher might use information derived from medical records (e.g. hospital admits), as well as sales of over the counter and prescription medications (Shmueli & Fienberg, 2006). Note that the presence (or lack thereof) of a single indicator may be enough to accept or reject a single hypothesis, resulting in increased data efficiency. Specifically, if there are many hospital admits from a single locality, the probability of disease is extremely high regardless of the over the counter and prescription drug sales. Moreover, while we believe that the underlying cause affects each of the signals we observe, we do not necessarily believe that the signals themselves are correlated. For instance, colds result in significant over the counter sales with few hospital visits or prescription sales. However, anthrax attacks will affect all three data streams.

There are many other examples of the multiple model setting. Here, we focus on finding $1-\alpha$ confidence regions for statistical analyses involving multiple related data sets. Traditionally, the combination of statistical evidence has been achieved in the sciences in a somewhat ad-hoc fashion. For instance, a joint analysis can be performed by (loosely) intersecting the confidence regions of several studies. Additionally, results from one publication might be used to guide the selection of parameters in future experiments, possibly in the form of a prior.

Appearing in *Proceedings of the 25^{th} International Conference on Machine Learning*, Helsinki, Finland, 2008.

A more rigorous and efficient approach is to consider multiple experimental sources of evaluation simultaneously and choose samples in light of their contribution to the combined target function. This target function is the composition of the "observable" test functions: one for each data set and model pair. We assume that the observable functions share the same parameter space, but are functionally independent. As such, hierarchical models do not apply. Moreover, whereas multi-task learning problems are based on learning the commonality between the constituent models, the task of locating confidence regions benefits from the discrepancies between the models to efficiently accept or reject a parameter vector. While in theory we could check each point in the parameter space to determine whether or not it should be included within our 1–α confidence region, in practice each experiment is too expensive.

As such, we develop active learning algorithms to learn the confidence regions. Active learning using informed choices of future experiments has long been known to drastically decrease a problem's sample complexity (Angluin, 1988). Many sampling heuristics have been developed to learn either the entire target function (e.g. MacKay (1992); Guestrin, C., et al. (2005)) or some feature of the target function, such as its level sets (e.g. Bryan, B., et al. (2005); Ramakrishnan, N., et al. (2005)). While we cannot directly observe the value of the target function, we can use the observable functions to infer its value. By measuring all observable functions at a particular parameter setting, we can compute the value of the target function, reducing the problem to a standard active learning problem. However, such an approach disregards any strong evidence provided by a single statistical test, and hence may result in extraneous sampling of the remaining statistical models.

Rather, we are interested in active learning algorithms which use information about each observable function to learn some composite target function. In Section 2, we propose a heuristic for actively learning level sets of composite functions of sums for continuous valued input spaces. In Section 3, we show that this heuristic performs the level-set discovery task more efficiently than both random and sequential sampling of the constituent functions using state of the art heuristics. In Section 4, we discuss how the task of finding joint confidence regions can be formulated as a level set problem, where the target function is the sum of several observable functions. Section 5 concludes by demonstrating the computation of 95% confidence regions for eight cosmological parameters using our algorithm.

2. Active Learning Algorithm

Let f be a target function we are interested in learning on the domain $\Theta \subseteq \mathbb{R}^d$. Suppose that f is the linear combination of m observable functions, f_i $(i = 1, \ldots, m)$. Without loss of generality, we can drop the coefficients from the summation (as they can be included in the f_i's) and write $f(\theta) = \sum_{i=1}^{m} f_i(\theta)$ for all $\theta \in \Theta$. We are now interested in finding the level set, $\mathcal{S}$, of f at the threshold t:

$$\mathcal{S} = \left\{ \theta \in \Theta \,\middle|\, \sum_{i=1}^{m} f_i(\theta) = f(\theta) = t \right\}.$$

In general, computing the value of each f_i may not incur the same cost. However, we begin by assuming that the costs are similar, and hence try to minimize the total number of samples of observable functions required to accurately estimate $\mathcal{S}$. Moreover, we assume that f cannot be directly sampled, and that neither f nor any of the f_i's is invertible. That is, the only way to estimate a level-set of f is to sample points from the f_i's and infer f. As we will see in Section 4, this formulation accurately mimics combining p-values using Fisher's method, as the method for finding the individual p-values may be entirely unknown.

We must now determine how best to choose samples both among and within the f_i's. Ideally, we want to sample the observable function f_i at the point $\tilde{\theta} \in \Theta$ which best increases our prediction accuracy (e.g. whether another point is above or below the threshold) over f. Since the parameter space is continuous and multi-dimensional, we cannot afford to test all possible points and observable functions.

Instead, we model each of the observable functions independently given the current samples taken from that function, as illustrated in Figure 1. For each experiment, we randomly select a small subset of the parameter space (usually 1000 points drawn uniformly at random, although other distributions are possible based on domain knowledge) and choose the best point and observable function pair upon which to experiment from among these candidates. We find the value of the observable function at the selected point and add it to the data set used to model that function. The process is then repeated.

There are several methods one could use to model each of the f_i's, notably some form of parametric regression. However, we chose to approximate the f_i's using Gaussian process regression, as other forms of regression may over smooth the data, ignoring subtle features of the function that may become pronounced with more data. While much work has been done studying Gaussian processes, we only touch on the basic concepts here; we refer interested readers to Cressie (1991); Rasmussen and Williams (2006).

Gaussian processes are non-parametric forms of regression. Predictions for unobserved points are computed by using a weighted combination of the function values for those points which have already been observed, where a distance-based kernel function is used to determine the relative weights. These distance-based kernels generally weight

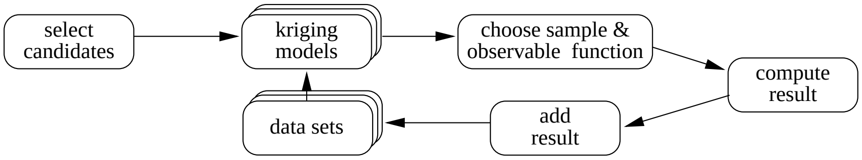


Figure 1. Outline of our sampling algorithm. Given an initial set of points (typically empty), we randomly select a set of candidates and score them using a set of Gaussian process models. The best scoring point and observable function pair is chosen, and we evaluate the selected observable function at the given point. This data is added to the corresponding data set.

nearby points significantly more than distant points. Thus, assuming the underlying function is continuous, Gaussian processes will perfectly describe the function given an infinite set of unique data points. While, in many applications the assumption of continuity is violated, Gaussian processes have been successfully used to model response surfaces in many domains with limited smoothness guarantees (Cressie, 1991; Santner et al., 2003).

In this work we use ordinary kriging (Cressie, 1991), which assumes a linear semivariance as a function of distance, as it is both data and computationally efficient. While other forms of Gaussian Processes could be used — most notably adaptive kernel methods (e.g. Kersting, K. et al. (2007)) — we find that a learned model based upon a simple kriging approximator performs well in practice and ensures that we do not spend more time computing the next sample than we do running the experiment.

Regardless of the kernel used, Gaussian processes predict that the value of a target point, $\tilde{\theta}$, will be Normally distributed with a mean and variance ($f_i(\tilde{\theta})$ and $\sigma_i^2(\tilde{\theta})$, respectively) given by:

$$f_i(\tilde{\theta}) = \bar{f}_i + \vec{\Sigma}_{i,\tilde{\theta}}^T \mathbf{\Sigma}_i^{-1} \vec{\mathcal{F}}_i \tag{1}$$

$$\sigma^2(\tilde{\theta}) = \vec{\Sigma}_{i,\tilde{\theta}}^T \mathbf{\Sigma}_i^{-1} \vec{\Sigma}_{i,\tilde{\theta}} \tag{2}$$

where $\mathcal{T}_i$ is the set of observed experiments of f_i,

$$\bar{f}_i = \frac{1}{|\mathcal{T}_i|} \sum_{j=1}^{|\mathcal{T}_i|} f_i(a_j),$$

$$\mathcal{F}_i[j] = f_i(\theta_j) - \bar{f}_i,$$

$\mathbf{\Sigma_i}$ denotes the covariance matrix between the elements of $\mathcal{T}_i$, and $\vec{\Sigma}_{i,\tilde{\theta}}$ is the covariance vector between elements of $\mathcal{T}_i$ and $\tilde{\theta}$.

For a set of n_i observed points ($|\mathcal{T}_i| = n_i$), prediction with a Gaussian process requires $O(n_i^3)$ time, as a $n_i \times n_i$ linear system of equations must be solved. However, for many Gaussian processes — and ordinary kriging in particular — the correlation between two points decreases as a function of distance. Thus, the full Gaussian process model is approximated well by a local Gaussian process in which only the k nearest neighbors of the query point are used, for some fixed constant k. This reduces the computation time to $O(k^3 + k\log(n_i))$ per prediction. Here, we let $k = 1000$.

2.1. Choosing Experiments

Given this active learning framework, we must now decide how to choose sample / observable function pairs. We consider the following heuristics:

Random One of the candidate points and an observable function pair is chosen uniformly at random. This method serves as a baseline for comparison of the other heuristics.

Variance The candidate point and observable function pair which has the highest predicted variance (out of all the candidate / observable function pairs) is selected. Using model variance to pick the next experiment is common for active learning methods whose goal is to map out the target function over a parameter space (MacKay, 1992; Guestrin, C., et al., 2005). In particular, (Guestrin, C., et al., 2005) showed that greedily picking experiments based upon model variance performs nearly as well as using a mutual information heuristic when learning the target over the entire parameter space; this is significant, as the mutual information heuristic can be shown to be $(1 - 1/e)$ optimal (Guestrin, C., et al., 2005). Since variance is closely related to distance for kriging models, this heuristic samples points which are far from their nearest neighbors. However, when searching for level-sets, we are less interested in the function away from the level-set boundary, and instead want to focus our sampling resources near the predicted boundary. In particular, sampling based solely on variance results in substantially worse performance than heuristics that concentrate on the function level-set (Bryan, B., et al., 2005).

Information Gain Information gain is a common myopic metric used in active learning. Computing the information gain over the whole state space for each observable function provides an optimal 1-step experiment choice. In some discrete or linear problems this can be done, but it is intractable for continuous non-linear spaces. As such we do not consider a traditional information gain heuristic, but rely on efficient point estimates which act as proxies for global information gain.

Sequential-Straddle As noted in Section 1, the problem can be simplified to a standard active learning problem if one sequentially samples each of the observable functions in order to directly compute f. (Bryan, B., et al., 2005) showed that in a setting where experiments yield the (approximately) true values of the target function, a good heuristic for level set identification is the straddle heuristic: $\text{straddle}(\tilde{\theta}) = 1.96\sigma^2(\tilde{\theta}) - |f(\tilde{\theta}) - t|$. This heuristic balances the need to explore uncertain parts of parameter space, with the desire to refine the model's estimate around those regions already known to be close to the level-set boundary; the constant 1.96 ensures that points with negative scores are far from the desired level set with at least a 95% probability. This heuristic leverages the straddle heuristic by choosing the candidate point with the highest combined straddle score,

$$\text{combined-straddle}(\tilde{\theta}) = 1.96 \sum_{i=1}^{m} \sigma_i^2(\tilde{\theta}) - \left| \sum_{i=1}^{m} f_i(\tilde{\theta}) - t \right|, \tag{3}$$

and then sequentially sampling all m observable functions at this point.

Variance-Straddle While (Bryan, B., et al., 2005) showed that the straddle heuristic works well when directly sampling the target function, we can hope to do better by considering the output from each observable function individually. For instance, if a sample point results in a very large value for one of the observable functions, it may be unlikely that the results of the other f_i's will be such that the resulting value of f is near the level-set. In particular, when dealing with χ^2 models (see Section 4), we know that $f_i \geq 0$ for all i. Thus, if a single f_i is greater than the level-set boundary, the target function will also be greater than the level-set boundary, and hence it may be more efficient to sample elsewhere. This heuristic simply chooses the next sample from among the candidates based on the combined-straddle score, and then selects the observable function with the largest variance at that point.

Variance-MaxVarStraddle Finally, we consider a variant of the straddle heuristic. This heuristic tries to mimic the information gain of choosing a particular point and observable function pair. Note that after observing a point, the variance of the kriging model is effectively zero at that point (since we have set c to be a very small positive value). The original straddle heuristic balances the expected gain in the model fit ($\sigma(\tilde{\theta})$) with the expected distance of the point to the level-set boundary.

However, with the multiple model formulation, we do not expect the model variance to decrease by $\sigma^2(\tilde{\theta}) = \sum_{i=1}^{m} \sigma_i^2(\tilde{\theta})$, but rather by $\sigma_i(\tilde{\theta})$ where f_i is the observable function we pick. Thus, a more accurate proxy for the information gain of a candidate point and observable function pair is:

$$\begin{aligned} &\text{variance-maxvarstraddle}(\tilde{\theta}) \\ &\qquad = \max_i \left\{ 1.96\sigma_i^2(\tilde{\theta}) \right\} - \left| \sum_{i=1}^{m} f_i(\tilde{\theta}) - t \right|. \end{aligned} \tag{4}$$

We choose the candidate point that maximizes this heuristic and the corresponding f_i.

3. Experiments

We now assess the accuracy with which our active learning model reproduces synthetic target functions for the sampling heuristics just described. This is done by computing the fraction of test points in which the predictive model (the sum of the kriging models associated with each observable function) agrees with the true target function about on which side of the threshold the test points lie. This process was repeated 20 times to account for variations due to the random nature of the candidate generation process. The first three target functions considered were sums of two observable functions, while the fourth was a sum of four observable functions. The kriging parameters for each model were computed *a priori* from the observable functions. The considered functions are:

Gaussian This problem consisted of determining the 95% acceptance region of two axis aligned perpendicular two dimensional Gaussian distributions centered at the origin. Both Gaussians had diagonal covariance matrices with on diagonal elements of 1 and 16. Since working in probability space results in many near-zero values, the problem was considered in log-space. As such, the target function was a 2 dimensional symmetric quadratic function, and the level-set was a circle centered at the origin. The range of the parameter space was ($\theta_1, \theta_2 \in [-3.4, 3.4]$)

Sin2D The second problem consists of finding where the two 2D sinusoidal observable functions

$$\begin{aligned} f_1(\theta_1, \theta_2) &= \sin(10\theta_1) + \cos(4\theta_2) - \cos(3\theta_1\theta_2) \\ f_2(\theta_1, \theta_2) &= \sin(10\theta_2) + \cos(4\theta_1) - \cos(3\theta_1\theta_2) \end{aligned}$$

sum to zero where $\theta_1, \theta_2 \in [0, 2]$. These observable functions were chosen because 1) the target threshold winds through the plot giving ample length to test the accuracy of the approximating model, 2) the boundary is discontinuous with several small pieces, 3) there is an ambiguous region around $(0.9, 1)$, where the true function is approximately equal to the threshold, and the gradient is small and 4) there are areas in the domain where the function is far from the threshold and hence we can see whether algorithms refrain from oversampling in these regions.

Table 1. Number of samples required to achieve a 99% accuracy on the Gaussian and SimpleSin2D tests, and a 90% accuracy on the Sin2D and 4-Sin2D tests based on 20 trials. The variance-maxvarstraddle heuristic consistently performs better than competitors.

	Gaussian	SimpleSin2D	Sin2D	4-Sin2D
random	> 1000	> 1000	> 1000	> 1000
variance	95.0±11.0	> 500	105.0±11.5	188.6±32.2
variance-straddle	89.5±5.0	157.9±12.3	90.4±9.0	72.5±12.0
sequential-straddle	76.2±3.5	150.3±6.5	87.0±7.3	98.1±14.0
variance-maxvarstraddle	71.7±3.3	127.3±6.8	82.9±10.2	54.9±16.9

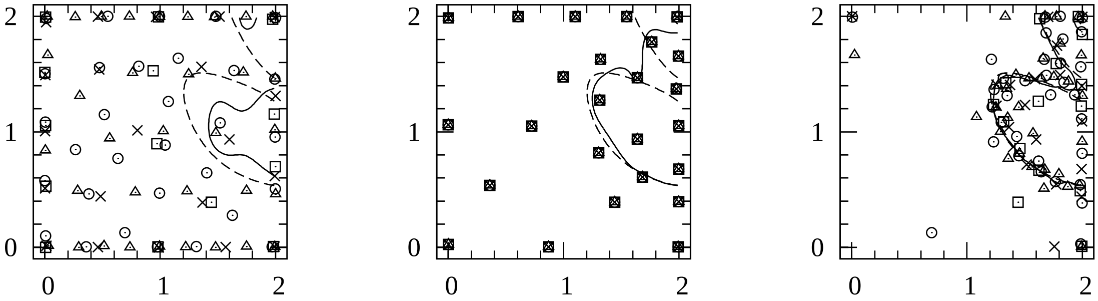


Figure 2. Predicted level-set (solid), true level-set (dashed) and experiments (squares, circle, triangles and x's) for the 4-Sin2D function after sampling 100 points using the Variance heuristic (left), the sequential-straddle heuristic (center), and the variance-maxvarstraddle heuristic (right).

SimpleSin2D This problem is a simplified version of the previous problem, where the observable functions

$$\begin{aligned} f_1(\theta_1,\theta_2) &= \sin(4\theta_1)+\cos(4\theta_2)-\cos(\theta_1\theta_2) \\ f_2(\theta_1,\theta_2) &= \sin(4\theta_2)+\cos(4\theta_1)-\cos(\theta_1\theta_2) \end{aligned}$$

were chosen to reduce the problem's semi-variances (again $\theta_1, \theta_2 \in [0:2]$). Since problems with large semi-variances result in large model variance estimates in the kriging models, such problems require extensive sampling to correctly identify function level-sets. Performance on this function highlights an algorithm's ability to quickly rule out portions of the function.

4-Sin2D This task consisted of finding where four 2D sinusoids sum to -2. The sinusoids chosen for this problem were similar to those of the SimpleSin2D problem:

$$\begin{aligned} f_1(\theta_1,\theta_2) &= \sin(4\theta_1)+\cos(2\theta_2)-\cos(3\theta_1) \\ f_2(\theta_1,\theta_2) &= \sin(2\theta_2-2)+\cos(2\theta_1)-\cos(3\theta_1) \\ f_3(\theta_1,\theta_2) &= \sin(3\theta_1\theta_2)+\cos(2\theta_1)+1 \\ f_4(\theta_1,\theta_2) &= \cos(\theta_1\theta_2)-\sin(\theta_1\theta_2) \end{aligned}$$

The resulting target function contains regions with both high and low derivatives near the specified threshold.

Classification accuracy results for the four tests are given in Table 1. variance-maxvarstraddle outperforms all of the other heuristics on each of the target functions. Unsurprisingly, the straddle-based heuristics beat the random and variance-weighted heuristics, as both the random and variance-weighted heuristics choose samples (roughly) uniformly throughout the parameter space, while the straddle-based heuristics focus on the level-set of interest. Additionally, the advantage of variance-maxvarstraddle over sequential-straddle grows as the number of observable functions increases, as the relative cost of a bad choice is increased. These results demonstrate that learning the models independently allows for better overall prediction.

One surprising result of our experimentation is that the sequential-straddle performs as well as the variance-straddle heuristic on the test functions which are sums of two observable functions. We believe that this result illustrates the fact that the variance-straddle heuristic is over estimating the importance of the variance component of the candidate points to the information gain of a point, while the fact that there are only two observable functions reduces the efficiency of the sequential-straddle heuristic only by a factor of two. The variance-straddle heuristic will be as likely to choose a candidate point where the predicted observable functions are moderate but equal, as it is to choose a point with a large predicted variance for one of the observable functions, and zero variance for the other observable functions. However, the second candidate has much more information than the first, as selecting the second candidate will give us the (approximately) exact value of the target function, while selecting the first will only reduce the overall variance by a moderate amount. On the 4-Sin2D task the variance-straddle heuristic is able to make use of the individual observable functions, but still does not do as well as the variance-maxvarstraddle heuristic.

To illustrate the differences in sampling patterns between these heuristics, we plot the samples chosen for the observable functions (with squares, circles, triangles and x's, respectively) with the true (dashed) and predicted (solid) function level-sets for the 4-Sin2D task in Figure 2. The variance-maxvarstraddle heuristic is much better at picking points than the other two heuristics. Note that the variance-maxvarstraddle heuristic is able to learn that some regions of the space are poor by sampling just one of the observable functions; as such, its samples lie much closer to the target level-set. This reinforces our hypothesis that modeling the observable functions separately results in additional learning opportunities.

4. Joint Statistical Analyses

Now let us look at a concrete application of this sampling algorithm: joint statistical analyses. Let X_i be a random variable denoting a data source and x_i be a generic observation of X_i. For each data set, X_i, let m_i be a corresponding model of X_i given some $\theta \in \Theta$. We are interested in constructing a confidence region for the true value of the parameter, denoted $\theta^\star$, based on the observation that $X_i = x_i$ for each model / data set pair.

For a single data set, consider testing the hypothesis that $\theta^\star = \theta$ at level α for some arbitrary $\theta \in \Theta$. The associated acceptance region for the test, $\mathcal{A}_i(\theta)$, is the set of data values (model outputs) for which the test will not reject the hypothesis $\theta^\star = \theta$ for model m_i. Since we are interested in tests with significance level α, we require $P_\theta(X_i \in \mathcal{A}_i(\theta)) \geq 1 - \alpha$. We can then use $\mathcal{A}_i$ to construct a $1-\alpha$ confidence region, $\mathcal{C}_{\mathcal{A}_i}(x_i)$, for $\theta^\star$ based on the observed data x_i: $C_{\mathcal{A}_i}(x_i) = \{\theta \in \Theta | x_i \in \mathcal{A}_i(\theta)\}$.

We consider two approaches to combine the individual confidence tests above into joint confidence regions. In the first we create a statistical model which simultaneously considers all data sets. For instance, when performing an analysis on two data sets using χ^2 tests, we will have one χ^2 test for data set A and a second for data set B. Since the χ^2 test assumes that each of the data points have dependencies given by the covariance matrix, we can combine the two tests into a single χ^2 test of the form

$$\begin{bmatrix} \vec{x}_A - \vec{m}_A \\ \vec{x}_B - \vec{m}_B \end{bmatrix}^T \begin{bmatrix} \Sigma_A & \Sigma_{AB} \\ \Sigma_{AB} & \Sigma_B \end{bmatrix}^{-1} \begin{bmatrix} \vec{x}_A - \vec{m}_A \\ \vec{x}_B - \vec{m}_B \end{bmatrix} \sim \chi^2_{(a+b)}$$

where $m_\dagger, x_\dagger$ and $\Sigma_\dagger$ are the associated test model, observed data and observed covariance of data set † given some vector from the parameter space, a and b are the degrees of freedom of the tests associated with data sets A and B respectively, and Σ_{AB} is the covariance of the data points between data sets A and B. If data sets A and B are independent, then all elements of Σ_{AB} are zero and we can write the above expression as:

$$(\vec{x}_A - \vec{m}_A)^T \Sigma_A^{-1} (\vec{x}_A - \vec{m}_A) + (\vec{x}_B - \vec{m}_B)^T \Sigma_B^{-1} (\vec{x}_B - \vec{m}_B) \sim \chi^2_{(a+b)}.$$

That is, the target function is merely the sum of the two observable functions: the variance weighted sum of squares for both data sets.

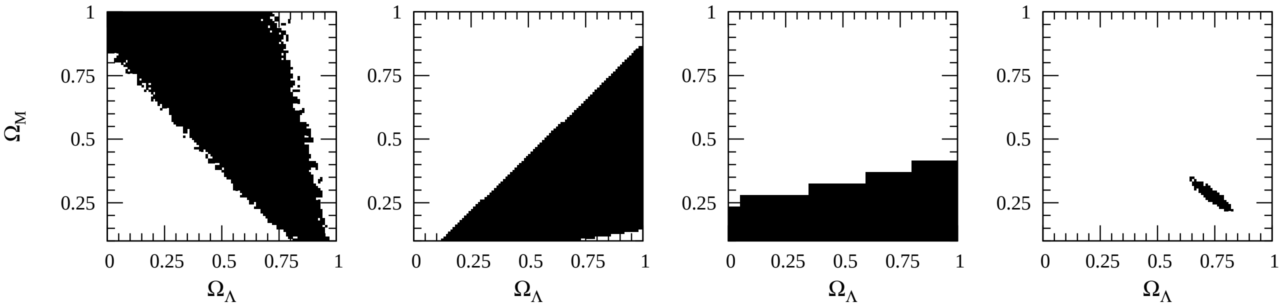
Another approach to performing simultaneous joint analysis is to combine the models' p-values. There are many ways to combine test procedures, including using Bonferroni corrections, the inverse normal method, and inverse logit methods (Hedges, 1985). A common method to combine p-values is Fisher's method (Fisher, 1932). Fisher noted that since a p-value, p_i, has a Uniform distribution, $-2\log(p_i)$ will have a $\chi^2_{(2)}$ distribution. Again, using the fact that the sum of independent χ^2 random variables has a χ^2 distribution, the test becomes: reject H_0 if and only if $-2\sum_{i=1}^{k} \log(p_i) \geq C$ where C is the critical value of a $\chi^2_{(2k)}$ distribution for some particular level α. Again, we see that the target function is the sum of observable functions.

Thus, given the models m_i and data sets X_i, we are interested in locating those $\theta \in \Theta$, such that the the resulting models m_i $(i = 1, \ldots, m)$ are accepted by the chosen hypothesis test. This, in turn, reduces to testing whether the sum of a set of observable functions is below a specified threshold. Specifically, given a threshold t, we want to find the set of points, Θ', where the target function f is equal or less than the threshold: $\Theta' = \{\theta \in \Theta | f(\theta) \leq t\}$. However, note that we need only discover the boundary, $\mathcal{S} = \{\theta \in \Theta | f(\theta) = t\}$, as $\mathcal{S}$ implicitly defines Θ'. Therefore, using either χ^2 tests or Fisher's method, we can apply the algorithm described in Section 2 to locate the boundaries of the $1-\alpha$ confidence region.

5. Cosmological Data Example

To illustrate our algorithm and its application to joint statistical analyses, we show how it can be applied to an analysis of eight cosmological parameters that affect the formation and evolution of our universe using three data sets: the Comic Microwave Background (CMB) power spectrum as observed by Wilkinson Microwave Anisotropy Project (WMAP) (Bennett, C. L., et al., 2003), the Davis, T. M., et al. (2007) supernovae (SN) survey and a large scale structure survey (LSS) from Tegmark, M., et al. (2006).

While models for each of these data sets try to determine what the Universe is formed of and how it has evolved, they measure significantly different aspects of the Universe. The CMB data set records temperature fluctuations in the Uni-



(a) WMAP data (b) Supernova data (c) LSS data (d) WMAP + SN + LSS data

Figure 3. Comparison of the confidence regions derived for WMAP (a), supernova (b), and LSS (c) data sets with those derived using all three data sets together (d). Regions of solid color indicate values for Ω_M and Ω_Λ for which some combination of the remaining parameters results in a model with probability greater than $1-\alpha$. The WMAP and LSS models are 7 parameter models, while the supernova is a 3 parameter model, and the combination model is an 8 parameter model.

verse just after the Big-Bang. The size and spatial proximity of these temperature fluctuations depict the types and rates of particle interactions in the early universe and hence characterize the formation of large scale structure (galaxies, clusters, walls and voids) in the current observable universe. Meanwhile, the supernovae data measures the expansion of the universe as a function of time, in order to constrain the total mass and eventual fate of the Universe. Finally, the large scale structure survey measures the degree of galaxy cluster clumping in order to determine the relative importance of dark matter and Baryonic (normal) matter. Combined, these data sets can be used to determine the age, composition and eventual fate of the Universe, as well as provide strong evidence for the presence of dark energy — a large-scale negative gravitational force.

In this analysis we look at an eight dimensional parameter space comprised of the optical depth (τ), dark energy mass fraction (Ω_Λ), total mass fraction (Ω_m), baryon density (ω_b), dark matter density (ω_dm), neutrino fraction (f_n), spectral index (n_s) and galaxy bias (b). The CMB model constrains the first seven parameters while the supernova model constrains ω_dm, ω_B, Ω_M and Ω_Λ. The LSS model constrains all of the parameters except for τ.

Fisher's method was used to combine p-values from each of the three models. While for small p-values the log of the p-value goes to infinity, note that the algorithm is interested in determining where the sum of the p-values corresponds to the 95% quantile of a $\chi^2_{(6)}$ distribution. Since this results in $t \approx 12.6$, the algorithm has no incentive to select points which are expected to have near zero p-values.

Computing expected observations given parameter vectors is fast for the supernovae and large scale structure models, and hence we can quickly compute the p-values associated with these two models using χ^2 tests. However, computing the expected observations for the CMB data set is much more time consuming. Typically one employs a numerical solver, such as CMBFast to approximate the Boltzmann equation and yield the expected power spectrum.

To alleviate the problem posed by the computational costs of CMBFast, we initialize the Gaussian process model associated with the WMAP data using the one million p-values derived by Bryan, B., et al. (2005). Bryan, B., et al. (2005) uses confidence balls — a statistical procedure similar to χ^2 tests, generally with better inference properties — to map out the level-set associated with the 95% confidence region of the seven CMB parameters. Additional models were selected using the variance-maxvarstraddle heuristic with one small change: If the heuristic selects the observable function associated with the CMB data, we first compute the p-values associated with the supernova and large scale structure data sets to see if we can exclude the parameter vector without needing to run CMBFast. That is, we determine whether the sum of the log p-values from the supernovae and large scale structure data sets alone is larger than the threshold for the combined model. This modification allows us to reduce the number of CMBFast computations by about a factor of five. Using this modified variance-maxvarstraddle heuristic, we sampled roughly 1.5 million additional parameter vectors, about 300,000 of these points resulted in CMBFast runs. Note that 1.5 million parameter vectors corresponds to a grid with roughly six elements per side. Since the variance-based metrics sample the entire parameter space, their prediction performance is typically similar to this naive gird. Thus, using an active learning metric that focuses on the boundary that we are interested in (and ignores large parts of the parameter space which can be proved to be infeasible) significantly reduces the computational complexity of the algorithm.

In Figures 3(a)-3(c) we depict 95% confidence regions derived using only a single data set projected into the Ω_M versus Ω_Λ space. Confidence regions are derived by binning the samples selected by the algorithm and including those bins in the confidence region which contain points where

$f \leq t$, resulting in the blockiness in the diagrams. The figures illistrate that the shapes of the 95% confidence regions for each of the data sources are quite different, validating our supposition that different observable functions can be used to efficiently reject parts of parameter space.

In Figure 3(d), depicts the 95% confidence region found using the joint analysis for all three data sets; one and two dimensional projections onto the other parameters can be found in Bryan (2007). It is clear that using the combination of all three data sets dramatically improves the inferences that can be made on the cosmological parameters' values. In particular, note that the derived confidence region is significantly smaller than what would have been obtained using a simple intersection. As a result, we cannot blindly combine the WMAP p-values of Bryan, B., et al. (2005) with p-values derived for the supernova and large scale structure data sets, as the surface of the combined target function is drastically different from the surfaces of each of the models independently. Specifically, all of the models in the Bryan, B., et al. (2005) data set can be rejected at the 95% confidence level by the supernova and large scale structure data. This is not surprising; the analysis of Bryan, B., et al. (2005) used only CMBFast one the WMAP data, and it is well known that CMBFast only loosly fits the WMAP data (Spergel, D. et al., 2003). Thus in order to accurately compute the 95% confidence regions of the joint model (using all three data sets), we must sample new models in the multiple model framework, as we did in Figure 3(d). Only then will we correctly learn the true level-set of the composite target function.

6. Conclusions

We have described the problem of learning a target function based on a set of related observable functions. This problem naturally arises in many situations including the joint analysis of multiple data sets which describe a single physical phenomenon. We have developed an algorithm for locating the level set of this target function while minimizing the number of experiments necessary. We described and showed how several different heuristics for choosing experiments from a set of candidates perform on synthetic target functions. Our experiments indicate that variance-maxvarstraddle outperforms both random and variance-weighted heuristics typically applied to active learning problems. Moreover, variance-maxvarstraddle is better than both the sequential- and variance-straddle heuristics, as it appears to better approximate the information gain of a candidate point.

Using the variance-maxvarstraddle heuristic, we were able to efficiently learn the level set of an eight dimensional surface. This level-set corresponds to the 95% confidence region of a joint analysis between three data sources. Using the CMB, supernovae and large scale structure data sets results in much tighter confidence regions than those obtained using only a single source of data, allowing for stronger scientific inferences. Standard ad hoc techniques for combining evidence, such as intersecting the data, or using weak priors do not result in such a significant reduction in the accepted parameter space.

References

Angluin, D. (1988). Queries and concept learning. *Machine Learning*, *2*, 319–342.

Bennett, C. L., et al. (2003). First-Year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Foreground Emission. *Astrophysical Journal Supplemental*, *148*, 97–117.

Bryan, B. (2007). *Actively learning specific function properties with application to statistical inference*. Doctoral dissertation, Carnegie Mellon University.

Bryan, B., et al. (2005). Active learning for identifying function threshold boundaries. In *Advances in neural information processing systems 18*. Cambridge, MA: MIT Press.

Cressie, N. (1991). *Statistics for spatial data*. New York: Wiley.

Davis, T. M., et al. (2007). Scrutinizing Exotic Cosmological Models Using ESSENCE Supernova Data Combined with Other Cosmological Probes. *Astrophysical Journal*, *666*, 716.

Fisher, R. (1932). *Statistical methods for research workers*. London: Oliver and Boyd. 4 edition.

Guestrin, C., et al. (2005). Near-optimal sensor placements in gaussian processes. *ICML 2005: Proceedings of the 22nd International Conference on Machine learning*. ACM Press.

Hedges, L. V. (1985). *Statistical methods for meta-analysis*. Academic Press.

Kersting, K. et al. (2007). Most likely heteroscedastic gaussian process regression. *ICML '07: Proceedings of the 24th International Conference on Machine Learning*. ACM Press.

MacKay, D. (1992). Information-based objective functions for active data selection. *Neural Computation*, *4*, 590.

Ramakrishnan, N., et al. (2005). Gaussian processes for active data mining of spatial aggregates. *Proceedings of the SIAM International Conference on Data Mining*.

Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian processes for machine learning*. MIT Press.

Santner, T. J., Williams, B. J., & Notz, W. (2003). *The design and analyis of computer experiments*. Springer. 1 edition.

Shmueli, G., & Fienberg, S. E. (2006). *Statistical methods in counterterrorism*, chapter Current and Potential Statistical Methods for Monitoring Multiple Data Streams for Biosurveillance, 109. New York: Springer.

Spergel, D. et al. (2003). First-Year Wilkinson Microwave Anisotropy Probe Observations: Determination of Cosmological Parameters. *Astrophysical Journal Supplemental*, *148*.

Tegmark, M., et al. (2006). Cosmological constraints from the SDSS luminous red galaxies. *Physical Review D*, *74*, 123507.