

# Optimizing the Limit Setting Potential of a Multivariate Analysis Using the Bayes Posterior Ratio

G.C. Hill<sup>1</sup>, F. Lu<sup>2</sup>, P. Desiati<sup>1</sup>, G. Wahba<sup>2</sup>

1. Department of Physics, 2. Department of Statistics,  
University of Wisconsin, Madison, 53706, USA

In this work we consider the problem of optimal cut selection in a multivariate analysis. When we wish to place an upper limit on the normalisation of a theoretical flux model, we show how the best detector sensitivity is found by optimizing the ratio of the average upper limit to the expected signal. In a multidimensional observable space, we find the constant Bayes posterior surface that defines an acceptance region of events yielding the best limit setting power. The calculation of the posterior using a penalized likelihood method is described.

## 1. INTRODUCTION

In this paper, we consider the problem of optimizing the limit setting power of a multivariate analysis. A limit optimisation technique, the *model rejection potential* [1] technique, has been proposed and used in various analyses (see e.g. ref. [2]). The best limiting power is found by minimizing the *model rejection factor*, the ratio of the average upper limit to the expected signal. In these analyses, event selection proceeds by first cutting on several variables, then performing the optimisation in the final variable with best discriminating power between signal and background. In this paper, we propose the simultaneous optimisation across all variables, by finding the multi-dimensional constant Bayes posterior hypersurface that yields the lowest model rejection factor. A method is proposed here (modified penalized likelihood estimation) for identifying the level curves of the posterior. This is an extension of previous penalized likelihood methods to the use of simulated training data that is drawn from biased distributions via importance sampling, a common practice in particle physics and astrophysics.

## 2. OPTIMIZING EXPERIMENTAL LIMIT SETTING POWER

We will use the example of setting a limit on a diffuse flux of extraterrestrial neutrinos in an underground detector to illustrate how limits are set and optimized in the field of particle astrophysics. Suppose one has a neutrino detector which observes atmospheric neutrinos, produced via interactions of cosmic rays in the earth's atmosphere. These neutrinos are produced with a power law spectrum that goes approximately as  $E^{-3.7}$ . Models of an extraterrestrial flux of neutrinos from the sum of all active galaxies have a somewhat flatter power law, with an energy dependence of  $E^{-2}$ . We will write  $\phi_s \Phi_s(E)$  for the extraterrestrial signal neutrino flux and  $\phi_b \Phi_b(E)$  for the background atmospheric neutrino flux, where  $\Phi_s(E)$

and  $\Phi_b(E)$  are p.d.f.s (i.e.  $\int_E \Phi_s(E) dE = 1$ ). We denote the detector response to the flux of neutrinos by the probability  $P(x | E)$ , where  $x$  is a possibly multidimensional vector of observables describing an event. Then the p.d.f.s for signal and background in the event space are  $h_s(x) = \int_E P(x | E) \Phi_s(E) dE$  and  $h_b(x) = \int_E P(x | E) \Phi_b(E) dE$ . Over the space of all possible events, we therefore expect  $\phi_s$  signal and  $\phi_b$  background events. Then the number of signal events,  $N_s(r)$  expected in some yet to be defined subregion of  $x$ , denoted  $\Psi_r$ , can be written as

$$N_s(r) = (\phi_s + \phi_b) \pi_s \int_{x \in \Psi_r} h_s(x) dx \quad (1)$$

where the prior probability for signal is defined as  $\pi_s = \phi_s / (\phi_s + \phi_b)$ . After reducing the data by cutting on some of the variables, thereby leaving a subregion  $\Psi_r$  of events, we wish to set a limit on the normalisation scale factor  $\phi_s$ . This involves determining an experimental signal event upper limit  $\mu(N_{obs}(r), N_b(r))$ , which is a function of the number of observed events,  $N_{obs}(r)$ , and expected background,  $N_b(r)$ , after the cuts are applied. The limit on the normalisation of the source flux will then be  $\phi_{lim}(r) = \phi_s \times \mu(N_{obs}(r), N_b(r)) / N_s(r)$ . The choice of final cut is optimized before examining the data by minimizing the average "model rejection factor", where  $MRF(r) = \bar{\mu}(N_b(r)) / N_s(r)$  [1], where the as yet unknown experimental event limit  $\mu(N_{obs}(r), N_b(r))$  is replaced by the *average* upper limit  $\bar{\mu}(N_b(r))$  [3]. Over an ensemble of hypothetical repetitions of the experiment, this choice of cut will lead to the best average limit  $\bar{\phi}_{lim}$ . Importantly, the choice of the optimal region  $\Psi_r$  is independent of the original assumption of the normalisation of the source flux to be tested ( $\phi_s$  cancels in the expression for  $\bar{\phi}_{lim}$  leaving  $\bar{\phi}_{lim}(r) = \bar{\mu}(N_b(r)) / \int_{x \in \Psi_r} h_s(x) dx$ ). This method has been applied to the analysis of data from the AMANDA-B10 detector [2], where preliminary cuts were made to isolate atmospheric neutrinos, then the model rejection potential method was applied to the most energy sensitive variable, the number of detector optical modules that had registered Cherenkov pho-

tons in a given event. The final region  $\Psi_r$ , is essentially a “rectangular” region in the space of the observables, with only the final cut optimized to give the best limit setting potential.

Rather than finally optimize with respect to a single variable, it is desired to find a region in the multidimensional variable space for which the inclusion of events leads to the optimized model rejection potential and best limit. We use the Neyman-Pearson Lemma as a guide to defining the optimal region. It states that the critical region defining the most powerful test of one hypothesis against an alternative is given by taking all events with a p.d.f. ratio greater than some constant. Following this reasoning, we only allow regions of the form  $\Psi_r$  containing all  $x$  such that  $h_s(x)/h_b(x) \geq r$ . For a particular cut determined by  $r$ , we can then calculate the limit on the flux normalisation  $\phi_{lim}(r) = \phi_s \bar{\mu}(N_b(r))/N_s(r)$  as a function of  $r$  and seek to find  $r$  which minimizes it. In the next section, we discuss how a model of  $\Psi_r$ , equivalently the level curves of  $h_s(x)/h_b(x)$ , may be obtained using a penalized likelihood estimation method.

### 3. MODIFIED PENALIZED LIKELIHOOD ESTIMATION

Let  $x$  be a possibly multidimensional vector of event observables derived from a reconstructed event. Let  $h_s(x)$  be the probability density function for signal vectors and  $h_b(x)$  be the probability density for background vectors, and let  $\pi_s$  and  $\pi_b$  be prior probabilities of a signal and background observation, respectively. Then the posterior probability that  $x$  is a signal vector is  $p(x) = \pi_s h_s(x) / (\pi_b h_b(x) + \pi_s h_s(x))$ . The logit  $f(x)$  is defined as  $\log[p(x)/(1-p(x))] \equiv \theta + \log[h_s(x)/h_b(x)]$ , where  $\theta = \log \pi_s / \pi_b$ . We will estimate the unbounded  $f(x)$  (rather than  $p(x)/(1-p(x))$  or the bounded  $p(x)$ ) for a particular (implicit) value of  $\theta$ , but since the end result is to obtain level curves of  $f$ , the particular value of  $\theta$  is not important for the calculations. A modified form of the penalized likelihood estimate [4–6] will be used.

Let  $y_i$  be a random variable that is 1 (signal) with probability  $p(x_i)$  and 0 (background) with probability  $1 - p(x_i)$ . Then the likelihood of a single observation  $y_i$  is:  $\mathcal{L} = p(x_i)^{y_i} (1 - p(x_i))^{1-y_i}$ . The negative log likelihood of (independent) data  $y_1, \dots, y_n$  is then, in terms of the logit given by

$$Q(y, f) = \sum_{i=1}^n [\log(1 + e^{f(x_i)}) - y_i f(x_i)] \quad (2)$$

We want to find  $f \cong \sum c_k B_k \in H_K$  (a reproducing kernel Hilbert space (RKHS)[4, 5, 7]) which minimizes the penalized log-likelihood:  $I_\lambda(c) = Q(y, f) + \lambda \|f\|_{H_K}^2$ , where  $B_k$ 's are basis functions in  $H_K$  and  $\|\cdot\|_{H_K}$  is the function norm in  $H_K$ .

This is essentially the penalized log likelihood estimate of  $f$  proposed in O’Sullivan, Yandell and Raynor [8], and in common usage in some fields. Under rather general conditions, which include a proper choice of  $\lambda$ , penalized log likelihood estimates in many RKHS are known to converge to the “true”  $f$  as the sample size becomes large [9]. RKHS’s are discussed in Aronszajn [7] and their use in statistical model building in Wahba [4] and elsewhere, and a wide variety of these spaces are available. An RKHS is characterized by a unique positive definite function  $K(\cdot, \cdot)$ , and once  $K$  is chosen, the exact minimizer of  $I_\lambda(c)$  is known to be in the span of a certain set of basis functions determined from  $K$  [10]. In Section 4 below we will select a particular  $K$ , known to be a good general purpose choice, and use an approximating subset of this set of basis functions. Estimating  $f$  rather than  $p$  directly gives a strictly convex optimization problem whose gradient and Hessian are simple to compute, which then makes the numerical analysis easier and suitable for very large data sets. It is possible to estimate  $p$  directly [11] but this estimate is harder to compute in large data sets and is believed to be not as accurate.

The form of the negative log likelihood in equation 2 applies where the simulated training data is distributed as  $h_b(x)$  and  $h_s(x)$  through sampling directly from the generating distributions  $\Phi_s(\tilde{E})$  and  $\Phi_b(\tilde{E})$ , then processing the events  $\tilde{E}$  through the simulation chain to give events  $x$ . Here,  $\tilde{E}$  means a vector of generating parameters, e.g. neutrino energy, position and arrival direction. Often, we wish to emphasize more interesting regions of the event space, by e.g. sampling from biased energy and arrival directions, or by forcing events to occur close to the detector. Suppose these biased distributions in the generating parameters may be summarized as  $g_b(\tilde{E})$  and  $g_s(\tilde{E})$ , or there may be a single biased sampling distribution,  $g(\tilde{E})$ . We “unbias” the events by applying weight factors throughout any subsequent procedure. The weight for a given signal event  $x_i$  will be  $w_s(x_i) = \Phi_s(\tilde{E}_i)/g_s(\tilde{E}_i)$  and for a background event  $w_b(x_i) = \Phi_b(\tilde{E}_i)/g_b(\tilde{E}_i)$ . In the case that a single sampling spectrum is used each event is re-weighted to both signal and background energy spectra. In either case the weights satisfy  $\sum_{i=1}^n w_s(x_i) = N_s$  and  $\sum_{i=1}^n w_b(x_i) = N_b$ , i.e. the predicted numbers of events from the weighted simulation is the same as that from an un-weighted simulation. Now, if we have multiple unbiased observations at some  $x_i$  as  $y_{ij}, j = 1, \dots, m(i)$ , the likelihood of all these observations is:  $\mathcal{L} = p(x_i)^{\sum_{j=1}^{m(i)} y_{ij}} (1 - p(x_i))^{\sum_{j=1}^{m(i)} (1-y_{ij})}$ . If the samplings at  $x_i$  are biased, then the exponent sums are weighted by  $w_s(x_i)$  and  $w_b(x_i)$  respectively

leading to a modified likelihood

$$Q(w, f) = \sum_{i=1}^n \sum_{y_i=0}^1 w_{y_i} [\log(1 + e^{f(x_i)}) - y_i f(x_i)] \quad (3)$$

where  $w_{y_i} = w_s(x_i)$  for  $y_i = 1$  and  $w_{y_i} = w_b(x_i)$  for  $y_i = 0$ . The incorporation of weighted events is thus simply accounted for by weighting the terms in the logarithmic likelihood sum. Further, we can substitute  $w_s(x_i)$  and  $w_b(x_i)$  to obtain an alternative form of the likelihood

$$Q(w, f) = \sum_{i=1}^n \{w_t(x_i) [\log(1 + e^{f(x_i)}) - \tilde{p}(x_i) f(x_i)]\} \quad (4)$$

where  $w_t(x_i) = w_s(x_i) + w_b(x_i)$  and  $\tilde{p}(x_i) = w_s(x_i)/w_t(x_i)$ .

## 4. IMPLEMENTATION OF THE MODIFIED PLE METHOD

After getting the modified penalized likelihood formulation, we now can move on to look for a ‘good’ estimate of  $f(x)$  whose level curves can be obtained. In our implementation, we use radial basis functions plus constant and linear terms. So,

$$f(x) = \beta_0 + \beta^T x + \sum_{k=1}^N c_k K_\sigma(x, x_{i_k}), \quad (5)$$

where  $K_\sigma(\cdot, \cdot)$  is the Gaussian kernel with isotropic variance  $\sigma^2$ ,  $N$  is the total number of basis functions and the  $N$   $x_{i_k}$ ,  $k = 1, \dots, N$  will be chosen as a subset of the  $x_i$ ,  $i = 1, \dots, n$  as described below. Thus,  $f$  will be specified as long as all coefficients, i.e.  $\beta_0$ ,  $\beta$  and the  $c_i$ ’s are determined (note that  $\beta$  is a vector). By letting  $\lambda \|f\|_{H_K}^2 = \lambda \sum_{i,j=1}^N c_i c_j K_\sigma(x_i, x_j)$ , we put a penalty only on the  $c_i$ ’s.

We used a sequence of simulated data driven procedures to fit the model in the sense that we let the simulated data choose the ‘best’ combination of smoothing parameter  $\lambda$ , scale parameter  $\sigma$  and number of basis functions  $N$ . Five dimensional simulated data ( $x_i$ ’s) are first rescaled using their own sample weighted standard deviation after a log transformation. Then, the whole simulated data set is randomly divided into three subsets of almost the same size, one as training set, one as tuning set and the last one as testing set. After that, we randomly, but according to weights (large-weight simulated data points have higher chance to be selected), choose the  $N$   $x_{i_k}$ ’s which determine basis functions as a subset of the training set. We solve the minimization problem on a coarse 2-D parameter grid of  $\lambda$  (usually on a log scale) and  $\sigma^2$  using the training set. For each parameter

pair (each point on the grid), a Newton-Raphson iteration is used to solve this convex minimization problem [4]. After the algorithm converges we calculate the Kullback-Leibler (KL) distance between tuning simulated data and fitted model, which is essentially just the first term of  $I_\lambda(c)$  for tuning simulated data with  $f$  replaced by  $\hat{f}$ . We then find the best parameter combination based on the KL distance over the coarse grid. Starting from there, a direct-searching simplex method [12] is used to search for a locally best parameter combination according to the KL distance criterion. The procedure is repeated using  $2N$  bases, then  $4N$  bases and so on, until the improvement on the KL distance is smaller than some preset threshold. We use the coefficients corresponding to the then-best combination of parameters to construct our final estimate of the logit function. Next, the testing set is used to check the goodness of fit of this final model and to determine the optimal cut on  $p(x_i)$  and thus the limit setting power of the analysis. Finally, the real data can be analysed by applying the optimal  $p(x_i)$  cut, and the limit on the signal model determined.

## 5. CONCLUSIONS

In this work, we have described how the limit setting potential of a multivariate analysis is optimized by choosing an acceptance region of all  $x$  for which the estimated posterior probability  $p(x)$  is greater than some specified threshold. The main contribution of this paper is to introduce the well known penalized log likelihood estimation procedure for estimating  $f$  and hence  $p$  to an audience to which it is apparently unfamiliar, and to develop numerical algorithms for efficient computation and testing of the estimate that are appropriate for large multivariate data sets obtained via importance sampling.

## Acknowledgments

This research was supported by the National Science Foundation under Grants DMS-0072292 (G. Wahba and F. Lu) and OPP-9980474 (G.C. Hill and P. Desiati).

## References

- [1] G. C. Hill and K. Rawlins, “Unbiased cut selection for optimal upper limits in neutrino detectors: the model rejection potential technique”, *Astropart. Phys.* **19**, 393, 2003.
- [2] J. Ahrens *et al.* “Limits on diffuse fluxes of high energy extraterrestrial neutrinos with the

- AMANDA-B10 detector”, *Phys. Rev. Lett.* **90**, 251101, 2003.
- [3] G. J. Feldman and R. D. Cousins, “Unified approach to the classical statistical analysis of small signals”, *Phys. Rev. D* **57**, 3873, 1998.
- [4] G. Wahba, “Spline Models for Observational Data”, CBMS-NSF Regional Conference series in applied mathematics, **59**, 1990.
- [5] G. Wahba, “Soft and Hard Classification by Reproducing Kernel Hilbert Space Methods”, *Proceedings of the National Academy of Sciences*, **99**, 16524-16530, 2002.
- [6] G. Wahba, Y. Wang, C. Gu, R. Klein, and B. Klein, “Smoothing Spline ANOVA for Exponential Families, with Application to the Wisconsin Epidemiological Study of Diabetic Retinopathy”, *Ann. Statist.*, **23**, 1865-1895, 1995.
- [7] N. Aronszajn, “Theory of reproducing kernels” *Trans. Am. Math. Soc.*, **68**, 337-404, 1950.
- [8] F. O’Sullivan, B. Yandell and W. Raynor, “Automatic smoothing of regression functions in generalized linear models”, *J. Amer. Statist. Assoc.*, **81**, 96-103, 1986.
- [9] D. Cox, and F. O’Sullivan, “Asymptotic analysis of penalized likelihood and related estimators” *Ann. Statist.* **18**, 1676-1695, 1990.
- [10] G. Kimeldorf and G. Wahba, “Some results on Tchebycheffian spline functions”, *J. Math. Anal. Applic.*, **33**, 82-95, 1971.
- [11] M. Villalobos and G. Wahba, “Inequality constrained multivariate smoothing splines with application to the estimation of posterior probabilities”, *J. Am. Statist. Assoc.*, **82**, 239-248, 1987.
- [12] J.C. Lagarias, J. A. Reeds, M. H. Wright, and P. E. Wright, “Properties of the Nelder-Mead Simplex Method in Low Dimensions”, *SIAM Journal of Optimization*, **9**, Number 1, 112-147, 1998.