From data to constraints

S. Mukhopadhyay, E. Parzen, and S. N. Lahiri

Citation: AIP Conference Proceedings **1443**, 32 (2012); doi: 10.1063/1.3703617
View online: http://dx.doi.org/10.1063/1.3703617
View Table of Contents: http://scitation.aip.org/content/aip/proceeding/aipcp/1443?ver=pdfcov
Published by the AIP Publishing

Articles you may be interested in

Instabilities in dark coupled models and constraints from cosmological data
AIP Conf. Proc. **1241**, 1016 (2010); 10.1063/1.3462595

Reconciling Solar Interior Models and Helioseismological Data: Constraints on the Neon Content of the Sun from Nearby B Stars
AIP Conf. Proc. **948**, 225 (2007); 10.1063/1.2818975

Constraints on ocean internal wave spectra from longrange acoustic transmission data
J. Acoust. Soc. Am. **103**, 2789 (1998); 10.1121/1.422293

Gamma-Ray burst redshift constraints from BATSE spectral data
AIP Conf. Proc. **384**, 482 (1996); 10.1063/1.51708

The black hole mass in Xray Nova Muscae: Constraint from SIGMA annihilation line data
AIP Conf. Proc. **280**, 423 (1993); 10.1063/1.44312

This article is copyrighted as indicated in the article. Reuse of AIP content is subject to the terms at: http://scitation.aip.org/termsconditions. Downloaded to IP: 155.247.166.234 On: Sat, 05 Jul 2014 15:34:39

From Data To Constraints

S.Mukhopadhyay, E.Parzen and S.N.Lahiri

Texas A&M University, Department of Statistics

Abstract. Jaynes' Maximum Entropy (MaxEnt) inference starts with the assumption that we have a set of known constraints over the distribution. In statistical physics, we have a good intuition about the conserved macroscopic variables. It should not be surprising that in a real world applications, we have no idea about which coordinates to use for specifying the state of the system. In other words, we only observe *empirical data* and we have to take a decision on the constraints from the data. In an effort to circumvent this limitation, we propose a nonparametric quantile based method to extract relevant and significant facts (*sufficient statistics*) for the maximum entropy *exponential model*.

Keywords: Maximum entropy, mid-rank transformations, exponential model, quantile function, nonparametric Entropy estimation.
PACS: 02.50.cW , 02.50.tT.

INTRODUCTION

One of the profound questions of stochastic modeling is "*What comes first - a Parametric Model or Sufficient Statistics* ?". We propose the following philosophy of modeling which goes from measurement to parameter through sufficient statistics. MaxEnt is such a modeling framework where we can start modeling by first specifying the appropriate sufficient statistics in terms of the constraints and then performing nonparametric Score test to finally arrive at the parametric model. But unfortunately Maximum entropy principle is silent on the role of finding the proper constraints from data. The aim of this paper is to introduce a unified framework for answering (i) How to derive constraints from measurements? and (ii) How many of them to use ! Note that the distribution derived using the MaxEnt principle assumes that we have a proper set of constraints that can explain the phenomena under study. But in practice, we rarely have this situation and we have to take a decision about the proper choice of constraints. In this paper we introduce a novel way of building nonparametric robust score functions (sufficient statistics) and describe a goodness of fit framework using entropy statistics for selecting proper number of constraints. We believe that, this novel systematic approach for inference , will help the maximum entropy to go beyond the conventional "*exploratory phase*" and become an *objective inferential paradigm* for practitioners. We demonstrate the steps for efficient representation, processing and data analysis using microarray gene expression data.

MAXENT AND NONPARAMETRIC STATISTICS

Maximum entropy is a probability modeling principle which converts a nonparametric problem into a parametric one by putting the entropy criteria under *suitable* constraints.

Bayesian Inference and Maximum Entropy Methods in Science and Engineering
AIP Conf. Proc. 1443, 32-39 (2012); doi: 10.1063/1.3703617
© 2012 American Institute of Physics 978-0-7354-1039-8/$30.00

This article is copyrighted as indicated in the article. Reuse of AIP content is subject to the terms at: http://scitation.aip.org/termsconditions. Downloaded to IP: 155.247.166.234 On: Sat, 05 Jul 2014 15:34:39

Exponential Model and Maximum Entropy

Let X be a continuous random variable with density f, and our aim is to find the probability law or the distribution of X on the basis of a random sample $X_1, X_2, \ldots X_n$ in a fully nonparametric way. One elegant solution for this problem is the Maximum Entropy which has it's root in Thermodynamics. Rather than directly dealing with the raw data, it starts with few *known* macroscopic summary statistics (sufficient statistics) of the empirical data as form of constraints and then maximize the entropy (defined as, $H(f) = \int_{-\infty}^{\infty} -\log f(x)\, f(x)\, \mathrm{d}x$) of X to find the distribution uniquely. In general, let $S_1(X), S_2(X), \ldots, S_m(X)$ be the sufficient statistics and we want to find the distribution for which we have,

$$\max_f H(f) \quad \text{subject to} \quad \mathbb{E}_f[S_k(X)] = n^{-1} \sum_{i=1}^{n} S_k(X_i), \ \text{ for } k = 1, 2, \ldots, m. \tag{1}$$

This characterizes the density $\widehat{f}(x) = Z^{-1} \exp[\sum_{k=1}^{m} \hat{\theta}_j S_k(x)]$ which belongs to the exponential family.

Note 1. The MaxEnt distribution heavily depends on the *form* of the score functions $S_k(X)$ and the *number* of such score function, i.e, m.

Note 2. There exists standard procedure to find sufficient statistics for *specified* exponential family models. For example, if we assume the underlying law is Gaussian then it is enough to summarize the data using $S_1(X) = X$ and $S_2(X) = X^2$ and perform parametric score test to decide how many of them we need. The point worth emphasizing at this point is that, in MaxEnt *nonparametric density estimation*, we face the challenging *inverse problem* of designing basis functions (possibly non-linear) from the data without characterizing the underlying distribution.

The rest of this article is devoted to building a unified theoretical framework and algorithm to answer *how to pick the moment constraints and how many of them we should pick ?*, thus widening the scope of MaxEnt.

Note 3. One of the major successes of our proposed methodology is its simplicity and generality. We can incorporate any definition of entropy, for example $\int f \log f$ or $\int \log f$ into our analysis (yields respectively two popular nonparametric likelihoods , MaxEnt and Empirical likelihood), broadening the applicability and importance of our methodology.

Unified Methodology using Quantile Technology

In this section we introduce the necessary concepts on quantile based machinery to design the score function, initiated in Parzen (1979, 1983, 1991, 2004, 2009) and latter we show how to build a nonparametric goodness of fit measure for selecting the appropriate number of these score functions.

This article is copyrighted as indicated in the article. Reuse of AIP content is subject to the terms at: http://scitation.aip.org/termsconditions. Downloaded to IP: 155.247.166.234 On: Sat, 05 Jul 2014 15:34:39

Mid-Quantile Transformation

One of the key intermediate step for building the sufficient statistics is mid-rank transformation, which plays an unifying role in non-parametric data analysis. For a discrete random variable with probability mass function (pmf) $p(x) = \Pr(X = x)$ and cumulative distribution function (cdf) $F(x)$, the Mid-distribution function, defined as

$$F^{\text{mid}}(x) = F(x) - .5\, p(x),\ x \in \mathbb{R}. \tag{2}$$

Our data analysis starts with transforming the raw data $x_1, x_2, \ldots x_n$ to $u_1, u_2, \ldots, u_n$, where $F^{\text{mid}}(x_i) = u_i$. The elegant formula for the mean and variance of $W = F^{\text{mid}}(X)$ is given by $\mathbb{E}(W) = .5$ and $\text{Var}(W) = \sigma_{\text{mid}} = 1/12\,[1 - \mathbb{E}(p^2(X))]$ (Parzen, 2004). Our whole framework depends on the quantile function which traditionally defined as $Q(u) = \inf_t\{F(t) \geq u\}$, where $F(x)$ is the distribution function. But this definition face roadblock for discrete/grouped data. The primary reason for introducing the mid-distribution function is to unify the analysis of continuous and discrete data (with or with out ties). To understand further the role played by mid-distribution function, let us assume we are having data from some underlying absolutely continuous distribution F. Our job then boils down to estimating the underlying unknown *continuous* density from observed *discrete* finite sample by turning the MaxEnt crank. Mid-distribution transformation (contrary to $F(x)$) improves the accuracy of approximation of discrete random variable by continuous random variable as the inversion formula of a distribution function from characteristic functions actually holds for mid-distributions (at all x, not just x a continuity point of F).

Building Score Functions

Novelty of our approach is in the construction of the basis functions. In contrast to the standard practice of taking the basis as powers of x, here we use orthonormal score functions based on ranks through mid-distribution transform. We define orthonormal score functions as

$$S_j(u) = \mathscr{L}_j\left[\widetilde{F}^{\text{mid}}(\,\widetilde{Q}(u)\,)\right] \text{ for } j = 1, 2, \ldots, m. \tag{3}$$

Here $\mathscr{L}_1, \mathscr{L}_2, \ldots, \mathscr{L}_m$ are Legendre Polynomials on $[0,1]$ and the $\widetilde{}$ signifies the empirical estimate of the population version. In stead of Legendre Polynomials one can choose any orthogonal polynomials like cosine or Hermite polynomials. The choice solely case specific; for example Legendre polynomials turns out to be very useful for building score functions (sufficient statistics) in the case of pattern classification problems as we will demonstrate in the next section. This approach of constructing the score functions has the following added advantages : (i) robust; (ii) works for both continuous, discrete data ; (iii) bounded score functions avoids the problems with non-integrable densities in the exponential model and lastly, (iv) the use of orthogonal polynomials for the basis significantly improves the accuracy and stability the MaxEnt algorithm.

This article is copyrighted as indicated in the article. Reuse of AIP content is subject to the terms at: http://scitation.aip.org/termsconditions. Downloaded to IP: 155.247.166.234 On: Sat, 05 Jul 2014 15:34:39

Application to Probabilistic Classification

Here we will show how to apply the theoretical concepts of previous section to classification and variable selection. Consider the problem of classifying the patients into two group sick ($Y = 1$) and healthy ($Y = 0$) on the basis of p gene expression values $X_1, X_2, \ldots X_p$. Let us denote $F(x|Y = 1) = F(x)$ and $F(x|Y = 0) = G(x), x \in \mathbb{R}$. Let $H(x)$ denote the polled cdf, given be $H(x) = \pi F(x) + (1 - \pi) G(x)$, where $\pi = \Pr(Y = 1)$, proportion of sick people in the population. Typically we have $p \approx 5000$ in this type of classification setup, involving microarray gene expressions. So the first challenging task is to reduce the number of variables and our specially designed score functions help to achieve the goal in the following way. For each variable X_k, we have the following result,

Theorem 1 (Representation of Wilcoxon Statistics)**.** *Wilcoxon rank sum statistics can be represented by* Wil *(linearly equivallent version of it),*

$$\text{Wil}(X_k) = \mathbb{E}\left[S_1(\widetilde{H}^{mid}(X_k)) \,|\, Y = 1 \right].$$

For proof see Mukhopadhyay et al. 2011. Theorem 1 indicate the interpretation of Wilcoxon statistics as $< Y, S_1(u_k) >$, where S_1 is specially designed according to Eq [3] and $u_k = \widetilde{H}^{\text{mid}}(x_k)$. This motivates us to propose a nonlinear variable selection measure

$$C_k = \sum_{j=1}^{m} < Y, S_j(u_k) >^2 \quad \text{for } k = 1, 2, \ldots, p. \tag{4}$$

Using this importance score we can select the discriminative variables and thus reduce the complexity of the model. This demonstrate the dual role played by the score functions $S_1, S_2, \ldots, S_m$ for (i) creating appropriate summary statistics (ii) utilizing those for nonparametric robust variable selection. In the following section we will discuss the issue "How to choose m, the number of basis function" for a variable, developing a completely nonparametric goodness of fit measure.

Nonparametric Estimation of Kullback-Liebler Information

A well known measure of similarity between statistical models is Kullback-Liebler (KL) information number that can be expressed in terms of comparison density $d(u) = f(H^{-1}(u))/h(H^{-1}(u)), 0 < u < 1$ as

$$\begin{aligned} I(H:F) &= \int_{-\infty}^{\infty} \log\left[h(x)/f(x)\right] h(x)\, \mathrm{d}x. \\ &= -\int_0^1 \log\left[f(H^{-1}(u))/h(H^{-1}(u)) \right] \mathrm{d}u = -\int_0^1 \log d(u)\, \mathrm{d}u,\ 0 < u < 1 \end{aligned} \tag{5}$$

where last line follows from the substitution $x = H^{-1}(u)$. The reason to call it a density due to the fact that $\int_0^1 d(u)\, \mathrm{d}u = 1$. Thus to estimate the KL number we have to estimate

This article is copyrighted as indicated in the article. Reuse of AIP content is subject to the terms at: http://scitation.aip.org/termsconditions. Downloaded to IP: 155.247.166.234 On: Sat, 05 Jul 2014 15:34:39

the comparison density of the corresponding variable. Here we will use the previously designed score functions to build the exponential density given by,

$$\widehat{d}(u;\theta_1,\ldots\theta_m) = Z_m^{-1}\exp[\sum_{k=1}^{m}\widehat{\theta}_k S_j(u)], \quad 0<u<1. \tag{6}$$

Combining Eq. (4) and (5) immediately gives the following important Corollary.

Corollary 1. *Relation of Entropy number and log Partition function of the fitted exponential model*

$$\widehat{I}_m(H:F) = -\int_0^1 \log\widehat{d}(u;\theta_1,\ldots\theta_m)\,\mathrm{d}u = \log Z_m. \tag{7}$$

It gives a clear strategy for model selection comparing the nonparametric estimates of information numbers $\widehat{I}_1,\ldots,\widehat{I}_m$. For MaxEnt models (Eq. 6) we just need to compare the log partition numbers of various models to find the best yet simple model that fits the data.

Note 4. *Neyman (*1937*)* $\rightarrow$ *Akaike (*1973*)* $\rightarrow$ *Parzen (*1983*)*

Akaike Information Criteria (AIC) (Akaike 1973) is an *asymptotically* unbiased estimator of the expected KL information. In our setup we directly estimated nonparametrically the KL information for all the completing models and selected the one having minimum value (see Fig.3). The Kth model class, $\mathscr{M}_k$ is characterized by $h(x) = f(x)\left[Z_k^{-1}\exp[\sum_{i=1}^{k}\theta_i S_i(x)]\right]$, which is famously known as Neyman model (Neyman 1937). Which says that the pooled density is the product of density under class 1 and a exponential 'modification factor'. We are interested to find what is the best k using KL distance. The trick is to express the KL information in the quantile domain (replacing $X = Q(X;H)$) as a functional of comparison density (Parzen 1983), represented by Eq. 7. Our treatment thus *unifies* exponential smooth test, information theoretic approach and nonparametric quantile domain data analysis.

Real Data Example

In this section we will illustrate our methodology using Colon cancer microarray data (Alon et al.1999), consists of 2000 gene expressions (number of variables) measured in 62 samples (22 normal and 40 tissue samples). An important intermediate step to build efficient probabilistic classifier for medical decision making is variable selection, which we accomplish through our detector introduced in Eq (4) utilizing our specially designed score functions. First we create the score function using our recipe (see Eq 3), as shown in Figure 1. Figure 2. shows the ranked ordering of the variables according to their importance for classification. There is a significant gap after variable number 8, as shown in Figure 4b, which drastically reduce the original dimension of the problem.

Utilizing the proper number of score functions (using Corollary 1) , Figure 3 present the corresponding MaxEnt estimates of comparison density which plays a *fundamental* role in classification (See Parzen et al. 2011).

This article is copyrighted as indicated in the article. Reuse of AIP content is subject to the terms at: http://scitation.aip.org/termsconditions. Downloaded to IP: 155.247.166.234 On: Sat, 05 Jul 2014 15:34:39

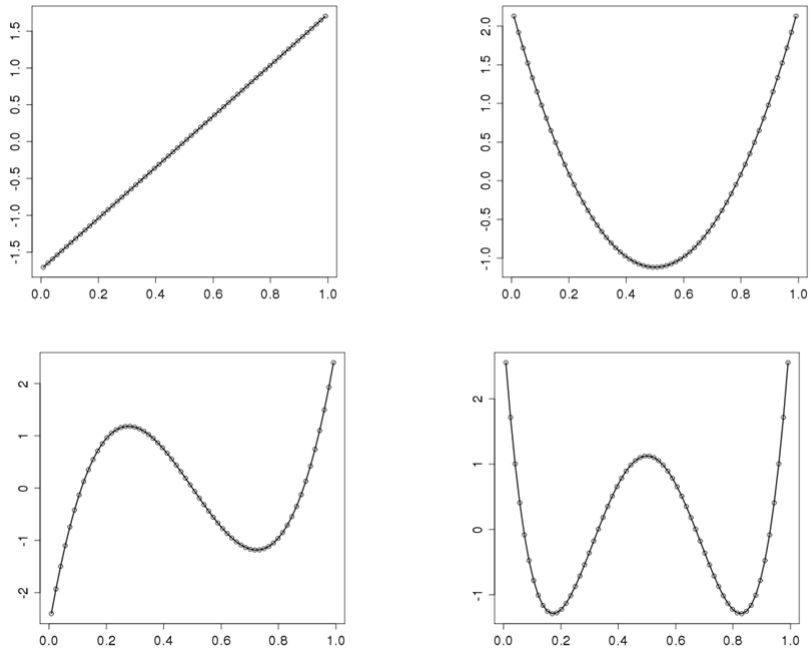

FIGURE 1. The shape of first four orthogonal mid-distribution score functions.

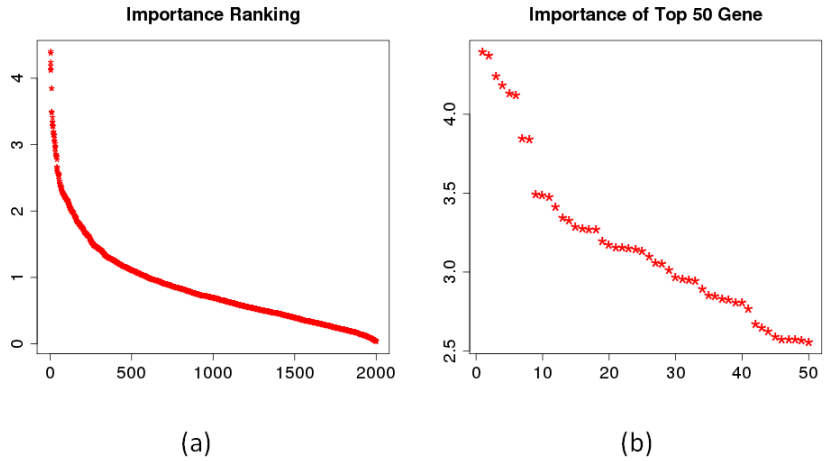

FIGURE 2. Ranking the variables according to their discriminative information using the measure introduced in Eq 4. Here we have used using $m = 4$ to generate the rankings. Fig. 4b, clearly separates the interesting variables from the the rest..

This article is copyrighted as indicated in the article. Reuse of AIP content is subject to the terms at: http://scitation.aip.org/termsconditions. Downloaded to IP: 155.247.166.234 On: Sat, 05 Jul 2014 15:34:39

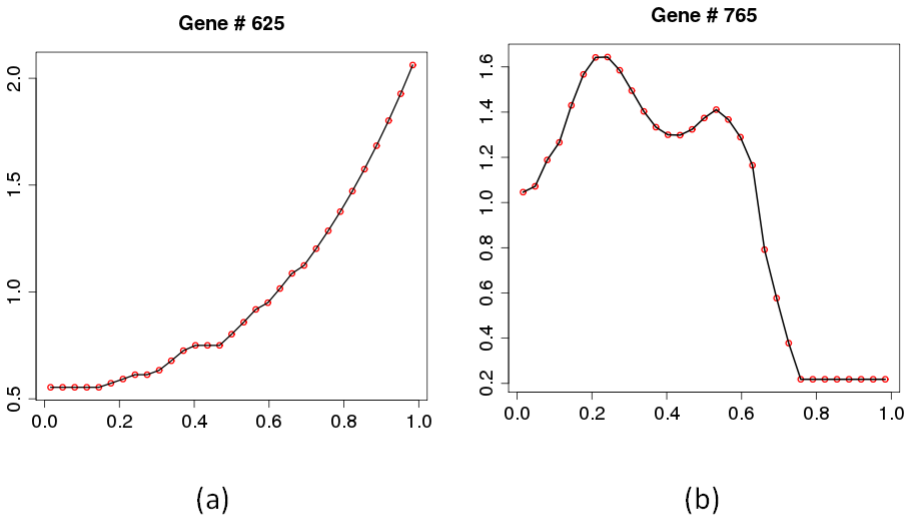

FIGURE 3. MaxEnt Comparison Density Estimates.

DISCUSSION

This paper lays the groundwork for building relevant constraints from data, in a unified manner through modern nonparametric methods and thus, takes care of the issues of model uncertainty and model fidelity. The key idea is to estimate the exponential model that uses a novel construction of orthogonal basis functions from the F^{mid}-transformed values which gives the method extra robustness. One more pleasing aspect of our method is that, it unifies the theory for discrete and continuous data (Parzen, 1991).
Our proposed method uses the information theoretic inference in conjunction with quantile based approach to describe a unified approach which uses optimization and approximation to develop methods which encompasses nonparametric, maximum entropy, estimation, testing parametric hypothesis and goodness of fit for model selection.
Although we have only illustrated our method for the classification setup, several interesting features are apparent. Our method opens up the possibility to generate useful features in terms of score functions for conditional or discriminative probabilistic (exponential family distribution) models like logistic regression (Mukhopadhyay, 2011), conditional random field etc.

ACKNOWLEDGMENT

The first author like to convey a heartfelt thanks to Prof. Parzen for suggesting this topic as well as for his inimitable guidance and infinite patience. The author also

This article is copyrighted as indicated in the article. Reuse of AIP content is subject to the terms at: http://scitation.aip.org/termsconditions. Downloaded to IP: 155.247.166.234 On: Sat, 05 Jul 2014 15:34:39

appreciates Dr. Philip Goyal (organizing committee of MaxEnt 2011) for all his help and support.

REFERENCES

1. Akaiyke, H. *Information theory and an extension of the maximum likelihood principle.* 2nd International Symposium on Information Theory (B. N. Petrov and F. CzAki, eds), 1973, pp. 267–281.
2. Ma, Y. and Genton, M. G. and Parzen, E. *Asymptotic properties of Sample Quantiles of Discrete Distributions*. Annals of the Institute of Statistical Mathematics, pp. 227–243
3. Neyman, Jerzy *'Smooth' Test for Goodness of Fit*, Skandinavsk Aktuarietidskrift, 1937, pp. 149–199.
4. Parzen, E. *Nonparametric statistical data modeling*. Journal of the American Statistical Association, 1979,pp. 105–131.
5. Parzen, E. *Quantiles, parametric-select density estimation, and bi-information parameter estimators*. New York: Springer Verlag, Proceedings of the 14th Annual Symposium on the Interface of Computer Science and Statistics, 1983,pp. 241–245.
6. Parzen, E. , *Unification of statistical methods for discrete and continuous data*. Michigan State University . Computer Science and Statistics: Proceedings of the Symposium on the Interface, 1991, pp. 235–242.
7. Parzen, E. , *Quantile Probability and Statistical Data Modeling*. Statistical Science, 2004, pp. 652–662.
8. Parzen, E. , S.Mukhopadhyay and S.N.Lahiri *Nonparametric Quantile based High dimensional Classifier: A Comparison Density Approach*. Technical Report, Texas A&M , 2010.
9. Parzen, E. , S.Mukhopadhyay and S.N.Lahiri *Revisiting Logistic Regression : A Modern Nonparametric Approach*. Technical Report, Texas A&M , 2011.
10. Mukhopadhyay, S., Parzen, E. and S.N.Lahiri *Quantile Based Variable Mining: Detection, FDR Extraction and Interpretation*. Technical Report, Texas A&M , 2011.
11. U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, and Levine Mack, D. *Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays*. PNAS, 1999,pp. 6745–6750.

This article is copyrighted as indicated in the article. Reuse of AIP content is subject to the terms at: http://scitation.aip.org/termsconditions. Downloaded to IP: 155.247.166.234 On: Sat, 05 Jul 2014 15:34:39