SELECTION OF DATA-GENERATING EXPERIMENTS, IDENTIFIABILITY AND EXPECTED P-VALUES

Yannis G. Yatracos

Yau Mathematical Sciences Center

Tsinghua University, Beijing

and

Beijing Institute of Mathematical Sciences and Applications

e-mail: yatracos@tsinghua.edu.cn yannis.yatracos@gmail.com

January 5, 2022

Running Head: Black-Box Selection

Some key words: Black-Box; Data-Generating Experiment; EDI-graphs; Empirical Discrimination Index (EDI); *g-and-k* model; Identifiability; Intractable models; Learning Machines; Proportion of P-values Index ($PPVI$); P-values; Tukey's *g-and-h* model

Summary

In a Data-Generating Experiment (*DGE*), the data, $\mathbf{X}$, is often obtained either from a Black-Box with inputs θ and $\mathbf{Y}$, or from a Quantile function or a learning machine, $f(\mathbf{Y}, \theta)$; θ is unknown, element of metric space (Θ, ρ), $\mathbf{Y}$ is random. If $\mathbf{X}$ has intractable or unknown *c.d.f.*, F_θ, non-identifiability of θ cannot be confirmed and when present, among others, limits the predictive accuracy of the learned model, $f(\mathbf{Y}, \hat{\theta})$; $\hat{\theta}$ estimate of θ. In Machine Learning, non-identifiability of θ is ubiquitous and its extent is a criterion for selecting a learning machine. Empirical indices, EDI and $PPVI$, are introduced using P-values of Kolmogorov-Smirnov tests: *i)* to confirm almost surely, using generated data, *the discrimination of* θ from θ^*, namely that the Kolmogorov distance, $d_K(F_\theta, F_{\theta^*})$, is positive, *ii)* to confirm identifiability of $\theta(\in \Theta)$ by repeating *i)* for θ^* in a sieve of Θ, since neighboring parameter values are in practice indistinguishable, and *iii)* most important, to compare *EDI-graphs* of *DGEs*, preferring more discrimination and less non-identifiability among parameters, and select one *DGE* to use. In applications, EDI-graphs confirm non-identifiability in mixture models and in models parametrised with sums of parameters. EDI and $PPVI$ explain why Tukey's *g-and-h* model (*DGE1*) has better g-discrimination than the *g-and-k* model (*DGE2*), unless the sample size is extremely large; $h_0 = k_0$. EDI-graphs indicate that Normal learning machines have better parameter discrimination than Sigmoid learning machines and their parameters are non-identifiable.

1 Introduction

Statistical modeling is used in numerous fields, from Economics and Psychology to Biology, Engineering and Machine Learning, among others. Often, it is assumed the data, $\mathbf{X}$, has known and tractable cumulative distribution function (*c.d.f.*), $F_\theta; \theta$ is unknown, element of metric space (Θ, ρ). Recently, $\mathbf{X}$ is also obtained either from a Quantile function or a learning machine, $f(\mathbf{Y}, \theta)$, with f known and intractable F_θ, or more generally from a "Black-Box", with unknown data-generating mechanism depending on θ and $\mathbf{Y}$; input, $\mathbf{Y}$, is either observed or latent.

Breiman (2001) called the Black-Box model *algorithmic,* observed that statisticians rarely adopt it, and commented: "If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools." Earlier, Tukey (1962, p.60) wrote: "Procedures of *diagnosis*, and procedures to *extract indications* rather than extract conclusions, will have to play a large part in the future of data analyses and *graphical techniques* offer great possibilities in both areas." Both suggestions have been widely adopted nowadays in Data Science. Such tools and their theoretical justifications are presented for Black-Box models in this work.

Modeling goals include estimation of θ, and for the Black-Box with output approximated by learning machine, $f(\mathbf{Y}, \theta)$, the accurate and *reproducible* prediction of future outputs of f. These goals depend on parameter *identifiability*, confirmed so far only when *c.d.fs* are tractable. Note that the parameter in model $\mathcal{F} = \{F_{\theta^*}, \theta^* \in \Theta\}$ is identifiable when for each θ and θ^* in Θ, $\theta \neq \theta^*$ implies $F_\theta \neq F_{\theta^*}$. A problem not frequently studied for intractable or algorithmic models is that, often, due to the shapes of F_{θ^*} in $\mathcal{F}$, the sample size, n, needed for small estimation error may be excessively large and the statistician may not be aware about it. There is no direct data-tool measuring *discrimination* of θ from θ^* by evaluating the (strong) distance between intractable F_θ and F_{θ^*}, and also confirm θ-identifiability. For example, using Tukey's *g-and-h* model and the *g-and-k* model, the difficulty in the discrimination (and estimation) of parameters has been studied via Maximum Likelihood estimates (Rayner and MacGillivray, 2002). The extent of non-identifiability and the level of discrimination of parameters in a model, $\mathcal{F}$, are criteria for the choice of the data-

generating machine used. Answers to the emerged questions will be provided.

In practice, θ is not distinguished from θ^*-values in an open ρ-ball, $N_\epsilon(\theta)$, centered at θ with small radius ϵ, and for intractable models the sample size needed for estimation error, ϵ, is unknown. Taking advantage of samplers, tools are introduced herein: *i)* to study for a Black-Box model, the discrimination of parameters θ and θ^* at ρ-distance greater than or equal to ϵ, by evaluating with data the (strong) distance $\tilde{\rho}(F_\theta, F_{\theta^*})$, *ii)* to confirm identifiability of θ using *i)* for all θ^* in $\Theta^*(\subset \Theta)$, and *iii)* most important, to select among several data-generating or learning machines with non-identifiable parameters which one to use, preferring less non-identifiability and more discrimination, namely larger $\tilde{\rho}(F_\theta, F_{\theta^*})$.

The main tool used for discriminating θ from θ^* is the empirical discrimination index, EDI_M, estimating the expected P-value, $EPV(\theta, \theta^*; n)$, for the Kolmogorov-Smirnov test of hypotheses $F_\theta = F_{\theta^*}$ against $F_\theta \neq F_{\theta^*}$, due to *EPV*'s special properties (Proposition 3.1); M samples of size n are used, $\tilde{\rho}$ is Kolmogorov distance, d_K. The motivation for EDI_M and its connection with *EPV* are described in section 3. Since in practice, Θ can be assumed to be compact or ρ-totally bounded, let $\theta_1^* = \theta$ and form a partition $N_\epsilon(\theta_2^*), \ldots, N_\epsilon(\theta_m^*)$ of $N_\epsilon^c(\theta)$; A^c denotes the complement of A. ϵ-identifiability of θ is confirmed *almost surely* by confirming its discrimination from each θ^* in *sieve*

$$\Theta^* = \{\theta_1^*, \ldots, \theta_m^*\}. \tag{1}$$

For practical purposes, ϵ-identifiability of θ will imply identifiability for all the parameters in $N_\epsilon(\theta)$. Global identifiability in Θ can be confirmed by repeating the approach for $\theta = \theta_i^*, i \neq 1$. Alternatively, EDI-graphs for one θ, with θ^* in sieve Θ^* and various sample sizes, n, may be enough as observed in Examples for which θ-identifiability implies $\tilde{\theta}$-identifiability for every $\tilde{\theta}$ in Θ.

This work was motivated from Statistical Learning of θ with *Matching* in a Data-Generating Experiment (*DGE*) (Yatracos, 2020, 2021). A *DGE* consists of the sample space of $\mathbf{X}$, the parameter space Θ and the data-generating mechanism, f, with inputs θ and random $\mathbf{Y}$. Matching Estimates of $\theta = (g, h)$ were satisfactory at g_0 for Tukey's *g-and-h* model (*DGE1*) but not for the *g-and-k* model(*DGE2*), for the same values of h, k. The questions were "Why?" and "How this problem could be predicted in a *DGE*?". The

second tool, used for *DGE1* and *DGE2* with same inputs, $(\mathbf{Y}, \theta)$, in $f_i, i = 1, 2$, applicable for learning machines, is the proportion, $PPVI$, of P-values of *DGE1* exceeding those of *DGE2*. If $PPVI$ is smaller than .5, *DGE1* is preferred, at least locally, at θ. In some cases, $PPVI$ gets much smaller than .5 before n gets too large.

In Example 4.1, EDI-graphs are presented for the normal and Cauchy models, to compare the discrimination of their parameters and to observe the form of *EDI*-graphs for identifiable parameters. In Examples 4.2 and 4.3, non-identifiability is confirmed with EDI-graphs, respectively, for the parameters in a normal model with their sum the mean, and for the mixture of two normal distributions with known variance. In Examples 4.4 and 4.5, Tukey's *g-and-h* model ($DGE1$) and the *g-and-k* model ($DGE2$) are compared using EDI and $PPVI$, to confirm that the former has better discrimination and is to be preferred as data-generating or learning machine, unless the sample size, n, is extremely large. In Example 4.6, EDI-graphs are depicted for Normal and Sigmoid data-generating machines, showing that both have unidentifiable parameters but the former has better parameter discrimination.

Rothenberg (1971) established conditions for parameter identifiability using F_θ's Fisher Information Matrix (*FIM*) for tractable $F_\theta, \theta \in \Theta$. Non-identifiable statistical models include mixture models (Hartigan, 1985), autoregressive moving averages (Veres, 1987) and change point problem (Csörgo and Horvath, 1997). In Statistical Machine Learning, non-identifiability of θ is ubiquitous and has deep influence in particular on the output, $f(\mathbf{Y}, \theta)$, of the Learning Machine, f. The estimate, $\hat{\theta}$, affects the capability of the *learned* model (or *representation*), $f(\mathbf{Y}, \hat{\theta})$, to predict future data (Ran and Hu, 2017). For example, in Deep Neural Networks it is preferable that when a network relearns with data from the same model, the obtained learned representation is *nearly similar* to $f(\mathbf{Y}, \hat{\theta})$. This wish motivated studying *linear identifiability in function space* (Roeder *et al.*, 2021) for tractable models with general exponential form, extending results on distinct models (Hyvärinen and Marioka, 2016, Hyvärinen *et al.* 2018). Other tools include *FIM*, the asymptotic order of the Likelihood Ratio test statistic of the MLE and the Kullback-Leibler divergence of tractable models, used among others by Fukumizu (2003), Watanabe (2001), Fukumizu and Amari (2000) and Ran and Hu (2014). A detailed review is presented in Ran and Hu

(2017), endorsing in the "Summary and Perspective" (p. 1196) the view in Breiman (2001) for algorithmic models and *especially* the tools needed, by adding: "This will become one of the most important issues for machines in the future."

Dempster and Schatzoff (D&S, 1965) treated P-value as random and used its expected value, the Expected Significance Level (*ESL*), as sensibility index for comparing several multivariate tests. D&S viewed *ESL* as "reasonable compromise" to the Neyman-Pearson theory that uses the power of the test which depends on the α-level. Sackrowitz and Samuel-Cahn (S&SC, 1999) used Expected P-Value (*EPV*) instead of *ESL*, to stress that P-value is random, suggested EPV as test's performance measure when it is difficult to evaluate the power function, and examined the P-values under the alternative for location and scale models. Since then, *EPV* and its estimates have been used to study the power of tests for parametric tractable models, but not for studying identifiability or discrimination of parameters with the Kolmogorov-Smirnov test for general, intractable models. Additional references on the use of P-values, related controversies and the Bayesian approach can be found in Shi and Yin (2021), where the connection between P-value and posterior probability is presented.

Empirical discrimination indices for θ based on tests' comparisons, provide useful information in estimation problems since in the elementary estimation problem, Θ consists of two parameters, θ and θ^*, and can be solved by testing. Stein (1964) showed inadmissibility of the usual estimator for the variance of a normal distribution with unknown mean using a test of hypotheses to obtain the improved estimate. Le Cam (1973) and Birgé (2006) used successfully multiple, simultaneous tests of simple hypotheses for estimation with infinite dimensional parameter space, Θ.

In section 2, previous results related indirectly to parameter discrimination are presented. In section 3, EPV's properties are presented and EDI and $PPVI$ are introduced along with the use of EDI-graphs. Applications follow in section 4. The reader may proceed directly to Figures 1-8 that depict EDI-graphs with informative captions.

2 Model Shapes and Parameter Discrimination

Tukey (1960,1962, 1977) used the *g-and-h* model to better fit data, $\mathbf{X}$, with heavy tails and skewness, and the λ-distribution to fit both symmetric and asymmetric data; the sample $\mathbf{X} = \{X_1, \ldots, X_n\}$. The idea is to model the quantiles of $\mathbf{X}$ directly and not via a density (Yan and Genton, 2019). Thus, $\mathbf{X}$ is seen as *modification* of data, $\mathbf{Y}$, using known data-generator f, with unknown parameter $\theta (\in R^p)$, *e.g.*, $\theta = (g, h)$,

$$X_i = f(Y_i, \theta), \ i = 1, \ldots, n; \tag{2}$$

$p \geq 1, \mathbf{Y} = \{Y_1, \ldots . Y_n\}$ is observed from a known model with known parameter η.

Tukey's quantile-modeling approach has been used since then in several fields, (2) evolved and has led to an abundance of *models' shapes* that would fit better $\mathbf{X}$. However, this abundance of model shapes near the underlying distribution increases the difficulty in estimating θ, and f in (2) should be examined before its use as data-generator.

For the *g-and-k* model (Haynes *et al.*, 1997) and the generalized *g-and-h* models, Rayner and MacGillivray (2002) confirmed the difficulty of the MLE *to discriminate* distributional shapes and parameters' values with small and moderate n : "... computational Maximum Likelihood procedures are very good for very large sample sizes, but they should not necessarily be assumed to be safe for even moderately large sample sizes" (p. 58); also, "... with moderately large positive (*i.e.* to the right) skewness, the MLE method fitting to the *g-and-k* distribution cannot efficiently discriminate between moderate positive values and small negative values of the kurtosis parameter." (p. 64).

For Tukey's asymmetric λ-distributions, with wider variety of distributional shapes than *g-and-h* and using the Moments estimation method it is observed: "An additional difficulty with the use of this distribution when fitting through moments, is that of *nonuniqueness*, where more than one member of the family may be realized when matching the first four moments ... " (Ramberg *et al.* 1979, Rayner and MacGillivray, 2002, p. 58).

These findings indicate, for *DGEs* with intractable or unknown models, the need to study the discrimination of parameters using tools independent from estimation methods.

3 Empirical Discrimination and Identifiability of parameters in Black-Box Models

Assume the sample, $\mathbf{X}$, in a *DGE* has either intractable or unavailable *c.d.f.* $F_\theta; \theta \in \Theta, \rho$ is metric on $\Theta, \mathbf{X} = \{X_1, \ldots, X_n\}; X_i \in R^d, d \geq 1, \Theta \subseteq R^p, p \geq 1$.

Definition 3.1 *For any two distribution functions F, G in $R^d, d \geq 1$, their Kolmogorov distance*

$$d_K(F, G) = \sup\{|F(y) - G(y)|; y \in R^d\}. \tag{3}$$

Definition 3.2 *For any sample $\mathbf{U} = (U_1, \ldots, U_n)$ of random vectors in $R^d, n\hat{F}_{\mathbf{U}}(u)$ denotes the number of U_i's with all their components smaller or equal to the corresponding components of $u(\in R^d)$. $\hat{F}_{\mathbf{U}}$ is the empirical c.d.f. of $\mathbf{U}$.*

θ is identifiable when for *any* $\theta^* \in \Theta, \theta^* \neq \theta$, it holds that $F_\theta \neq F_{\theta^*}$, verified, *e.g.*, when the Kolmogorov distance, $d_K(F_\theta, F_{\theta^*})$, is not zero. However, this cannot be confirmed with unavailable or intractable *c.d.fs*, F_θ and F_{θ^*}.

Definition 3.3 *θ is discriminated from θ^* when $\tilde{\rho}(F_\theta, F_{\theta^*}) > 0; \tilde{\rho}$ is a strong probability distance,* i.e., *when $\tilde{\rho}(F_\theta, F_{\theta^*}) = 0$, then $F_\theta = F_{\theta^*}$.*

The larger the distance $\tilde{\rho}(F_\theta, F_{\theta^*})$ is, the better the discrimination between θ and θ^* is. If θ is discriminated from all $\theta^* \in \Theta$, then θ is identifiable. In the sequel, $\tilde{\rho} = d_K$ is used.

Discrimination of θ and θ^* is studied for Black-Box models, taking advantage of samplers by drawing samples, $\mathbf{X}$ and $\mathbf{X}^*$ both of size n, with unknown or intractable $c.d.fs$, F_θ and F_{θ^*}, respectively. $d_K(F_\theta, F_{\theta^*})$ is estimated by $d_K(\hat{F}_{\mathbf{X}}, \hat{F}_{\mathbf{X}^*})$. To evaluate how large $d_K(F_\theta, F_{\theta^*})$ is, since $d_K(\hat{F}_{\mathbf{X}}, \hat{F}_{\mathbf{X}^*})$ is random, one could rely on its P-value exceeding t^*, for various observed t^*-values, and in particular on their average value, that has crucial properties for discriminating θ and θ^* when several, independent $\mathbf{X}, \mathbf{X}^*$ and t^* are used. The average's estimand is an expected value that is also a probability, $P(T_n > T_n^*)$, with

T_n and T_n^* Kolmogorov distances of empirical *c.d.fs*, obtained from the testing problem that follows; see (9) and (11).

Let

$$\eta = d_K(F_\theta, F_{\theta^*}),\ 0 \le \eta \le 1, \tag{4}$$

and consider the hypotheses

$$H : \eta = 0\ (i.e.\ F_\theta = F_{\theta^*}) \text{ against } H^* : \eta = \eta^* > 0\ (i.e.\ F_\theta \neq F_{\theta^*}). \tag{5}$$

Independent samples $\mathbf{X}$ and $\mathbf{X}^*$ of size n are obtained, respectively, from F_θ and F_{θ^*}, using the *DGE*'s sampler. Let

$$T_n = d_K(\hat{F}_{\mathbf{X}}, \hat{F}_{\mathbf{X}^*}), \tag{6}$$

and with abuse of notation use for T_n's *c.d.f.*, G_0 under H and G_{η^*} under H^*, instead of using, respectively, $F_\theta = F_{\theta^*}$ and $F_\theta \neq F_{\theta^*}$. H is rejected if T_n is large, $T_n > t^*$, or instead if

$$P_0(T_n > t^*) = 1 - G_0(t) \tag{7}$$

is small; t^* is usually the observed T_n-value under H. Instead of the P-value calculated for an observed t^* under H, its expected value, EPV, is used calculated under both H and H^* with $\mathbf{X}$ and $\mathbf{X}^*$ independent,

$$EPV(\theta, \theta^*; n) = \int_0^1 P_0(T_n > t^*) dG_{\eta^*}(t^*) = 1 - E_{\eta^*} G_0(T_n^*),\ \eta^* \ge 0. \tag{8}$$

In the right side of (8), T_n^* is the random variable with observed value t^* obtained also when $\eta^* = d_K(F_\theta, F_{\theta^*}) \neq 0$. Discrimination of θ and θ^* is confirmed almost surely with *EPV*'s estimate EDI, due to the *EPV*-properties *a)-c)* in Proposition 3.1, and is used to determine identifiability of θ. In D&S and S&SC, *a)* has been used as well as that, when T_n and T_n^* are independent, then

$$EPV(\theta, \theta^*; n) = 1 - E_{\eta^*} G_0(T_n^*) = P(T_n > T_n^*), \tag{9}$$

with T_n and T_n^* obtained, respectively, under H and H^*, or H and H.

Proposition 3.1 *Using independent T_n and T_n^* described as above and in (6)-(9),*
a) for every θ in Θ and for every sample size n, $E_0[1 - G_0(T_n^*)] = .5 = EPV(\theta, \theta; n)$,

b) for every $\eta^* \neq 0$, $\lim_{n\to\infty} E_{\eta^*}[1 - G_0(T_n^*)] = 0 = \lim_{n\to\infty} EPV(\theta, \theta^*; n)$, $\theta \neq \theta^*$,
c) if $\eta^* = d_K(F_\theta, F_{\theta^*}) < \eta^{**} = d_K(F_\theta, F_{\theta^{**}})$, *and* T_n^{**} *is defined as* T_n *in (6) using* $\mathbf{X}^{**}$ *from* $F_{\theta^{**}}$, *then for large* n,

$$EPV(\theta, \theta^*; n) = E_{\eta^*}[1 - G_0(T_n^*)] \geq E_{\eta^{**}}[1 - G_0(T_n^{**})] = EPV(\theta, \theta^{**}; n). \tag{10}$$

For each n, by the strong law of large numbers, the averages of P-values from M repeated, independent samples obtained by the sampler at θ and $\theta^* \in \Theta$, have *almost surely* the properties of their expected values in *a)-c)* of Proposition 3.1. These properties are the keys to discriminate θ from θ^*, for checking θ-identifiability using θ^* in sieve Θ^* of Θ and for choosing one among several data-generating machines.

Definition 3.4 *(EDI) Let* $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_M$ *be samples of size* n *obtained in a DGE with intractable or unavailable c.d.f.* F_θ, *and let* $\mathbf{X}_1^*, \ldots, \mathbf{X}_M^*$ *be samples of size* n *from* F_{θ^*}, *with* $\mathbf{X}_i, \mathbf{X}_i^*$ *independent,* $i = 1, \ldots, M$. *Let* PV_i *be the* P*-value for the two sided Kolmogorov-Smirnov test of* θ *against* θ^* *using* $\mathbf{X}_i, \mathbf{X}_i^*, i = 1, \ldots, M$. *The Empirical Discrimination Index (EDI) of* θ *and* θ^* *is*

$$EDI_M(\theta, \theta^*; n, DGE) = \frac{1}{M} \sum_{i=1}^{M} PV_i. \tag{11}$$

The smaller the EDI-value is, the easier θ *and* θ^* *are discriminated in the DGE.*
EDI *is also used instead of* EDI_M.

EDI in (11) can be used to compare two or more $DGEs$, in particular quantile experiments and learning machines generated with $\mathbf{Y}$ having the same distribution.

DGE Selection with EDI's Discrimination Criterion

Assume that $DGE1$ and $DGE2$ are indexed by a parameter $\theta \in \Theta$ (usually of similar nature, *e.g.*, location) and that samples as in Definition 3.4 are obtained. The discrimination of θ from θ^* is easier in $DGE1$ than $DGE2$ if

$$EDI_M(\theta, \theta^*; n, DGE1) < EDI_M(\theta, \theta^*; n, DGE2). \tag{12}$$

When (12) holds for some θ in Θ, then $DGE1$ is preferred. The EDI comparison can be used for $DGE1$ and $DGE2$ with parameters respectively θ, θ^* in Θ, ζ, ζ^* in $\mathcal{Z}$ and $\rho(\theta, \theta^*) = \rho_{\mathcal{Z}}(\zeta, \zeta^*)$; $\rho_{\mathcal{Z}}$ is distance in $\mathcal{Z}$.

Definition 3.5 *When $DGE1$ and $DGE2$ are generated each by quantile model (2) with latent variables $\mathbf{Y}$ from the same model, for each of the M tests in (11) the same $\mathbf{Y}, \mathbf{Y}^*$ can be used to generate the data and P-values are compared. The proportion of P-values index is*

$$PPVI_M(\theta, \theta^*; n, DGE1, DGE2) = \frac{\#\{PV_j(DGE2) < PV_j(DGE1), j = 1, \ldots, M\}}{M}. \quad (13)$$

$PPVI$ is also used instead of $PPVI_M$.

DGE Selection with $PPVI$ Discrimination Criterion

Using the previous notation, for $DGE1$ and $DGE2$ as in Definition 3.5, the discrimination of θ from θ^* is easier in $DGE1$ than $DGE2$ if

$$PPVI_M(\theta, \theta^*; n, DGE1, DGE2) < .5. \quad (14)$$

Remark 3.1 *When $\Theta \subseteq R^p$, in Definitions 3.4 and 3.5 indices are calculated at $\theta = (\theta_1, \ldots, \theta_i, \ldots, \theta_p)$ and $\theta^* = (\theta_1, \ldots, \theta_i + \epsilon^*, \ldots, \theta_p), i = 1, \ldots, p,\ \epsilon^* > 0$. Local discrimination of a subvector of θ parallels the concept of local identifiability of a subvector of θ; see,* e.g., *Ran and Hu (2017, Definition 5, p. 1161).*

When $d = 1$, the function *ks.test* in R provides the P-value for the Kolmogorov-Smirnov two-sample test of equality for θ and θ^*, with $\mathbf{X}$ and $\mathbf{X}^*$ from F_θ against F_{θ^*}, respectively. For $d = 2$ and $d > 2$, the approaches in Peacock (1983) and Polonik (1999) can be used to obtain P-values The theory in the latter was implemented by Glazer *et al.*(2012), who estimated high-density regions directly instead of using density estimates.

Use of EDI-graphs

For $\theta \in \Theta$ and $\epsilon(> 0)$ small enough such that parameters in $N_\epsilon(\theta)$ are indistinguishable, let $\Theta^* = \{\theta_1^*, \ldots, \theta_m^*\}$ be the sieve in (1), with $\theta \in \Theta^*$ but not necessarily equal to θ_1^*.

EDI-graphs of $EDI_M(\theta, \theta_i^*; n)$ against $\rho(\theta, \theta_i^*), i = 1, \ldots, m$, for various n :
A) show non-identifiability of θ (and other parameters) when either $EDI_M(\theta, \theta_i^*; n)$ does not decrease as $\rho(\theta, \theta_i^*)$ increases, or, when n increases, at least $EDI_M(\theta, \theta_i^* = \theta; n)$ and $EDI_M(\theta, \theta_j^*; n)$ remain larger from most of the remaining $EDI_M(\theta, \theta_k^*; n), i \neq j \neq k$,
B) indicate better discrimination of θ from θ^* with smaller $EDI(\theta, \theta^*; n)$, and
C) allow comparing data-generating machines with their EDI-graphs using *A)* and *B)*, preferring more discrimination and less non-identifiabiliy.
When $\Theta \subset R^p$, ρ is the usual Euclidean distance.

4 Applications

EDI-graphs for θ, with θ^* in the sieve Θ^*, are presented for models and learning machines with identifiable and non-identifiable parameters to see their differences and use. The y-axis is used for EDI-values and the x-axis for the Euclidean L_2-distance between θ and the sieve's elements. For the interpretation of EDI-graphs follow *A)-C)* at the end of the previous section. Figures 1-8 indicate that EDI-graphs for one θ and for θ^* in Θ^*, and for moderate and large sample size, n, can be sufficient for checking identifiability in Θ and parameter discrimination, thus confirming their usefulness.

EDI-graphs for statistical models are first presented.

Example 4.1 EDI-graphs for Normal (N) and Cauchy (C) models are depicted for studying discrimination and confirming identifiability of θ=$(\mu, \sigma) = (1.2, 1.6)$. The assumed parameter spaces for μ and σ are, respectively, $[0, 2]$ and $[0.4, 2.4]$, Θ is their Cartesian product. For $\epsilon = .4$, consider in each parameter space, respectively, sieve $\mu_j^* = .4(j-1), 1 \leq j \leq 6$ and $\sigma_k^* = .4 + .4(k-1), 1 \leq k \leq 6$, providing Θ-sieve, $\theta_i^* = (\mu_j^*, \sigma_k^*), i = 1, \ldots, 36$, that includes θ.

$M = 100$ independent samples $\mathbf{X}$ and $\mathbf{X}^*$, each of size n, are obtained with parameters, respectively, θ and θ_i^* in both models, and the corresponding $EDIs$ are calculated, $i = 1, \ldots, 36$, for $n = 100, 300, 600, 1000$. EDI-graphs appear in Figures 1 and 2. Identifiability of θ is indicated when $EDI(\theta, \theta; n)$ takes value near .5 and the other EDI values decrease

as the Euclidean distance of θ and θ^* increases, all eventually decreasing to zero as n increases except for $EDI(\theta, \theta; n)$.

Indicative $EDI(\theta, \theta^*; n)$-values are provided in Table 1, for $\theta^*_{22} = \theta = (1.2, 1.6)$ and $\theta^*_{23} = (1.2, 2)$, to observe better parameter discrimination for the Normal model.

EDI (θ, θ^*) **FOR NORMAL AND CAUCHY**				
n	$\theta^*_{22}(N) = \theta$	$\theta^*_{23}(N)$	$\theta^*_{22}(C) = \theta$	$\theta^*_{23}(C)$
100	.53	.41	.53	.48
300	.56	.20	.49	.35
600	.56	.05	.51	.21
1000	.47	.02	.53	.10

Table 1: EDI-values for θ^*_{23} indicate easier discrimination of θ with the Normal (N)-model than the Cauchy (C)-model, since their convergence rate to zero for the former is faster as n increases.

Example 4.2 EDI-graphs for Normal distributions of the form $N(a+b, 1)$ are depicted to confirm non-identifiability of the parameter $\theta = (a, b) = (1.2, 1.6)$. The assumed parameter spaces for a and b are, respectively, $[0, 2]$ and $[0.4, 2.4], \Theta$ is their Cartesian product. For $\epsilon = .1$, in each parameter space, consider $a^*_j = .1(j-1), 1 \leq j \leq 21$ and $b^*_k = .4 + .1(k-1), 1 \leq k \leq 21$, providing sieve $\theta^*_i = (a^*_j, b^*_k) \in \Theta, i = 1, \ldots, 21^2$.

$M = 100$ independent samples $\mathbf{X}$ and $\mathbf{X}^*$ each of size n are obtained with parameters, respectively, θ and θ^*_i, and the corresponding $EDIs$ are calculated, $i = 1, \ldots, 21^2$, for $n = 100, 1000, 30000, 100000$. EDI-graphs appear in Figure 3. For $n = 100$, several circles in the EDI-graph "jump" and have y-values near .5, as the distance on the x-axis from $\theta = (1.2, 1.6)$ increases. For $n = 1000$, the θ^*-values that indicate non-identifiability of θ have EDI values near .5 and the sum of the θ^* parameters is 2.8, as with θ. The θ^* with EDI-values near .2 indicate additional non-identifiable parameters and have sum of parameters 2.7. Even for $n = 100000$, the EDI-graph differs from the EDI-graph of Example 4.1 that indicates identifiability of θ.

Example 4.3 Let $N(\mu, \sigma^2)$ denote Normal *c.d.f.* with mean, μ, and variance, σ^2. Consider the normal mixture model, $\mathcal{F}$, with known variance, $\sigma^2 = 1$, and parameter $\theta = (p, \mu_1, \mu_2)$, which is non-identifiable,

$$\mathcal{F} = \{pN(\mu_1, 1) + (1-p)N(\mu_2, 1), (p, \mu_1, \mu_2) \in \Theta = [0,1]x[0,2]x[0,2]\}.$$

EDI is used to confirm algorithmically non-identifiability of $\theta = (.25, .4, 1.2)$. $S_p = \{0, .25, .5, .75, 1\}$ is a sieve for $[0,1]$ and $S_{\mu_1} = S_{\mu_2} = \{0, .4, .8, 1.2, 1.6, 2\}$ are sieves for $[0,2]$, and

$$\Theta^* = S_p x S_{\mu_1} x S_{\mu_2} = \{\theta_i^* : 1 \leq i \leq 180\} \tag{15}$$

is sieve for Θ, with $\theta = \theta_{46}^*$. $\theta_{128}^* = (.75, 1.2, .4)$ makes θ non-identifiable. Using $M = 100$, EDI-plots for $n = 100, 500, 15000, 30000$, appear in Figure 4. When $n = 100$, the EDI-values are spread in $[0, .5]$, but as n increases, *e.g.* when $n = 500$, several EDI-values are near zero. When $n = 15000$, there are 6 EDI-values far from 0. Those with similar values correspond to non-identifiable θ_{46}^* and θ_{128}^*, $\theta_{63}^* = (.25, 1.6, .8)$ and $\theta_{125}^* = (.75, .8, 1.6), \theta_{88}^* = (.5, .8, 1.2)$ and $\theta_{93}^* = (.5, 1.2, .8)$. When $n = 30000$, the EDI-values of the last two are nearly zero, those corresponding to indices 63 and 125 decrease, while those for indices 46 and 128 remain near .5, as expected.

In Figure 5, simulations are repeated when the means' coordinates in $\theta = (.25, .35, 1.25)$ are not in Θ^*. θ's closest element in the sieve is $\theta_{46}^* = (.25, .4, 1.2)$. The same pattern with Figure 4 is observed in Figure 5, except for the two larger EDI-values which decrease slower than the other EDI-values towards zero as n increases. Non-identifiable $\theta_{63}^*, \theta_{125}^*$ and $\theta_{88}^*, \theta_{93}^*$ have similar EDI-values, with those of the last two vanishing for $n \geq 15000$.

EDI-graphs are used also for studying data-generating machines.

Tukey's (1977) (a, b, g, h)*-model* accommodates data from non-Gaussian distribution, with g real-valued controlling skewness, non-negative h controlling tail heaviness and with location and scale parameters $a \in R, b > 0$. A vector of independent standard normal random variables, $\mathbf{Z} = (Z_1, \ldots, Z_n)$ and parameter values a, b, g, h are used to obtain

$$X_{1,i}(g,h) = a + b\frac{e^{gZ_i} - 1}{g} e^{.5hZ_i^2}, \ i = 1, \ldots, n; \tag{16}$$

$\mathbf{X}_1(g,h) = (X_{1,1}(g,h), \ldots, X_{1,n}(g,h))$.

The (a,b,g,k)-model (Haynes et al., 1997) includes distributions with more negative kurtosis than the normal distribution and some bimodal distributions (Rayner and MacGillivray, 2002, p. 58). Standard normal $Z_1, \ldots, Z_n$ and parameter values a, b, g, k are used to obtain

$$X_{2,i}(g,h) = a + b[1 + c \cdot \frac{1 - e^{-gZ_i}}{1 + e^{-gZ_i}}](1 + Z_i^2)^k Z_i, \ i = 1, \ldots, n; \tag{17}$$

$\mathbf{X}_2(g,k) = (X_{2,1}(g,k), \ldots, X_{2,n}(g,k))$, c is a parameter used to make the sample correspond to a density and usually $c = .8$.

Tukey's *g-and-h* model ($DGE1$) and the *g-and-k* model ($DGE2$) are now compared, initially *g-locally* with EDI and $PPVI$; see Remark 3.1.

Example 4.4 Normal sample, $\mathbf{Z}$, is used to obtain $\mathbf{X}_1(g,h), \mathbf{X}_2(g,k)$, respectively, from (16) and (17) with $g = 5, h = k = 2.5, a = 0, b = 1, c = .8$. Normal sample, $\mathbf{Z}^*$, also of size n, is used to obtain similarly $\mathbf{X}_1^*(g^*,h), \mathbf{X}_2^*(g^*,k)$ with $g^* = 3, h = k = 2.5, a = 0, b = 1$. The P-value for the Kolmogorov-Smirnov test for equality of g and g^* based on $\mathbf{Z}$ and $\mathbf{Z}^*$ is obtained for both *g-and-h* ($DGE1$) and *g-and-k* ($DGE2$) models. Both experiments are repeated $M = 1000$ times for $n = 50, 100, 200, 500, 1000, 1500, 2500, 5000, 10000, 35000, 40000$. The corresponding $EDIs$ for $DGE1$ and $DGE2$ are computed for each n, and counters measure for each n the number of times out of M the P-value for $DGE2$ is smaller than, or larger than, that of $DGE1$. Comparisons of $EDIs$ indicate that for the values g and g^* used and $n < 40000$, Tukey's *g-and-h* model has better g-discrimination than the *g-and-k* model. For 5% significant difference, a sample of size $n = 1500$ is needed for the *g-and-h* model, and $n = 2500$ is needed for the *g-and-k* model. The proportion of samples out of M for which the P-value(*g-and-k*) is smaller than the P-value(*g-and-h*) appears decreasing to zero as n increases and at $n = 40000$ there is a correction. It takes a sample of size $n_c = 40000$ for the parameters $g = 5$ and $g^* = 3$ for the *g-and-k* model to be discriminated as in Tukey's *g-and-h* model. The results appear in Table 2; those for $n = 40000$ hold also for $n = 50000, 100000$.

Remark 4.1 *The results in Example 4.4 for the g-and-k model suggested comparing also smooth histograms for this model. Visually, there is no discrimination between overlayed*

g-Local Discrimination: Tukey's *g-and-h* and *g-and-k* models				
n	*EDI (g-and-h)*	*EDI (g-and-k)*	*PPVI(g-and-h, g-and-k)*	*PPVI(g-and-k, g-and-h)*
50	5.05 e-01	5.31e-01	0.1540	0.2980
100	4.47e-01	4.92e-01	0.1762	0.4287
200	3.61e-01	4.38e-01	0.1864	0.5459
500	2.00e-01	3.15 e-01	0.1683	0.6845
1000	8.12e-02	1.836 e-01	0.1379	0.7807
1500	3.56e-02	1.083e-01	0.1150	0.8327
2500	6.21e-03	3.19e-02	0.0988	0.8745
5000	1.08e-04	1.86e-03	0.0680	0.9205
10000	7.31e-08	5.37e-06	0.0388	0.9379
35000	0	2.22e-20	0	1e-04
40000	0	0	0	0

Table 2: Model parameters: $g = 5, g^* = 3, h = k = 2.5$. EDIs and PPVIs for g are based on M=1000 repeats.

g-and-k smooth histograms with parameters, respectively, $g = 5$ *and* $g^* = 3.5$ *for various sample sizes less than or equal to* $n = 10000$ *and* $\mathbf{Z} = \mathbf{Z}^*$. *We did not use* $n > 10000$.

From the results in Example 4.4, Tukey's *g-and-h* model has better local discrimination than the *g-and-k* model, unless the sample size is very large. The findings in Table 2 extend those in Rayner and MacGillivray (2002), based on the data and not on a particular estimate, *e.g.* the MLE. The results are confirmed *g-globally* below, with EDI-graphs.

Example 4.5 Tukey's *g-and-h* and the *g-and-k* models are examined as Learning Machines with same parameters $a = 0, b = 1, \theta = (g, h) = (g, k) = (5, 2.5)$. It is assumed g is unknown, but $h = k = 2.5$ is known. To use the same R programs we considered parameter spaces for g and $h = k$, respectively, $[2, 5]$ and $[2.5, 2.5]$. The sieve for the first parameter space, with $\epsilon = .4$, is $g_j^* = 2 + .6(j-1), 1 \le j \le 6$, and $h_k^* = 2.5, 1 \le k \le 6$, therefore the Θ-sieve consists of $\theta_i^*, i = 1, \ldots, 36$.

$M = 100$ independent samples $\mathbf{X}$ and $\mathbf{X}^*$ each of size n are obtained with parameters,

respectively, θ and θ_i^* in both models, and the corresponding $EDIs$ are calculated, $i = 1, \ldots, 36$, for $n = 200, 1000$. EDI-graphs for both models appear in Figure 6 indicating identifiability. Comparison of the EDI-graphs for the same n value, indicates that Tukey's *g-and-h* model has better parameter discrimination than the *g-and-k* model. The findings confirm graphically the results in Example 4.4.

Example 4.6 Similar set-up as in Example 4.3 is used, with only difference that data, X, is obtained from the Normal learning machine that is convex combination of normal densities,

$$X = f(Z, \theta) = p\phi(Z - \mu_1) + (1 - p)\phi(Z - \mu_2). \tag{18}$$

(p, μ_1, μ_2) is element of $\Theta = [0, 1]x[0, 2]x[0, 2]$, with the same sieve, Θ^*. The model parameter $\theta = (.4, 1.2, 1.6) = \theta_{46}^*$, and Z is a standard Normal random variable. $M = 100$ learning samples are used for EDI and $n = 100, 500, 15000, 30000$. A similar experiment is examined for a Sigmoid learning machine, with ϕ in (18) replaced by

$$s(u) = \frac{1}{1 + e^{-u}},\ u \in R.$$

Figures 7 and 8 indicate non-identifiability and that the parameters are better discriminated with the data from the Normal learning machine.

5 Conclusion

For Black-Box and intractable data models, parameter identifiability cannot be confirmed. In Machine Learning, non-identifiability is ubiquitous and the resulting difficulty in the estimation of the parameters and the reproducibility of the learning models are not yet quantified using the data. Empirical discrimination index, $EDI(\theta, \theta^*)$, is used for θ^* in a sieve of Θ, to confirm with EDI-graphs almost surely identifiability of θ, and in Θ, or non-identifiability. EDI and $PPVI$ are useful tools: *a)* for selecting Data-Generating Experiments, in particular learning machines with non-identifiable parameters that have better parameter discrimination, and *b)* for indicating the sample size needed for a good estimate of a parameter to be informative, when identifiability holds.

6 Appendix

Lemma 6.1 *Assume X_n, Y_n are r.vs, $X_n \xrightarrow[n\to\infty]{Prob} 0, Y_n \xrightarrow[n\to\infty]{Prob} a > 0$, then,*
i)

$$\lim_{n\to\infty} P(X_n > Y_n) = 0. \tag{19}$$

ii) If in addition, $Y_n \xrightarrow[n\to\infty]{a.s.} a > 0$, and r.vs, $Y_n^* \xrightarrow[n\to\infty]{a.s.} a + \epsilon, \epsilon > 0$, then, for large n,

$$P(X_n > Y_n) \geq P(X_n > Y_n^*). \tag{20}$$

Proof: *i)* Let $0 < \epsilon < a$, then

$$P(Y_n < X_n) = P(Y_n < X_n, X_n > \epsilon) + P(Y_n < X_n, X_n \leq \epsilon) \leq P(X_n > \epsilon) + P(Y_n < \epsilon)$$

$$= P(X_n > \epsilon) + P(Y_n - a < \epsilon - a) \leq P(|X_n| > \epsilon) + P(|Y_n - a| > a - \epsilon) \xrightarrow[n\to\infty]{Prob} 0,$$

since X_n and Y_n converge in probability, respectively, to 0 and a, and $a - \epsilon > 0$.

ii) Let $0 < \delta < \epsilon$. Then, for large n, since *a.s.* $\lim_{n\to\infty}(Y_n^* - Y_n) = \epsilon > \delta > 0$, $(Y_n^* - Y_n)$ will eventually take positive values *a.s.* which are larger than δ,

$$P(X_n > Y_n^*) = P(X_n > Y_n + Y_n^* - Y_n) = P(X_n > Y_n) - P(Y_n < X_n \leq Y_n + Y_n^* - Y_n) < P(X_n > Y_n).$$

Proof of Proposition 3.1: *a)* $E_0 G_0(T_n^*) = .5$, either via integration by parts or since $G_0(T_n^*)$ is uniform random variable on $[0, 1]$ when $\eta^* = 0$.

b) Follows from Lemma 6.1 used for a *DGE*, under (6)-(9), with $X_n = T_n = d_K(\hat{F}_{\mathbf{X}}, \hat{F}_{\mathbf{X}^*})$ when $F_\theta = F_{\theta^*}$, and $Y_n = T_n^* = d_K(\hat{F}_{\mathbf{X}}, \hat{F}_{\mathbf{X}^*})$ when $F_\theta \neq F_{\theta^*}$. The assumptions of Lemma 6.1 hold by Glivenko-Cantelli Theorem using $a = d_K(F_\theta, F_{\theta^*}) \geq 0$, since

$$|d_K(\hat{F}_{\mathbf{X}}, \hat{F}_{\mathbf{X}^*}) - d_K(F_\theta, F_{\theta^*})| \leq d_K(\hat{F}_{\mathbf{X}}, F_\theta) + d_K(\hat{F}_{\mathbf{X}^*}, F_{\theta^*}). \tag{21}$$

From Lemma 6.1 *i)*, it follows by (9) that

$$\lim_{n\to\infty} EPV(\theta, \theta^*; n) = \lim_{n\to\infty} P[T_n > T_n^*] = 0.$$

c) Using the notation in *b)* and Lemma 6.1 *ii)*, with $Y_n = T_n^*, Y_n^* = T_n^{**}$ and, by Glivenko-Cantelli and (21), $a = \eta^* < \eta^{**} = a + \epsilon$, (10) follows using (9).

References

[1] Birgé, L. (2006) Model selection via testing: an alternative to (penalized) maximum likelihood estimators. *Ann. I. H. Poincaré* **42**, 273–325.

[2] Breiman, L. (2001) Statistical Modeling: The Two Cultures. *Stat. Science* **16**, **3**, 199-231.

[3] Csörgo, M. and Horvath, L. (1997) *Limit Theorems in Change-Point Analysis.* Wiley, New York.

[4] Dempster, A. P., and Schatzoff, M. (1965) Expected Significance Level as a Sensibility Index for Test Statistics, *JASA* **60**, 420-436.

[5] Fukumizu, K.(2003) Likelihood Ratio of non-identifiable Models and Multilayer Neural Networks. *Ann. of Stat*, **31**, 3, 833-851.

[6] Fukumizu, K. and Amari, S. (2000). Local minima and plateaus in hierarchical structures of multilayer perceptions. *Neural Networks* **13**, 317-327.

[7] Glazer, A., Lindenbaoum, M. and Markovitch, S. (2012) Learning High-Density Regions for a Generalized Kolmogorov-Smirnov Test in High-Dimensional Data. *Advances in Neural Information Processing Systems* **1**, 728-736.

[8] Hartigan, J. A. (1985). A failure of likelihood asymptotics for normal mixtures. *Proc. Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer (L. M. Le Camand R. A. Olshen, eds.)* **2**, 807–810. Wadsworth, Belmont, CA.

[9] Haynes M.A., MacGillivray H.L., and Mengersen K.L. (1997) Robustness of ranking and selection rules using generalized *g-and-k* distributions. *J. Statist. Plan. and Infer.* **65**, 45-66.

[10] Hyvärinen, A. and Morioka, H. (2016) Unsupervised Feature Extraction by Time-Contrastive Learning and Nonlinear ICA. *In Advances in Neural information Processing Systems,* 3765-3773.

[11] Hyvärinen,, A., Sasaki, H., and Turner, R. E. (2018) Nonlinear ICA Using Auxiliary Variables and Generalized Contrastive Learning. Arxiv Preprint Arxiv:1805.08651.

[12] LeCam, L. M. (1973) Convergence of estimates under dimensionality restrictions.*Ann. Statist.* **1**, 38–53.

[13] Lintusaari, J., Gutmann, M. U., Kaski, S. and Corander, J. (2016) On the Identifiability of Transmission Dynamic Models for Infectious Diseases. *Genetics* **202**, 911-918.

[14] Peacock, J. A. (1983) Two-dimensional goodness-of-fit testing in astronomy. *Monthly Notices Royal Astronomy Society* **202**, 615–627.

[15] Polonik, W. (1999) Concentration and goodness-of-fit in higher dimensions: (asymptotically) distribution-free methods. *Ann. of Stat.* **27**, 1210-1229.

[16] Ramberg, J. S., Tadikamalla, P. R., Dudewicz, E. J. and Mykytka, E. F. (1979) A probability distribution and its uses in fitting data. *Technometrics* **21**, 201–214.

[17] Ran, Z.-Y. and Hu, B.-G. (2017) Parameter Identifiability in Statistical Machine Learning: A Review. *Neural Computation* **29**, 1151-1203.

[18] Ran, Z.-Y. and Hu, B.-G. (2014) Determining parameter identifiability from the optimization theory framework: A Kullback-Leibler divergence approach. *Neurocomputing* **142**, 307-317.

[19] Rayner, G. D. and MacGillivray, H. L. (2002) Numerical maximum likelihood estimation for the *g-and-k* and generalized *g-and-h* distributions. *Statistics and Computing* **12**, 57-75.

[20] Roeder, G., Metz, L. and Kingma, D. P. (2021) On Linear Identifiability of Learned Representations. *Proceedings of the 38 th International Conference on Machine Learning,* PMLR 139, 2021. arXiv:2007.00810v3 [stat.ML] 8 Jul 2020

[21] Rothenberg, T. J. (1971) Identification in Parametric Models. *Econometrica* **39**, 577-591.

[22] Sackrowitz, H. and Samuel-Cahn, E. (1999) P-Values as Random Variables-Expected P-Values. *American Statistician* **53**, 326-331.

[23] Shi, H. and Yin, S. (2021) Reconnecting p-Value and Posterior Probability Under One-and Two-Sided Tests. *American Statistician* **75**, 3, 265-275.

[24] Stein, C. (1964) Inadmissibility of the usual estimator for the variance of a normal distribution with unknown mean. *Ann. Inst. Stat. Math.* **16**, 155–160.

[25] Tukey, J. W. (1986) Choosing techniques for the analysis of data. *The Collected Works of John Tukey— Philosophy and Principles of Data Analysis: 1965–1986 vol. 4*. Editor L. V. Jones, Ch. 24

[26] Tukey, J. W. (1977) Modern techniques in data analysis. NSF-sponsored regional research conference at Southeastern Massachusetts University, North Dartmouth, MA.

[27] Tukey, J. W. (1962) The Future of Data Analysis. *Ann. Math. Stat.* **33**, 1-67.

[28] Veres, S.(1987). Asymptotic distributions of likelihood ratios for overparameterized ARMA processes. *J. Time Ser. Anal.* **8**,3, 345–357.

[29] Watanabe, S. (2001) Algebraic Analysis of Nonidentifiable Learning Machines. *Neural Computation* **13**, 899-933.

[30] Yan, Y. and Genton, M. G. (2019) The Tukey g-and-h distribution. *Significance*, June 2019, 12-13.

[31] Yatracos, Y. G. (2021) Fiducial Matching for the Approximate Posterior: F-ABC. DOI: 10.13140/RG.2.2.20775.06568

[32] Yatracos, Y. G. (2020) Learning with Matching in Data-Generating Experiments. Submitted for publication. DOI: 10.13140/RG.2.2.30964.58245.

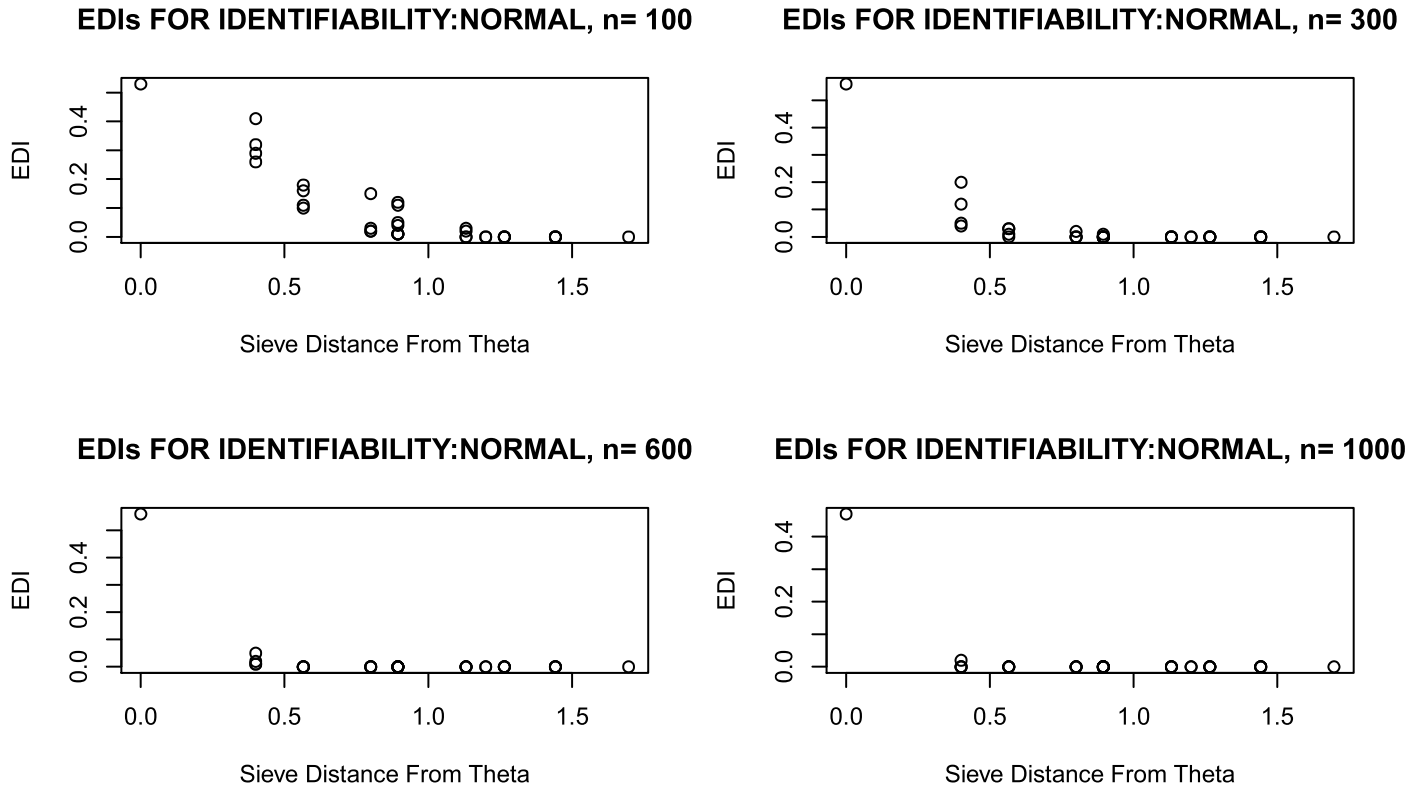

Figure 1: EDI-graphs indicate identifiability for the Normal model, seen also as learning machine. The graphs will be compared for discrimination with those of the Cauchy model in Figure 2.

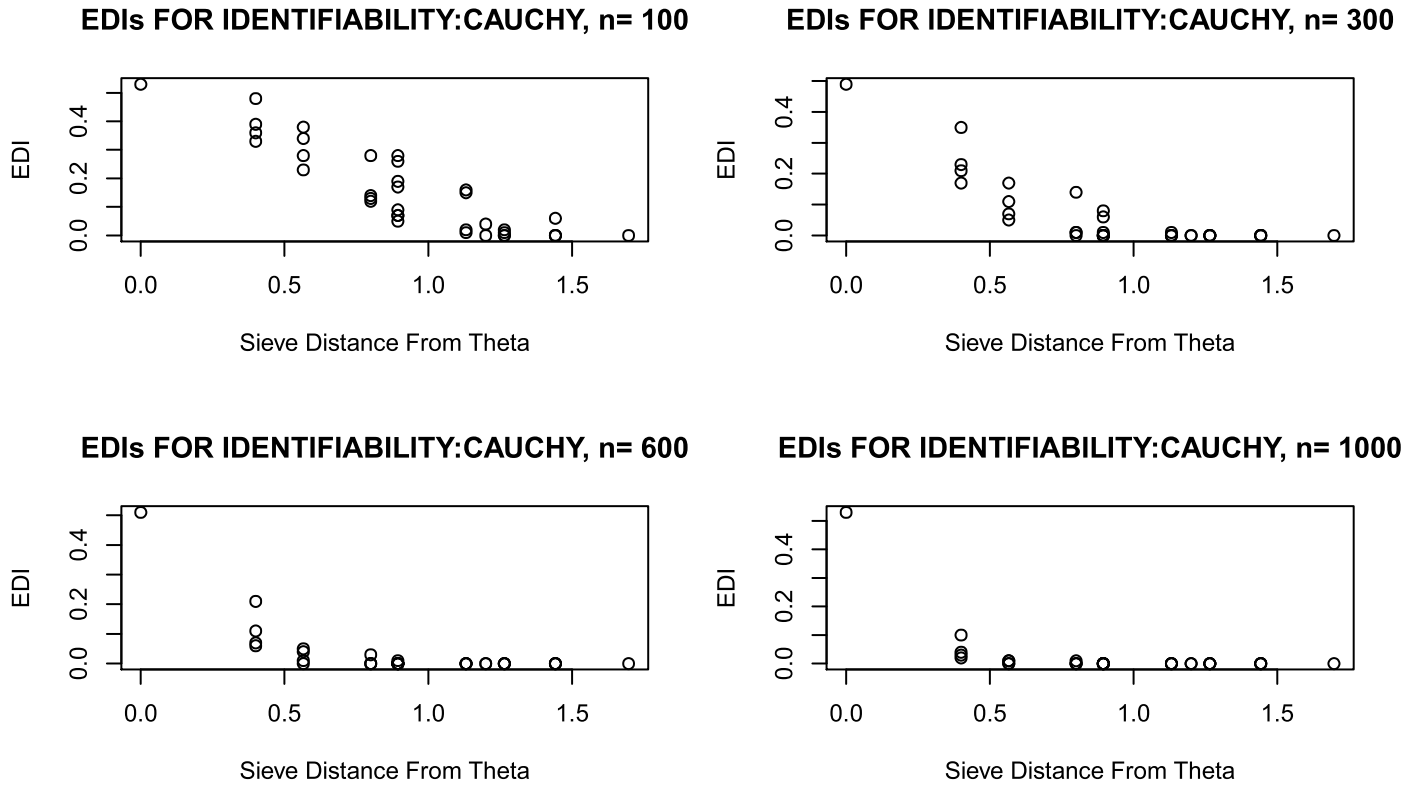

Figure 2: EDI-graphs indicate identifiability for the Cauchy model, seen also as learning machine. Comparison with Figure 1 indicates EDI-values of the Cauchy model are larger than or equal to those of the Normal model, for the same distance from θ and n-value, indicating better parameter discrimination for the latter.

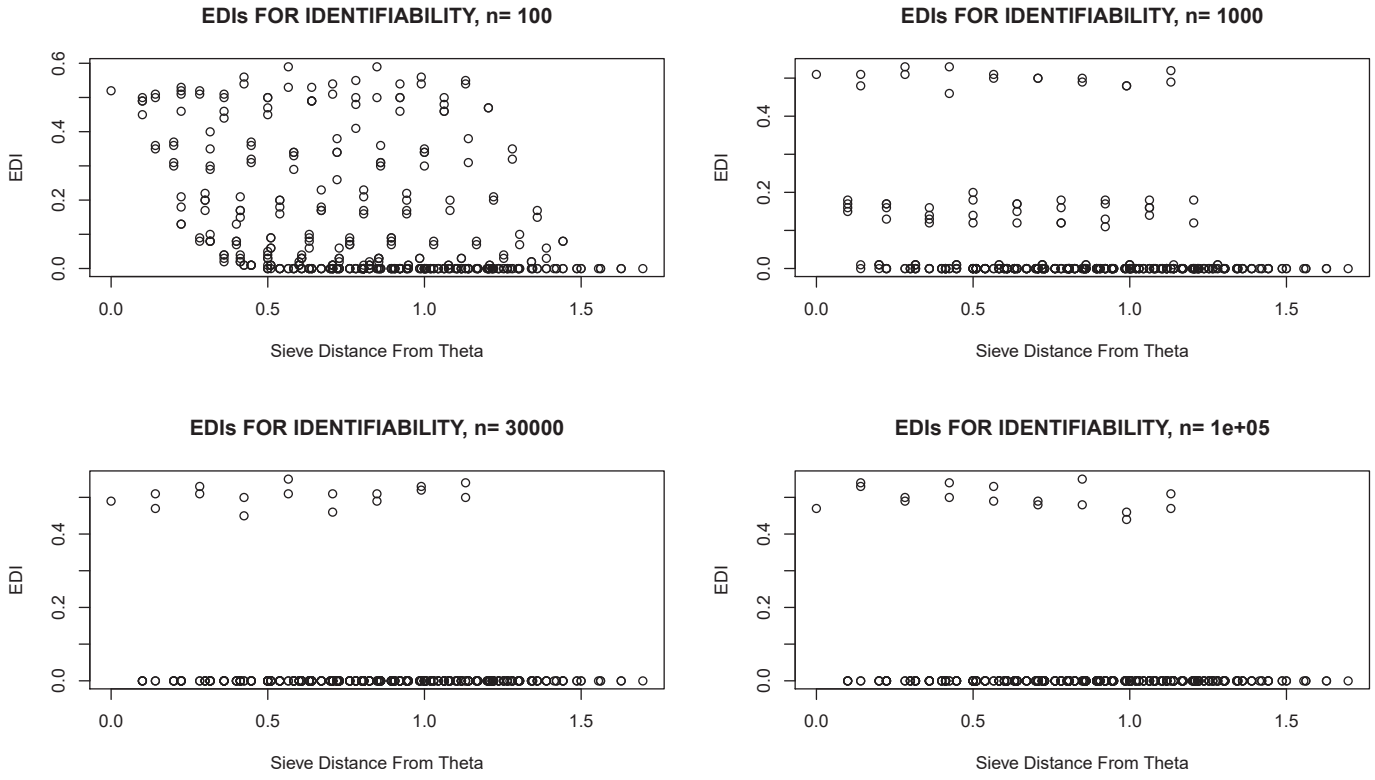

Figure 3: EDI-graphs indicate non-identifiable parameters for the normal model, $N(a+b,1)$, with $\theta = (a,b) = (1.2, 1.6)$. Several EDI-values near .5 as n increases, indicate non-identifiability, with $\theta^* = (a^*, b^*), a^* + b^* = 2.8$. When $n = 1000$, EDI-values near .2 correspond to non-identifiable θ^* with $a^* + b^* = 2.7$.

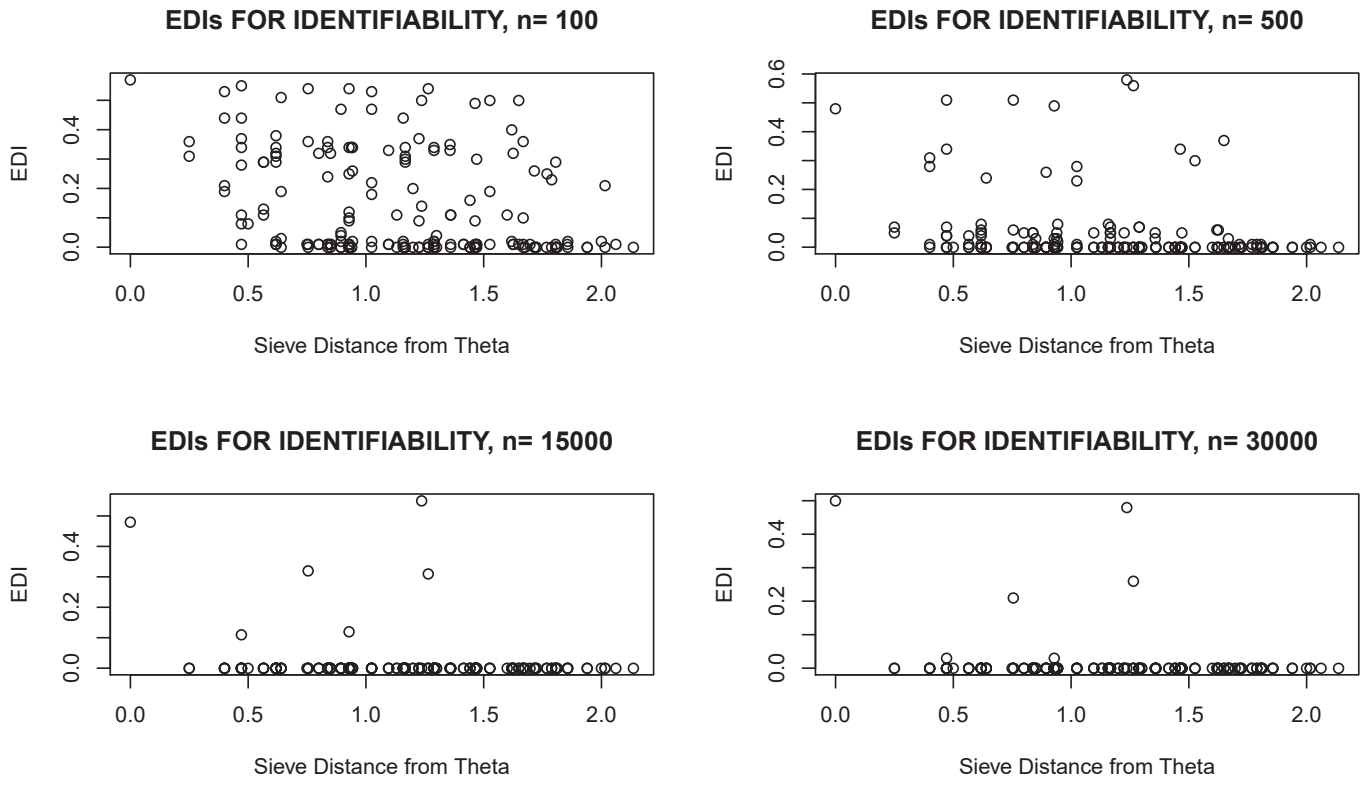

Figure 4: Non-identifiability of the normal mixture, $.25N(.4,1)+.75N(1.2,1)$, is confirmed by the two circles with y-coordinates near .5 for all sample sizes n. Circles with similar y-coordinates far from 0, as n increases, indicate non-identifiable parameters.

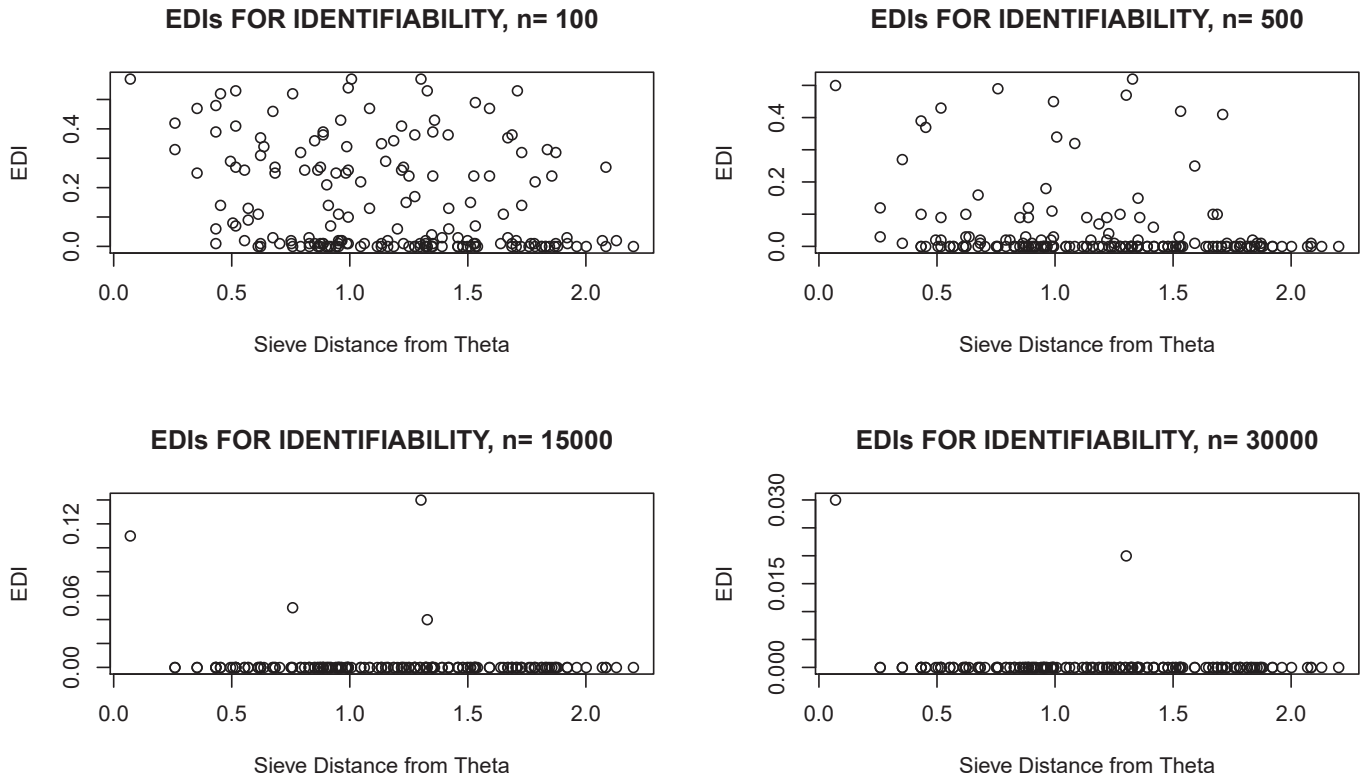

Figure 5: Non-identifiability of the normal mixture, $.25N(.35, 1) + .75N(1.25, 1)$, with the parameter $\theta = (.25, .35, 1.25)$ not included in the sieve, is confirmed by the two circles with the larger y-coordinates, which decrease as n increases. Circles with similar y-coordinates far from 0, as n increases, indicate non-identifiable parameters, as in Figure 4.

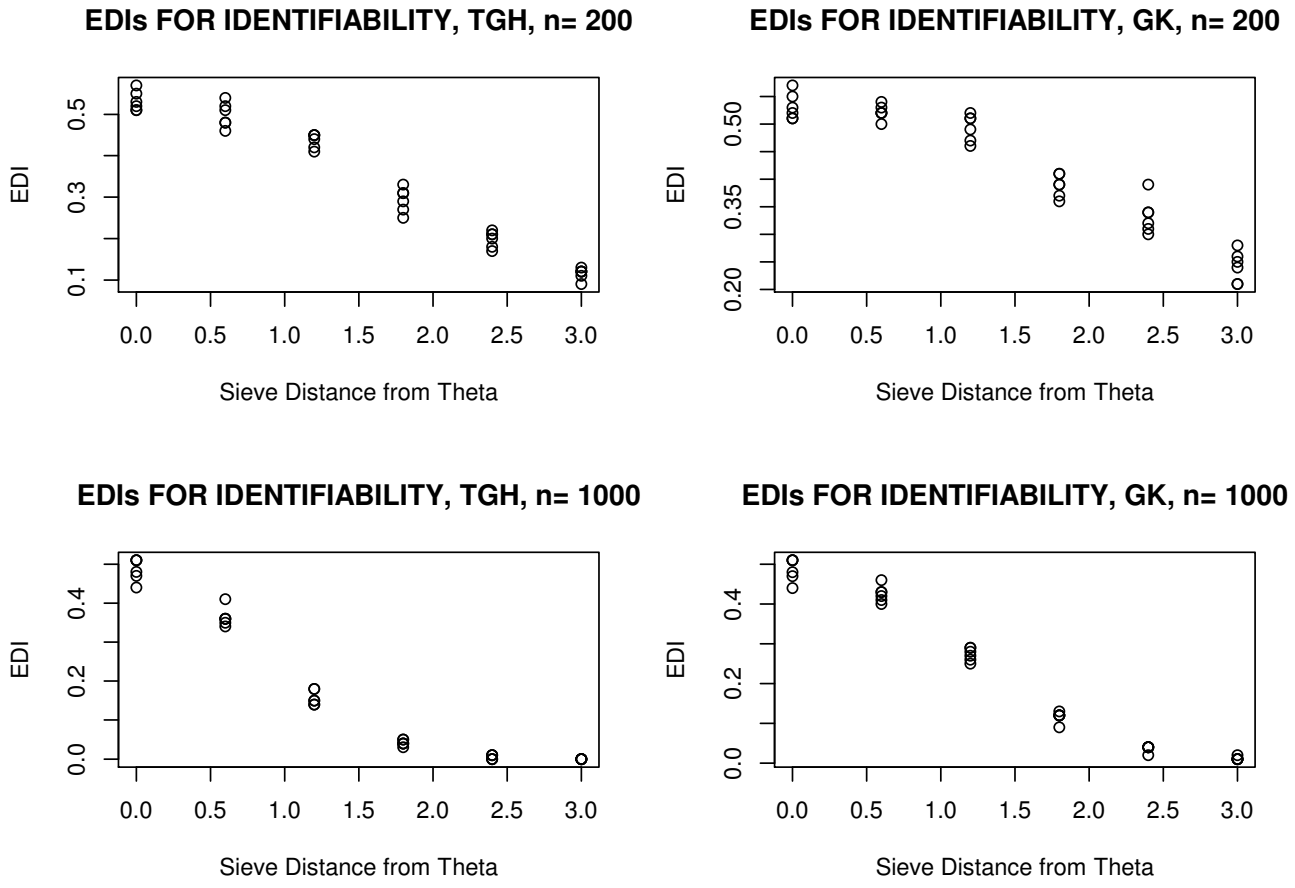

Figure 6: EDI-graph for comparison of learning machines: Tukey's *g-and-h* and the *g-and-k* models. Tukey's *g-and-h* has better parameter discrimination for moderate sample size, reconfirming Example 4.4.

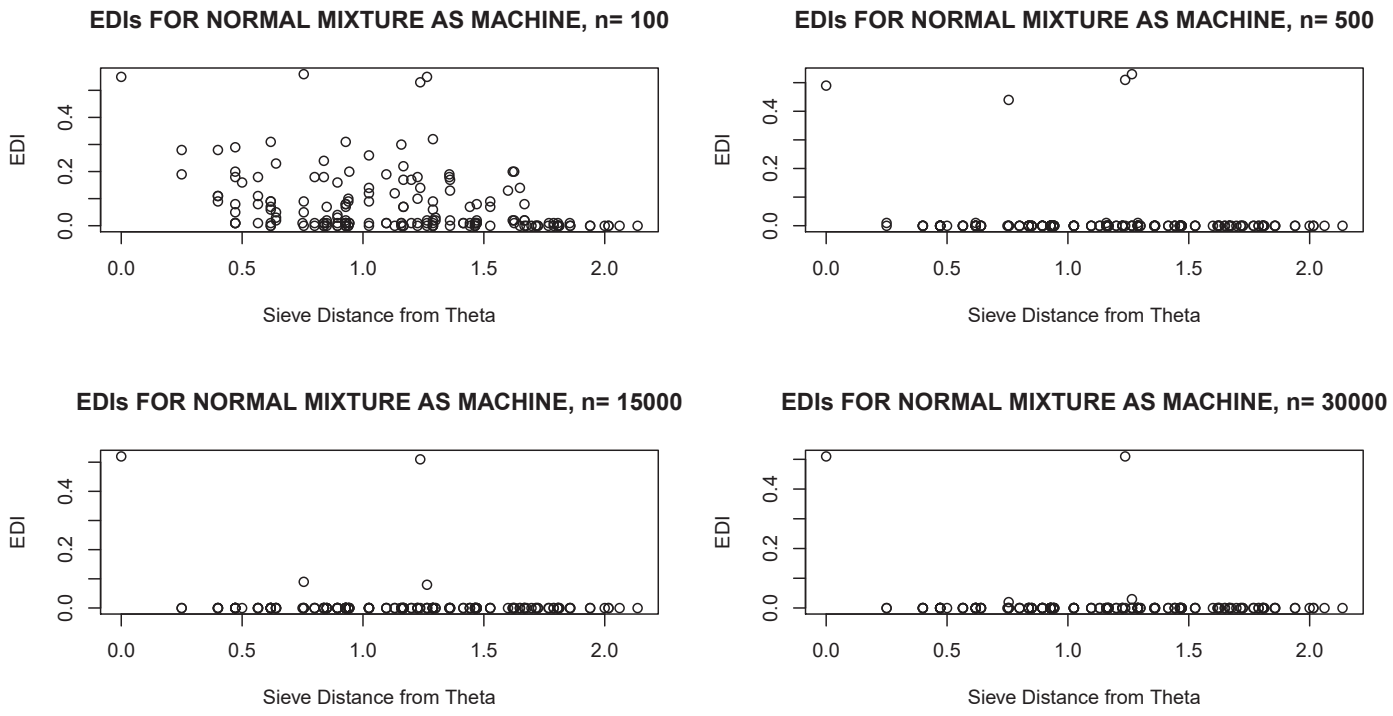

Figure 7: EDI-graphs showing non-identifiability for data, X, from a Normal learning machine, $X = f(Z, \theta) = p\phi(Z-\mu_1)+(1-p)\phi(Z-\mu_2), \theta = (p, \mu_1, \mu_2) = (.25, .4, 1.2) = \theta^*_{46}; Z$ is a standard Normal $r.v.$ with density ϕ.

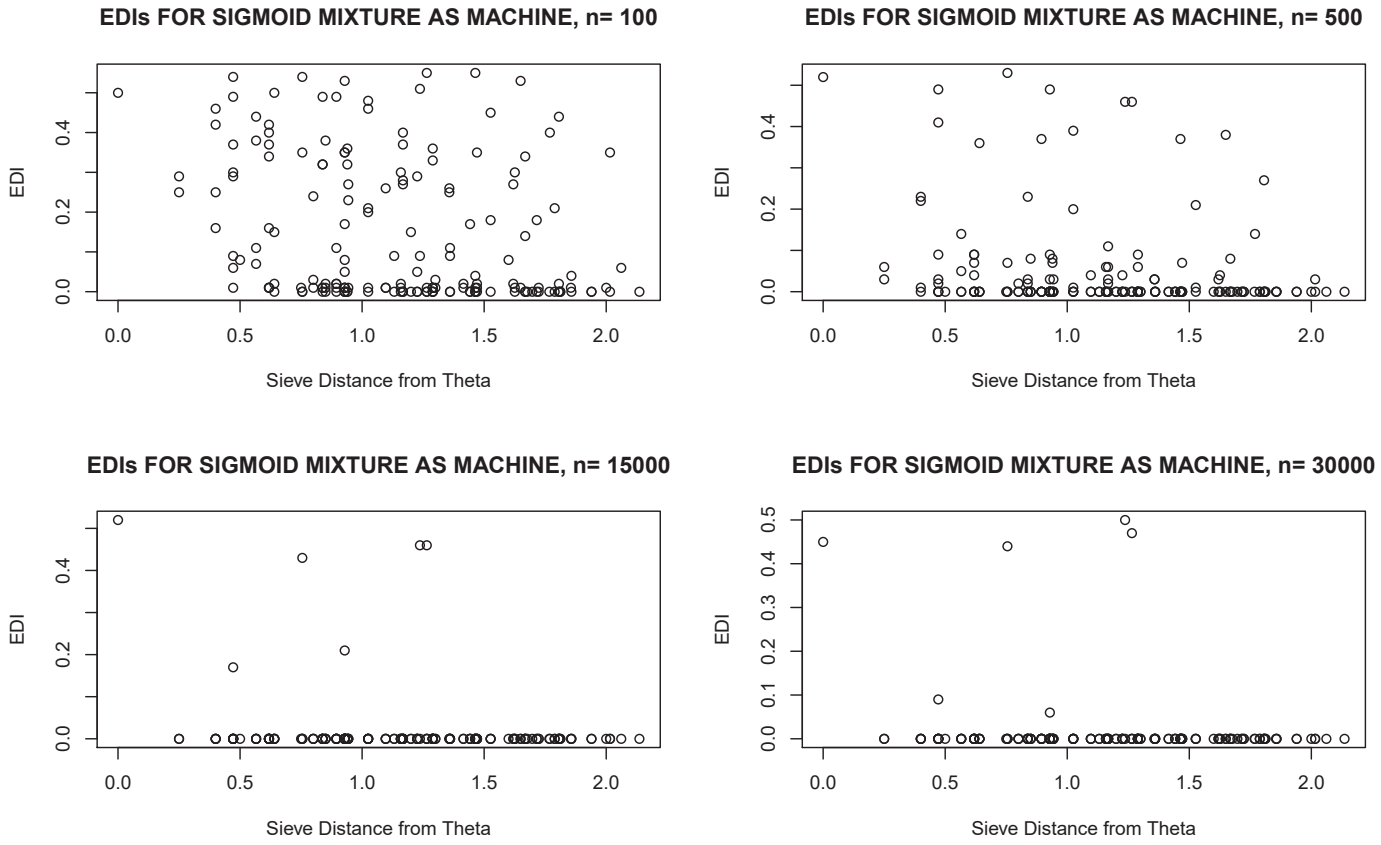

Figure 8: EDI-graphs showing non-identifiability for data, X, from a Sigmoid learning machine, $X = f(Z, \theta) = ps(Z-\mu_1)+(1-p)s(Z-\mu_2), \theta = (p, \mu_1, \mu_2) = (.25, .4, 1.2) = \theta^*_{46}; Z$ is a standard Normal $r.v.$, $s(u) = (1+e^{-u})^{-1}, u \in R$. Comparison with Figure 7 favors the Normal learning machine which shows better discrimination.