NONPARAMETRIC DETECTION OF CHANGES IN SYSTEM CHARACTERISTICS

Luc P. Devroye and Gary L. Wise
Department of Electrical Engineering
The University of Texas at Austin
Austin, Texas 78712

Abstract

A stochastic system with unknown structure and random inputs is considered. Two sequences of corresponding input-output pairs are observed over two disjoint observation intervals. It is desired to decide whether or not the system characteristics changed between the two observation periods. A localized version of the Kolmogorov-Smirnov statistic is introduced and discussed in this context.

1. INTRODUCTION

Frequently, the need arises to test whether the characteristics of a system are still the same, or whether they have changed. This problem is encountered in fault detection and quality control engineering. In this paper we consider a stochastic system transferring a random signal from $\mathbb{R}^d$ to $\mathbb{R}^c$. The structure of the system is unknown. We assume that the system is in operation and that we have observed (X_1,Y_1), $(X_2,Y_2),\ldots,$ (X_n,Y_n), a sequence of independent identically distributed $\mathbb{R}^{d+c}$-valued random vectors, where the X_i and Y_i are corresponding input-output pairs. The distribution of the X_i is governed by the input apparatus to the system and is assumed to be fixed but unknown. Later we observe (X_1',Y_1'), $(X_2',Y_2'),\ldots,(X_n',Y_n')$, another sequence of $\mathbb{R}^{d+c}$-valued random vectors, which represent corresponding input-output pairs during the latter observation period. If the characteristics of the system change, they are assumed to change between the two observation periods. The distribution of the input is assumed to be the same during both observation intervals. The system is said to have changed if (X_1,Y_1) and (X_1',Y_1') have different distribution functions. Notice that Y_1 and Y_1' may be identically distributed even though the system characteristics have changed.

A method based on the Kolmogorov-Smirnov statistic may be used in the above detection problem. This approach is briefly surveyed. Then a method based on a localized version of the Kolmogorov-Smirnov statistic is introduced and discussed. Its properties are illustrated with inequalities.

2. DISTANCE BETWEEN DISTRIBUTION FUNCTIONS

Let $Z = (X,Y)$ be an $\mathbb{R}^{d+c}$-valued random vector with distribution function F, and let $Z' = (X,Y')$ be an $\mathbb{R}^{d+c}$-valued random vector with distribution function G. Let H denote the distribution function of X. Define the metric ρ as

$$\rho(F,G) = \sup_z |F(z) - G(z)|.$$

Since X has the distribution function H in both cases, we intuitively feel that it must be possible to define the distance between F and G in terms of the conditional distribution functions of Y and Y' given X.

Presented at the Twentieth Midwest Symposium on Circuits and Systems, August 15-17, 1977; to be published in the Proceedings of the Symposium.

The regular conditional distribution function of Y given X=x, $F_x(y)$, which always exists [1], is for each x a distribution function on $\mathbb{R}^c$, and is for each $y\in\mathbb{R}^c$ a version of $P\{Y\le y|X=x\}$, that is, a Borel measurable function on $\mathbb{R}^d$ such that for all Borel sets A from $\mathbb{R}^d$,

$$\int_A F_x(y)\, dH(x) = P\{Y\le y,\ X\epsilon A\}.$$

Thus, we need not worry about the existence of F_x and G_x (the regular conditional distribution function of Y' given X=x).

One natural way to define the distance between F and G is as follows:

$$\begin{aligned} C(F,G) &= \operatorname*{ess\,sup}_H \ \sup_y |F_x(y) - G_x(y)| \\ &= \operatorname*{ess\,sup}_H \ \rho(F_x,G_x) \end{aligned}$$

where $\operatorname{ess\,sup}_H$ denotes the essential supremum with respect to the distribution function H. For the mapping C, as well as ρ, the triangle inequality holds. Also, C(F,G)=0 if and only if $\rho(F,G)=0$. Thus, C is a valid measure for the distance between F and G. Notice that

$$\begin{aligned} \rho(F,G) &= \sup_{x,y} \Big| \int_{w\le x} F_w(y)\, dH(w) - \int_{w\le x} G_w(y)\, dH(w)\Big| \\ &\le \sup_y \int |F_w(y) - G_w(y)|\, dH(w) \\ &\le \int \sup_y |F_w(y) - G_w(y)|\, dH(w) \\ &\le \operatorname*{ess\,sup}_H \rho(F_x,G_x) \\ &= C(F,G). \end{aligned}$$

Thus C tends to enhance the difference between F and G.

3. THE KOLMOGOROV-SMIRNOV STATISTIC

The reason for discussing distances between distribution functions is because they provide us with natural constructions for tests to detect changes in characteristics. Indeed, to do so, we use emperical estimates for $\rho(F,G)$. We recall that the (X_i,Y_i), $1\le i\le n$, have distribution function F; the (X_i',Y_i'), $1\le i\le n$, have distribution function G; and the X_i,X_i', $1\le i\le n$, have distribution function H. The empirical distribution functions with these samples are

$$\tilde{F}(x,y) = \frac{1}{n}\sum_{i=1}^{n} I_{\{X_i\le x,Y_i\le y\}} \tag{1}$$

and

$$\tilde{G}(x,y) = \frac{1}{n}\sum_{i=1}^{n} I_{\{X_i'\le x,Y_i'\le y\}},$$

where $I_{\{\cdot\}}$ is the indicator function. If $\varepsilon_n>0$ is a threshold, then the following detection rule is obvious:

decide $F\ne G$ if $\rho(\tilde{F},\tilde{G}) \ge \varepsilon_n$

decide F=G otherwise.

This test is attributed to Kolmogorov [2] and Smirnov [3] (see survey by Darling [4]).

The properties of the Kolmogorov-Smirnov statistic are well-known, in particular, its asymptotical properties as n grows large (for a survey, see Hajek and Sidak [5]). Less publicized are some strong inequalities valid for finite n. For instance, Devroye [6] has shown that for all F, $\theta>0$, $n\theta^2\ge(c+d)^2$,

$$P\{\rho(F,\tilde{F}) \ge \theta\} \le 2e^2(2n)^{c+d} \exp(-2n\theta^2), \tag{2}$$

whenever F is a distribution function on $\mathbb{R}^{c+d}$ and $\tilde{F}$ is defined by (1). (For other bounds, see [7-9].) But clearly, by the triangle inequality,

$$|\rho(F,G) - \rho(\tilde{F},\tilde{G})| \le \rho(F,\tilde{F}) + \rho(G,\tilde{G}).$$

Now assume first that F=G. Then if $n\varepsilon_n^2\ge4(c+d)^2$,

$$\begin{aligned} P\{\text{decide } F\ne G\} &= P\{\rho(\tilde{F},\tilde{G}) \ge \varepsilon_n\} \\ &\le P\{\rho(F,\tilde{F}) \ge \tfrac{\varepsilon_n}{2}\} + P\{\rho(G,\tilde{G}) \ge \tfrac{\varepsilon_n}{2}\} \\ &\le 4\ e^2(2n)^{c+d} \exp\left(-\tfrac{n}{2}\varepsilon_n^2\right). \end{aligned} \tag{3}$$

Thus, the probability of a false alarm decreases exponentially fast with n if ε_n is constant. Notice that (2) does not depend upon F. Assume next that $\rho(F,G)=\Delta>0$, that ε_n is so small that $\varepsilon_n\le\Delta/2$, and that n is so large that $n\Delta^2\ge16(c+d)^2$. Then

$$\begin{aligned} P\{\text{decide } F = G\} &= P\{\rho(\tilde{F},\tilde{G}) < \varepsilon_n\} \\ &\le P\{\rho(F,\tilde{F}) \ge \tfrac{\Delta}{4}\} + P\{\rho(G,\tilde{G}) \ge \tfrac{\Delta}{4}\} \\ &\le 4\ e^2(2n)^{c+d} \exp\left(-\tfrac{n}{8}\Delta^2\right). \end{aligned} \tag{4}$$

Since Δ is unknown, we must let ε_n decrease with n but not too fast so that (3) is small. That the bound (4) depends on $\Delta=\rho(F,G)$ is very normal.

Small changes require larger sample sizes to reduce the probability of making an erroneous decision. That the bound depends on nothing else but Δ is quite interesting.

From an engineering viewpoint, the computational requirements, as a function of n, for computing $\rho(\tilde{F},\tilde{G})$ are of the order of n^{d+c} (obtained by constructing the grids generated by the Z_i and the Z_i'). If d or c is large, this is clearly not feasible.

4. LOCALIZATION OF THE KOLMOGOROV-SMIRNOV STATISTIC

Recall that X_i and X_i' have the same distribution function H on $\mathbb{R}^d$. This fact will now be exploited. To do so, we need the distance C(F,G). Unlike in the previous section, it is impossible to compute

$$C(\tilde{F},\tilde{G}) = \operatorname{ess}_{\tilde{H}}\sup_{H} \rho(\tilde{F}_x,\tilde{G}_x)$$

for the simple reason that, although H is known, $\tilde{F}_x$ and $\tilde{G}_x$ are unknown. Thus the problem remains of the estimation of F_x and G_x. To solve this problem, we define permuted samples $(X_1^x,Y_1^x),\ldots,(X_n^x,Y_n^x)$ and $(X_1'^x,Y_1'^x),\ldots,(X_n'^x,Y_n'^x)$ where $x\in\mathbb{R}^d$. They are ordered such that

$$\|X_1^x-x\| \le \ldots \le \|X_n^x-x\|$$

$$\|X_1'^x-x\| \le \ldots \le \|X_n'^x-x\|$$

where $\|.\|$ denotes the L_2 norm on $\mathbb{R}^d$(for the case where $\|X_i-x\| = \|X_j-x\|$, we arbitrarily let X_i be closer to x if i<j). Estimate H by $\tilde{H}$ using $X_1,\ldots,X_n,X_1',\ldots,X_n'$. Estimate F_x and G_x by the following functions on $\mathbb{R}^c$:

$$\hat{F}_x(y) = \frac{1}{k_n}\sum_{i=1}^{k_n} I_{\{Y_i^x\le y\}},$$

and

$$\hat{G}_x(y) = \frac{1}{k_n}\sum_{i=1}^{k_n} I_{\{Y_i'^x\le y\}},$$

where $k_n\le n$ is a positive integer. What we are doing is assuming that F_x is close to $F_{X_i^x}$ if $\|X_i^x-x\|$ is small. Consider the statistic

$$C(\hat{F},\hat{G}) = \operatorname{ess}\sup_{\tilde{H}} \rho(\hat{F}_x,\hat{G}_x)$$

$$= \max_{1\le i\le n}\ \sup_y \left(|\hat{F}_{X_i}(y)-\hat{G}_{X_i}(y)| \vee |\hat{F}_{X_i'}(y)-\hat{G}_{X_i'}(y)|\right),$$

which, as a function of n, requires on the order of $n^2(k_n)^c(\log n)^2$ computations, a serious improvement over its counterpart $\rho(\tilde{F},\tilde{G})$. (The factor $n(\log n)^2$ arises from the search for the nearest neighbors [10].) Notice that the dimension d has disappeared from the exponent in the number of computations. Thus, the conditional distribution function approach seems adapted for multiple input single output systems.

5. PROPERTIES OF THE RULE BASED ON $C(\hat{F},\hat{G})$

In this section we consider the decision rule

Decide $F\ne G$ if $C(\hat{F},\hat{G}) \ge \varepsilon_n$

Decide F=G otherwise.

5.1 PROBABILITY OF ERROR WHEN F=G

Throughout we require two conditions, namely that the support B of H is compact (the support of a distribution function H is the smallest closed set B such that $\int_B dH(x) = 1$; equivalently, it is the set of all x such that every ε-sphere centered at x has positive probability), and that $\{F_x\}$ and $\{G_x\}$ are uniformly continuous collections of distribution functions with ρ, that is, for all $\varepsilon>0$, there exists a $\gamma(\varepsilon)>0$ such that $\|w-x\| < \gamma(\varepsilon)$ implies that

$$\rho(F_x,F_w) \le \varepsilon \quad\text{and}\quad \rho(G_x,G_w) \le \varepsilon.$$

These conditions are not too restrictive for practical systems.

We first need a lemma, relating F_x and G_x to

$$F_x^*(y) = \frac{1}{k_n}\sum_{i=1}^{k_n} F_{X_i^x}(y)$$

and

$$G_x^*(y) = \frac{1}{k_n}\sum_{i=1}^{k_n} G_{X_i'^x}(y).$$

A corollary of [6] is:

Lemma 1: Assume that $x\in\mathbb{R}^d$, $\varepsilon>0$, $1\le k_n\le n$, and $k_n\varepsilon^2\ge c^2$. Then both $P\{\rho(\hat{F}_x,F_x^*)\ge\varepsilon\}$ and $P\{\rho(\hat{G}_x,G_x^*)>\varepsilon\}$ are upper bounded by

$$2\,e^2\,(2k_n)^c\exp(-2k_n\varepsilon^2).$$

Thus we have the following:

$$P\{\text{decide } F\neq G\} = P\{C(\hat{F},\hat{G}) \geq \varepsilon_n\}$$

$$\leq P\{\operatorname{ess\,sup}_{\tilde{H}} |\rho(\hat{F}_x,\hat{G}_x)-\rho(F_x^*,G_x^*)| \geq \frac{\varepsilon_n}{2}\}$$

$$+ P\{\operatorname{ess\,sup}_{\tilde{H}} \rho(F_x^*,G_x^*) \geq \frac{\varepsilon_n}{2}\}$$

$$\leq 2n \sup_{x\varepsilon B} P\{\rho(\hat{F}_x,F_x^*)+\rho(\hat{G}_x,G_x^*) \geq \frac{\varepsilon_n}{2}\}$$

$$+ 2n \sup_{x\varepsilon B} P\{\rho(F_x^*,G_x^*) \geq \frac{\varepsilon_n}{2}\}$$

$$\leq 4n \sup_{x\varepsilon B} P\{\rho(\hat{F}_x,F_x^*) \geq \frac{\varepsilon_n}{4}\}$$

$$+ 4n \sup_{x\varepsilon B} P\{\|X_{k_n}^x - x\| \geq \gamma(\frac{\varepsilon_n}{4})\}$$

$$\leq 8n\, e^2\, (2k_n)^c \exp(-2\, k_n \frac{\varepsilon_n^2}{16})$$

$$+ 4n \sup_{x\varepsilon B} P\{\text{the number of } X_j \text{ in } S[x,\gamma(\frac{\varepsilon_n}{4})] \text{ is less than } k_n\}$$

when $k_n\, \varepsilon^2 \geq 16\, c^2$, where $S(x,\alpha)$ is a sphere in $\mathbb{R}^d$ centered at x with radius α. Let

$$m_n = \inf_{x\varepsilon B} \int_{S[x,\gamma(\frac{\varepsilon_n}{4})]} dH(w)$$

and assume that $k_n/n < m_n/2$. Then we have

$$P\{\text{decide } F\neq G\} \leq 8n\, e^2\, (2k_n)^c \exp(-k_n \frac{\varepsilon_n^2}{8}) + 4n \exp(-\frac{n\, m_n}{10})$$

If $\varepsilon_n = \varepsilon_1$ is constant, then $m_n = m_1$ is a constant also. Moreover, $m_n>0$ for all n, since for all $\alpha>0$

$$\inf_{x\varepsilon B} \int_{S(x,\alpha)} dH(w) > 0$$

by the compactness of B. The above inequality is valid if $k_n<n\, m_n/2$ and $k_n\, \varepsilon_n^2 \geq 16\, c^2$. In the derivation, use was made of Bennett's inequality for sums of independent identically distributed {0,1}-valued random variables [11].

5.2 PROBABILITY OF ERROR WHEN F≠G

Assume that $C(F,G)=\Delta>0$, that $\varepsilon_n<\Delta/2$, and that $k_n\, \Delta^2 \geq 144\, c^2$. Then

$$P\{\text{decide } F=G\}$$

$$\leq P\{\operatorname{ess\,sup}_{\tilde{H}} \rho(\hat{F}_x,\hat{G}_x) \leq \operatorname{ess\,sup}_{H} \rho(F_x,G_x) - \Delta/2\}$$

$$\leq P\{\operatorname{ess\,sup}_{\tilde{H}} |\rho(F_x,G_x) - \rho(F_x^*,G_x^*)| \geq \frac{\Delta}{6}\}$$

$$+ P\{|\operatorname{ess\,sup}_{\tilde{H}} \rho(F_x,G_x) - \operatorname{ess\,sup}_{\tilde{H}} \rho(F_x^*,G_x^*)| \geq \frac{\Delta}{6}\}$$

$$+ P\{\operatorname{ess\,sup}_{\tilde{H}} \rho(F_x,G_x) \leq \operatorname{ess\,sup}_{H} \rho(F_x,G_x) - \frac{\Delta}{6}\}$$

$$\leq 8n\, e^2\, (2k_n)^c \exp(-k_n \frac{\Delta^2}{72})$$

$$+ 4n \sup_{x\varepsilon B} P\{\|X_{k_n}^x - x\| \geq \gamma(\frac{\Delta}{12})\}$$

$$+ [1 - P\{\rho(F_{X_1},G_{X_1}) \geq \operatorname{ess\,sup}_{H} \rho(F_x,G_x) - \frac{\Delta}{6}\}]^{2n}$$

$$\leq 8n\, e^2 (2k_n)^c \exp(-k_n \frac{\Delta^2}{72}) + 4n \exp(-\frac{n\, m^*}{10}) + e^{-2n\delta}$$

where

$$m^* = \inf_{x\varepsilon B} \int_{S[x,\gamma(\frac{\Delta}{12})]} dH(w)$$

and $\delta = P\{\rho(F_{X_1},G_{X_1}) \geq \operatorname{ess\,sup}_{H} \rho(F_x,G_x) - \frac{\Delta}{6}\}$

are both positive numbers. If $\varepsilon_n=\varepsilon_1$ is constant, the inequalities show that we can meaningfully detect changes if $C(F,G) > 2\, \varepsilon_1$.

ACKNOWLEDGEMENT

This research was supported by the Department of Defense Joint Services Electronics Program under Contract F49620-77-C-0101.

REFERENCES

1. L. Breiman, Probability, Addison-Wesley, New York, 1968, pp. 77-79.
2. A. N. Kolmogorov, "Sulla Determinizione Empirica di una Legge di Distribuzione," Giorn. dell' Instituto Ital. degli Attuari, Vol. 4, pp. 83-91, 1933.
3. N. V. Smirnov, "Estimate of Deviation Between Empirical Distribution Functions in Two Independent Samples," (Russian) Bulletin Moscow Univ., Vol. 2, No. 2, pp. 3-16, 1939.
4. R. A. Darling, "The Kolmogorov-Smirnov, Cramér-von Mises Tests," Ann. Math. Stat., Vol. 28, pp. 823-838, 1957.
5. J. Hajek and Z. Sidak, Theory of Rank Tests, Academic Press, New York, 1967.

6. L. P. Devroye, "A Uniform Bound for the Deviation of Empirical Distribution Functions," to appear in J. Multivariate Analysis.

7. A. Dvoretzky, J. Kiefer, and J. Wolfowitz, "Asymptotic Minimax Character of the Sample Distribution Function and of the Classical Multinomial Estimator," Ann. Math. Stat., Vol. 27, pp. 642-669, 1956.

8. J. Kiefer and J. Wolfowitz, "On the Deviations of the Empiric Distribution Function of Vector Chance Variables," Trans. Am. Math. Soc., Vol. 87, pp. 173-186, 1958.

9. J. Kiefer, "On Large Deviations of the Empiric D.F. of Vector Chance Variables and a Law of the Iterated Logarithm," Pacific J. of Mathematics, Vol. 11, pp. 649-660, 1961.

10. J. H. Friedman, J. L. Bentley, and R. A. Finkel, "An Algorithm for Finding Best Matches in Logarithmic Time," Stanford Linear Accelerator Center Report No. SLAC-PUB-1549, Stanford University, February 1975.

11. W. Hoeffding, "Probability Inequalities for Sums of Bounded Random Variables," J. of the Am. Stat. Assoc., Vol. 58, pp. 13-30, 1963.

BIOGRAPHIES

Luc P. Devroye was born in Tienen, Belgium, on August 6, 1948. He obtained the M.S. and Ph.D. degrees from the Catholic University of Louvain, Belgium, and the University of Texas in 1971 and 1976, respectively. Supported by a Japanese Government Grant, he was involved in the study of probabilistic automata at the University of Osaka, Japan, from 1972 to 1974. In 1977 he became an Assistant Professor at the School of Computer Science, McGill University, Montreal, where he teaches nonparametric methods in discrimination and estimation. His research interests include random search, statistical pattern recognition, density estimation, and nonparametric statistics.

Gary L. Wise was born in Texas City, Texas, on July 29, 1945. He received the B.A. degree summa cum laude from Rice University in 1971 with a double major in electrical engineering and mathematics. He received the M.S.E., M.A., and Ph.D. degrees in electrical engineering from Princeton University in 1973, 1973, and 1974, respectively. He is presently an Assistant Professor of Electrical Engineering at the University of Texas at Austin. His research interests include statistical communication theory, random processes, and signal processing.

Dr. Wise is a member of Phi Beta Kappa, Tau Beta Pi, and Eta Kappa Nu.