Measuring the Influence of the \( k \)th Largest Variable on Functions over the Unit Hypercube

Jean-Luc Marichal and Pierre Mathonet

Mathematics Research Unit, FSTC, University of Luxembourg,
6, rue Coudenhove-Kalergi, L-1359 Luxembourg,
Grand Duchy of Luxembourg
jean-luc.marichal@uni.lu, pierre.mathonet@uni.lu

Abstract. By considering a least squares approximation of a given square integrable function \( f: [0, 1]^n \rightarrow \mathbb{R} \) by a shifted \( L \)-statistic function (a shifted linear combination of order statistics), we define an index which measures the global influence of the \( k \)th largest variable on \( f \). We show that this influence index has appealing properties and we interpret it as an average value of the difference quotient of \( f \) in the direction of the \( k \)th largest variable or, under certain natural conditions on \( f \), as an average value of the derivative of \( f \) in the direction of the \( k \)th largest variable. We also discuss a few applications of this index in statistics and aggregation theory.

1 Introduction

Consider a real-valued function \( f \) of \( n \) variables \( x_1, \ldots, x_n \) and suppose we want to measure a global influence degree of every variable \( x_i \) on \( f \). A reasonable way to define such an influence degree consists in considering the coefficient of \( x_i \) in the best least squares approximation of \( f \) by affine functions of the form

\[
g(x_1, \ldots, x_n) = c_0 + \sum_{i=1}^{n} c_i x_i.
\]

This approach was considered in [6,10] for pseudo-Boolean functions \( f: \{0, 1\}^n \rightarrow \mathbb{R} \) and in [9] for square integrable functions \( f: [0, 1]^n \rightarrow \mathbb{R} \). It turns out that, in both cases, the influence index of \( x_i \) on \( f \) is given by an average “derivative” of \( f \) with respect to \( x_i \).

Now, it is also natural to consider and measure a global influence degree of the smallest variable, or the largest variable, or even the \( k \)th largest variable for some \( k \in \{1, \ldots, n\} \). As an application, suppose we are to choose an appropriate aggregation function \( f: [0, 1]^n \rightarrow \mathbb{R} \) to compute an average value of \([0, 1]\)-valued grades obtained by a student. If, for instance, we use the arithmetic mean function, we might expect that both the smallest and the largest variables are equally influential. However, if we use the geometric mean function, for which the value 0
Similarly to the previous problem, to define the influence of the \(k\)th largest variable on \(f\) it is natural to consider the coefficient of \(x(k)\) in the best least squares approximation of \(f\) by symmetric functions of the form

\[
g(x_1, \ldots, x_n) = a_0 + \sum_{i=1}^{n} a_i x(i),
\]

where \(x(1), \ldots, x(n)\) are the order statistics obtained by rearranging the variables in ascending order of magnitude.

In this paper we solve this problem for square integrable functions \(f: [0, 1]^n \rightarrow \mathbb{R}\). More precisely, we completely describe the least squares approximation problem above and derive an explicit expression for the corresponding influence index (§2). We also show that this index has several natural properties, such as linearity and continuity, and we give an interpretation of it as an average value of the difference quotient of \(f\) in the direction of the \(k\)th largest variable. Under certain natural conditions on \(f\), we also interpret the index as an average value of the derivative of \(f\) in the direction of the \(k\)th largest variable (§3). We then provide some alternative formulas for the index to possibly simplify its computation (§4) and we consider some examples including the case when \(f\) is the Lovász extension of a pseudo-Boolean function (§5). Finally, we discuss a few applications of the index (§6).

We employ the following notation throughout the paper. Let \(\mathbb{I}^n\) denote the \(n\)-dimensional unit cube \([0, 1]^n\). We denote by \(L^2(\mathbb{I}^n)\) the class of square integrable functions \(f: \mathbb{I}^n \rightarrow \mathbb{R}\). For any \(S \subseteq [n] = \{1, \ldots, n\}\), we denote by \(1_S\) the characteristic vector of \(S\) in \(\{0, 1\}^n\) (with the particular case \(0 = 1_{\emptyset}\)).

Recall that if the \(\mathbb{I}\)-valued variables \(x_1, \ldots, x_n\) are rearranged in ascending order of magnitude \(x(1) \leq \cdots \leq x(n)\), then \(x(k)\) is called the \(k\)th order statistic and the function \(os_k: \mathbb{I}^n \rightarrow \mathbb{R}\), defined as \(os_k(x) = x(k)\), is the \(k\)th order statistic function. As a matter of convenience, we also formally define \(os_0 \equiv 0\) and \(os_{n+1} \equiv 1\). To stress on the arity of the function, we can replace the symbols \(x(k)\) and \(os_k\) with \(x_{k:n}\) and \(os_{k:n}\), respectively. For general background on order statistics, see for instance [1,4].

Finally, we use the lattice notation \(\wedge\) and \(\vee\) to denote the minimum and maximum functions, respectively.

2 Influence Index for the \(k\)th Largest Variable

An \(L\)-statistic function is a linear combination of the functions \(os_1, \ldots, os_n\). A shifted \(L\)-statistic function is a constant plus an \(L\)-statistic function. Denote by \(V_L\) the set of shifted \(L\)-statistic functions. Clearly, \(V_L\) is spanned by the linearly independent set

\[
B = \{os_1, \ldots, os_n, os_{n+1}\}
\]