THE TWO-SAMPLE PROBLEM IN $\mathbb{R}^m$ AND MEASURE-VALUED MARTINGALES

J.H.J. EINMAHL[1] AND E.V. KHMALADZE[2]

Eindhoven University of Technology and Eurandom, and University of New South Wales and A. Razmadze Mathematical Institute

The so-called two-sample problem is one of the classical problems in mathematical statistics. It is well-known that in dimension one the two-sample Smirnov test possesses two basic properties: it is distribution free under the null hypothesis and it is sensitive to 'all' alternatives. In the multidimensional case, i.e. when the observations in the two samples are random vectors in $\mathbb{R}^m$, $m \geq 2$, the Smirnov test loses its first basic property. In correspondence with the above, we define a solution of the two-sample problem to be a 'natural' stochastic process, based on the two samples, which is (α) asymptotically distribution free under the null hypothesis, and which is, intuitively speaking, (β) as sensitive as possible to all alternatives. Despite the fact that the two-sample problem has a long and very diverse history, starting with some famous papers in the thirties, the problem is essentially still open for samples in $\mathbb{R}^m, m \geq 2$. In this paper we present an approach based on measure-valued martingales and we will show that the stochastic process obtained with this approach is a solution to the two-sample problem, i.e. it has both the properties (α) and (β), for any $m \in \mathbb{N}$.

AMS subject classifications: 62G10, 62G20, 62G30; secondary 60F05, 60G15, 60G48.
Keywords and phrases: Dirichlet (Voronoi) tessellation, distribution free process, empirical process, measure-valued martingale, non-parametric test, permutation test, two-sample problem, VC class, weak convergence, Wiener process.

1 Introduction

Suppose we are given two samples, that is, two independent sequences $\{X_i'\}_1^{n_1}$ and $\{X_i''\}_1^{n_2}$ of i.i.d. random variables taking values in m-dimensional Euclidean space $\mathbb{R}^m, m \geq 1$. Denote with P_1 and P_2 the probability distributions of each of the X_i' and X_i'' and write $\widehat{P}_{n_1}$ and P_n for the empirical distributions of the first sample and of the pooled sample $\{X_i'\}_1^{n_1} \cup \{X_i''\}_1^{n_2}$ respectively, i.e.

$$\widehat{P}_{n_1}(B) = \frac{1}{n_1} \sum_{i=1}^{n_1} \mathbb{1}_B(X_i'), \tag{1.1}$$

$$P_n(B) = \frac{1}{n} \left(\sum_{i=1}^{n_1} \mathbb{1}_B(X_i' i) + \sum_{i=1}^{n_2} \mathbb{1}_B(X_i'') \right), \quad n = n_1 + n_2,$$

[1]Research partially supported by European Union HCM grant ERB CHRX-CT 940693.
[2]Research partially supported by the Netherlands Organization for Scientific Research (NWO) while the author was visiting the Eindhoven University of Technology, and partially by the International Science Foundation (ISF), Grant MXI200.