Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

# On normal approximations for the two-sample problem on multidimensional tori

Solesne Bourguin [a,*], Claudio Durastanti [b]

[a] Boston University, Department of Mathematics and Statistics, 111 Cummington Mall, Boston, MA 02215, USA
[b] Ruhr University Bochum, Faculty of Mathematics, D-44780 Bochum, Germany

ARTICLE INFO

*Article history:*
Received 11 May 2016
Received in revised form 10 October 2017
Accepted 14 October 2017
Available online xxxx

*MSC:*
60F05
60G55
62G09
62G10
62G20

*Keywords:*
Two-sample problem
*U*-statistics
Poisson processes
Wavelets
Needlets
Harmonic analysis on the torus
Stein–Malliavin method

ABSTRACT

In this paper, quantitative central limit theorems for *U*-statistics on the *q*-dimensional torus defined in the framework of the two-sample problem for Poisson processes are derived. In particular, the *U*-statistics are built over tight frames defined by wavelets, named toroidal needlets, enjoying excellent localization properties in both harmonic and frequency domains. The rates of convergence to Gaussianity for these statistics are obtained by means of the so-called Stein–Malliavin techniques on the Poisson space, as introduced by Peccati et al. (2011) and further developed by Peccati and Zheng (2010) and Bourguin and Peccati (2014). Particular cases of the proposed framework allow to consider the two-sample problem on the circle as well as the local two-sample problem on $\mathbb{R}^q$ through a local homeomorphism argument.

## 1. Introduction

The aim of this paper is to establish quantitative central limit theorems, by means of Stein–Malliavin techniques, for wavelet-based U-statistics on $q$-dimensional tori arising in the context of the two-sample problem for Poisson processes. The two-sample (or homogeneity) problem for Poisson processes can be described as follows: let $N_1$ and $N_2$ denote two independent Poisson processes observed on a measurable space $\mathbb{X}$, whose intensities with respect to some positive non-atomic $\sigma$-finite measure $\mu$ are denoted by $f_1$ and $f_2$, respectively. Given the observation of $N_1$ and $N_2$, the two sample problem aims at testing the null-hypothesis $(H_0) : f_1 = f_2$ versus the alternative hypothesis $(H_1) : f_1 \neq f_2$, see for instance Fromont et al. (2013) for an in-depth description. In such a problem, two-sample *U*-statistics arise very naturally (see for instance Fromont et al., 2013; van der Vaart, 1998; DasGupta, 2008) as they can be used to approximate both the distribution under the null as well as the alternative hypotheses (note that in the case of the alternative hypothesis, one has to deal with different underlying distributions for the Poisson processes, see e.g. Lee (1990, Example 1, Chapter 2, p. 38)).

[*] Corresponding author.
*E-mail addresses:* solesne.bourguin@gmail.com (S. Bourguin), claudio.durastanti@gmail.com (C. Durastanti).

https://doi.org/10.1016/j.jspi.2017.10.009
0378-3758/© 2017 Elsevier B.V. All rights reserved.

This paper assumes a slightly different, but equivalent framework in which the null-hypothesis $(H_0)$ is that observations of a unique Poisson process $N$, sampled over two disjoint $q$-dimensional tori $\mathbb{T}_1^q$ and $\mathbb{T}_2^q$, are distributed according to the same intensity with respect to $\mu$ (note when working under the null-hypothesis, this is equivalent to considering two independent Poisson processes over the same support). For this purpose, let $\{N_t : t \geq 0\}$ be a Poisson process over a $q$-dimensional torus $\mathbb{T}^q$ with control measure given by

$$\mu_t(d\theta) := R_t f(\theta)\, d\theta, \tag{1}$$

where $R_t > 0$ denotes, roughly speaking, the expected number of observations at time $t > 0$ and $f$ is a density function over $\mathbb{T}^q$ satisfying one of two possible mild regularity conditions on its spectral decomposition. In the framework of the single kernel test statistics, a natural way to compare $(H_0)$ and $(H_1)$ is to test $\|f_1 - f_2\| = 0$ versus $\|f_1 - f_2\| \neq 0$, where $\|\cdot\|$ is a given norm on $L^2(\mathbb{T}^q, d\mu)$. Therefore, as in Fromont et al. (2013), considering an increasing function of the norm of the projection onto a finite subspace of $L^2(\mathbb{T}^q, d\mu)$ can be viewed as a suitable candidate for a test statistic.

Now, consider the estimator defined by

$$\begin{aligned} U_j(t) := \sum_{k=1}^{K_j} \Bigg[ & \left( \int_{\mathbb{T}_1^q} \psi_{j,k}(\theta)\, N_t(d\theta) - \int_{\mathbb{T}_2^q} \psi_{j,k}(\theta)\, N_t(d\theta) \right)^2 \\ & - \int_{\mathbb{T}_1^q} \psi_{j,k}^2(\theta)\, N_t(d\theta) - \int_{\mathbb{T}_2^q} \psi_{j,k}^2(\theta)\, N_t(d\theta) \Bigg], \end{aligned} \tag{2}$$

where, given a scale parameter $B$, the set $\left\{\psi_{j,k} : j \geq 0,\ k = 1, \dots, K_j\right\}$ is the set of the $q$-dimensional toroidal needlets for which the index $j$ denotes the resolution level, while $K_j$ stands for the cardinality of needlets at a given $j$. Hence, each $k \in \left\{1, \dots, K_j\right\}$ denotes the location over $\mathbb{T}^q$ of the subregion (the so-called pixel) on which the corresponding needlet is non-negligible (see, for example, Narcowich et al., 2006a). Note that the estimator (2) is given in the form of a $U$-statistics, see Lemma 4.1. Heuristically, the higher $j$ is, the finer the resolution provided by the corresponding set of needlets. Therefore, the choice of $j$ establishes which and how many frequencies are involved in the construction of the finite subspace of $L^2(\mathbb{T}^q, d\mu)$ taken into account in the needlet decomposition (see also Baldi et al., 2009a). Whereas summing over the index $k$ in (2) guarantees the coverage of the whole $q$-dimensional torus, choosing a subset of $\left\{1, \dots, K_j\right\}$ establishes analogous results holding only on the corresponding subregions of $\mathbb{T}^q$. Furthermore, for $i = 1, 2$,

$$\mathbb{E}\left( \left( \int_{\mathbb{T}_i^q} \psi_{j,k}^2(\theta)\, N_t(d\theta) \right)^2 \right) = \left( R_t \int_{\mathbb{T}_i^q} \psi_{j,k} f_i(\theta)\, d\theta \right)^2 + R_t \int_{\mathbb{T}_i^q} \psi_{j,k}^2 f_i(\theta)\, d\theta.$$

As a straightforward consequence, each summand in (2) is an unbiased estimator of the projection of $f_1 - f_2$ (rescaled by $R_t^2$) onto the wavelet subspace labeled by the pair $j, k$ (see also Fromont et al., 2013).

The strategy for deriving the upcoming quantitative normal approximation results for the two-sample problem on multidimensional tori will be to make use of the celebrated Stein–Malliavin techniques obtained in Peccati et al. (2010) through the combination of Stein's method with the Malliavin calculus of variations for Poisson functionals.

Asymptotic results related to this framework are among notable innovations introduced in this paper. Analogously to Bourguin et al. (2014) for the estimation of the moments of an unknown distribution, the density to be tested is unknown in our framework, and so is the number of observations available at time $t$, which is in fact random and depends on the intensity of a Poisson process. Therefore, the asymptotic results produced here are intrinsically different from the ones making use of classical standard techniques such as, for instance, likelihood ratio tests or approximation kernels. Indeed, to establish central limit theorems and the corresponding rates of convergence requires a proper balance between the intensity $R_t$, the intrinsic properties of the chosen set of needlets at the level $j$, and the spectral parameter $\alpha$ controlling the rate of decay of the density function $f$. These bounds, through their dependence on the parameter $\alpha$ of the density function $f$ appearing in the control measure of the Poisson process, provide a new quantitative estimation of the impact of the regularity of $f$ in the quality of the normal approximation of the estimator (2). A rigorous formulation of these quantitative bounds, and of the role played by the regularity of $f$ are, up to our knowledge, the first quantitative bounds for the two-sample problem involving the regularity parameter of the density function $f$, hence providing a refined analysis of this well-known problem.

### *1.1. An overview of the literature*

The comparison between two probability distributions has been an important and long-lasting subject with applications in a wide range of fields, such as biology, medicine, physics and cosmology (among a lot of others). Following the seminal papers (Cox, 1953; Przyborowski and Wilenski, 1940), dealing with homogeneous Poisson processes, two-sample test statistics were introduced within many different settings: some kernel-based procedures were proposed in Anderson et al. (1994), Gretton et al. (2008) and Hall and Tajvidi (2002), while the study of asymptotic properties of U-statistics based on non-homogeneous Poisson processes was addressed, for instance, in Deshpande et al. (1999) and Fromont et al. (2013).

Please cite this article in press as: Bourguin S., Durastanti C., On normal approximations for the two-sample problem on multidimensional tori. J. Statist. Plann. Inference (2017), https://doi.org/10.1016/j.jspi.2017.10.009.