On testing the significance of a mode

Verifica della significatività di una moda

Federico Ferraccioli and Giovanna Menardi

Abstract We propose a nonparametric test for the significance of a mode, with the aim of evaluating whether a region of relatively high observed density reflects the actual presence of a mode in the true distribution underlying a set of data. The method leverages on Morse theory to characterize the local properties of the modes and the gradient. This allows the definition of an asymptotic test, based on the concept of gradient ascent paths and relying on resampling methods, to approximate the distribution of the test statistic under the null hypothesis. The performances of the proposed test statistic and the control of the Type-I error are shown via multiple simulation studies.

Abstract Al fine di valutare se una regione ad alta densità osservata riflette la presenza di una moda nella reale distribuzione sottostante i dati, in questo lavoro si propone un test di verifica della significatività di una moda. La procedura proposta sfrutta la teoria Morse per caratterizzare le proprietà locali delle mode e del gradiente di una funzione di densità. In questo modo, è possibile definire una procedura asintotica basata sull'ascesa del gradiente e che sfrutta una tecnica di ricampionamento per approssimare la distribuzione della statistica test sotto l'ipotesi nulla. Il comportamento del test è valutato rispetto alla probabilità di commettere un errore di I-tipo via simulazione.

Key words: bootstrap, mode, nonparametric inference, modal clustering

Federico Ferraccioli
Department of Statistical Sciences, University of Padova e-mail: federico.ferraccioli@unipd.it

Giovanna Menardi
Department of Statistical Sciences, University of Padova e-mail: menardi@stat.unipd.it

1 Introduction

Inference on the modes of a distribution has been historically overlooked with respect to other common location measures such as mean and median. In fact, especially when data exhibit non-Gaussian features as skewness or heavy tails, or some unlabeled heterogeneity occurring in the form of multimodal structures, modes represent useful tools to summarize distributions. Additionally, their understanding may represent a fundamental step to aid deciding how to subsequently approach the analysis the most fruitfully.

A first, well established, branch of literature on this topic addresses the problem of building statistical tests or confidence bounds on the true number of modes [see, for a review, **?**, and reference therein]. Alternatively, one may be interested in evaluating the position, rather than the number of modes, to understand whether the regions of relatively high observed density reflect the actual clustering of data in subpopulations; similarly, the observation of somewhat clumped data in the tails of an empirical distribution may induce to wonder if they are real or just a spurious effect of sample variability. These problems can be formalized in the - relatively neglected and fairly complicated - aim of testing the significance of a mode. The few contributions in this direction mostly rely on the study of density features like the gradient or the curvature. See **?**, **?** and more recently **?**. Consistently with this latter aim, we propose an asymptotic test to evaluate if a specific point is a true mode of the - unknown - probability density function underlying an observed set of data. The procedure borrows some tools from both the theory and the operational means addressing the modal formulation of the clustering problem and it is here applied by following a nonparametric approach. Specifically, we leverage on Morse theory to characterize the local properties of the modes, viewed as local maxima of a function, and their gradient. This formalization allows us to approximate the bootstrap distribution of a mode estimator based on the gradient ascent paths of the density, and used to define an asymptotically chi-squared test statistic.

After framing the problem in the context of Morse theory (Section 2), in the following we illustrate the test and its underlying rationale (Section 3), and show its behaviour with respect to the probability of type-I error via some simulations.

2 Modes as critical points of the density

While intuitively clear, the problem of testing mode significance is firstly definitional. The concept of mode itself is, indeed, ambiguous, as for example the Uniform distribution can be regarded to as both unimodal or without modes. To overcome this problem and formalize our framework without any elusiveness, we shall restrict the analysis to smooth distributions, and exclude non-standard ones as, for example, functions with plateaux. For our purpose, we resort to the framework provided by Morse Theory, a branch of differential topology which draws the relationship be-

tween the stationary points of a smooth real-valued functions on a manifold, and the global topology of the manifold. See **?** for an introduction.

Given a continuous random variable X, with probability density function $f : \mathbb{R}^d \to \mathbb{R}$, we then assume that f is a Morse function, i.e. a function having non-degenerate critical points. For any $\mathbf{x} \in \mathbb{R}^d$ it is possible to define the *integral curve* of the negative density gradient $-\nabla f$, as the path $\nu_{\mathbf{x}} : \mathbb{R} \mapsto \mathbb{R}^d$ such that

$$\begin{cases} \nu'_{\mathbf{x}}(t) & = -\nabla f(\nu_{\mathbf{x}}(t)) \\ \nu_{\mathbf{x}}(0) & = \mathbf{x}. \end{cases}$$

With that in mind, we identify the set of the local maxima, or modes, of f as

$$\boldsymbol{\Theta} = \{\mathbf{x} \in \mathbb{R}^d : \lim_{t\to\infty} \nu_{\mathbf{x}}(t) = \mathbf{x}\},$$

i.e. the set of points whose integral curve is degenerate at $\nu_{\mathbf{x}}(0)$.

A standard result in Morse theory is that there is a unique gradient ascent path starting at a point that eventually arrives at one of the modes (except for a set of measure 0). Hence, the set of integral curves of the negative gradient allows us to define a partition of $\mathbb{R}^d$ in "domains of attraction" of each mode, to be intended as the sets of points for which $\nu_{\mathbf{x}}(t)$ converges to that mode:

$$\mathscr{D}(\boldsymbol{\theta}) = \{\mathbf{x} \in \mathbb{R}^d : \lim_{t\to\infty} \nu_{\mathbf{x}}(t) = \boldsymbol{\theta} \in \boldsymbol{\Theta}\}.$$

The problem of finding the integral curve $\nu_{\mathbf{x}}(\cdot)$ and its limit $\lim_{t\to\infty} \nu_{\mathbf{x}}(t)$, can be approximated by the iterative scheme

$$\begin{cases} \mathbf{x}_{(0)} & = \mathbf{x}, \\ \mathbf{x}_{(s+1)} & = \mathbf{x}_{(s)} + A\frac{\nabla f(\mathbf{x}_{(s)})}{f(\mathbf{x}_{(s)})}, \end{cases} \tag{1}$$

where A is a $d \times d$ positive definite matrix chosen to guarantee the convergence. Operationally, the function f is unknown, and mode estimation is then performed by plugging in (1) a suitable estimate of both f and its gradient, built from a a sample $\mathscr{X} = (X_1, \ldots, X_n)$ of i.i.d realizations of X, so that the recurrence in (1) becomes

$$\mathbf{x}_{(s+1)} = \mathbf{x}_{(s)} + A\frac{\widehat{\nabla f}(\mathbf{x}_{(s)}; \mathscr{X})}{\hat{f}(\mathbf{x}_{(s)}; \mathscr{X})}. \tag{2}$$

The convergence properties of this gradient-ascent algorithm have been studied in **?** and **?**. A possible choice is to estimate the density and its gradient with a kernel estimator, leading to a particularly convenient iteration scheme known as the mean-shift [see **?**, Ch. 6, for a more detailed derivation]. In the next Section, we will use these properties to define a test statistic for the modes of a density function.

3 Methodology

In the lack of information about the true modal structure of f, testing the significance of a mode recasts to defining the system of hypotheses

$$H_0 : \theta_0 \in \Theta \quad \text{vs} \quad H_1 : \theta_0 \notin \Theta, \tag{3}$$

for some $\theta_0 \in \mathbb{R}^d$. While apparently composite, the null hypothesis is fact a simple one, as the - yet unknown - partition of $\mathbb{R}^d$ in the set $\{\mathscr{D}(\theta)\}_{\theta\in\Theta}$ allows us to intend H_0 as "θ_0 is the mode of the domain $\mathscr{D}(\theta)$ where it belongs".

Building on the sample $\mathscr{X}$, we first obtain an estimate $\hat{f}(\cdot;\mathscr{X})$ and $\widehat{\nabla f}(\cdot;\mathscr{X})$ of the density function and its gradient. For the subsequent developments, we consider a nonparametric kernel density estimator, which has been proven to provide consistent estimates of f under some regularity conditions on the function and the selected amount of smoothing [see, e.g. **?**]. In fact, other methods for density estimation - not necessarily nonparametric - with good general properties and producing differentiable estimates can be used.

To test (3) we then build an estimate $\hat{\theta}$ of the mode. This is obtained as the convergence point of the iteration scheme (2), with $\mathbf{x}_{(0)} = \theta_0$, and represents the mode of $\hat{f}$ associated with the domain $\mathscr{D}(\theta)$ to which θ_0 belongs. Afterwards, we obtain an approximation of the distribution of $\hat{\theta}$ under the null hypothesis. Here we propose to approximate such distribution with a resampling procedure, such as bootstrap or subsampling [**?**], together with the iteration scheme in (2). In particular, let $\mathscr{X}^*$ be a resampled version of the original data $\mathscr{X}$. With the obtained sample, using the initial condition $\mathbf{x}_{(0)} = \hat{\theta}$, we compute

$$\theta^* = \hat{\theta} + A\frac{\widehat{\nabla f}(\hat{\theta};\mathscr{X}^*)}{\hat{f}(\hat{\theta};\mathscr{X}^*)}.$$

Here we consider $A = \alpha I_d$, with $0 < \alpha < 1$, to guarantee the convergence. The underlying rationale is that, under the null hypothesis, since $\hat{\theta} \to \theta_0$ and $\widehat{\nabla f} \to \nabla f$, we expect $\widehat{\nabla f}(\hat{\theta};\mathscr{X}^*)$ to be close to zero. Hence, by iterating the process B times, we obtain a set $\{\theta_1^*,\dots,\theta_B^*\}$ of realizations from the bootstrap distribution of $\hat{\theta}$ under H_0. With that in mind, we define

$$\hat{\mu} = \frac{1}{B}\sum_{b=1}^{B}\theta_b^* \quad \text{and} \quad \hat{\Sigma} = \frac{1}{B}\sum_{b=1}^{B}(\theta_b^* - \hat{\mu})^2.$$

From the multivariate central limit theorem [**?**] it follows that under the null hypothesis $\sqrt{n}(\hat{\mu} - \theta_0) \dot{\sim} \mathscr{N}(0,\hat{\Sigma})$. We can therefore define a test statistic

$$T = (\hat{\mu} - \theta_0)^\top \hat{\Sigma}^{-1}(\hat{\mu} - \theta_0) \dot{\sim} \chi^2_d,$$

and reject H_0 for large values of T.

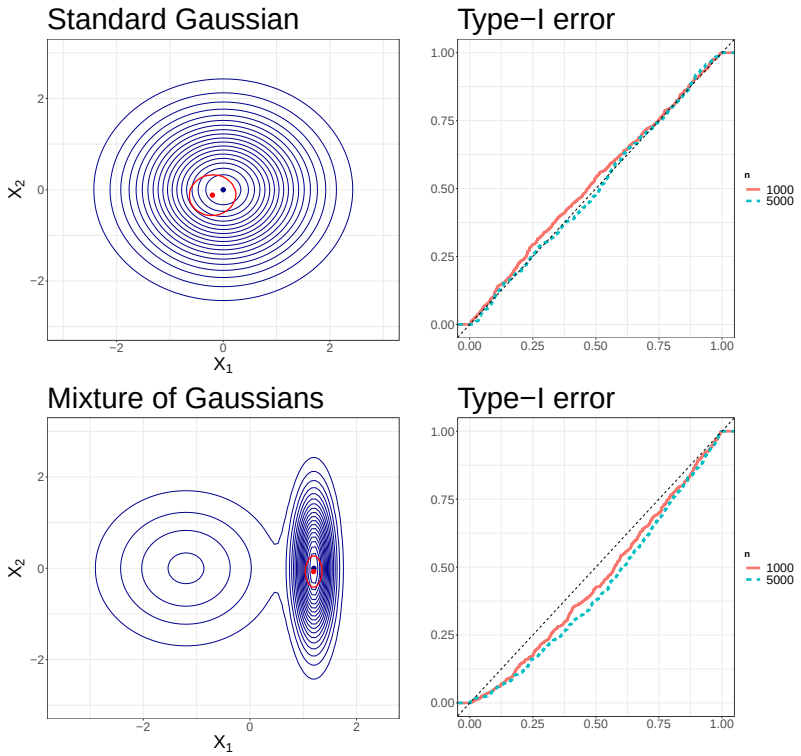

Fig. 1 In the first row, two dimensional standard Gaussian. On the second row, the Gaussian mixture. The left panels show the resampled distribution of the modes, with the black dot corresponding to θ_0. The right panels show the control of the Type-I error for two different sample sizes, $n = 1000$ and $n = 5000$.

4 Empirical study

To check the control of the Type-I error probability of the proposed test statistic, we have conducted a simulation study. For brevity, we report here the results of two bivariate settings of different complexity only, illustrated in the left panels of Figure 1, and referred to the mode of a standard Gaussian distribution and the most prominent mode of a balanced mixture of two Gaussian distributions with even variance components.

In both cases we generated 500 samples of size $n = 1000$ and $n = 5000$, and we compared the p-value curves vs increasing values of Type-I error probabilities.

In the first scenario, where the distribution is unimodal and isotropic, the test shows very good performances and the control of the Type-I error is almost perfect, even with a smaller sample size. Although this case is fairly simple, it is nonetheless informative on the behaviour of the proposed test in a benchmark setting. In the second, more complex, scenario, we focused on the right-most mode. As clear in

Figure 1, the region of interest is highly anisotropic in the vertical direction, with very steep gradients in the horizontal direction. In this case the true distribution of the mode might have a smaller variability in the horizontal direction with respect to the resampled distribution, thus leading to a more conservative test.

The proposed test shows fairly good performances and control of the Type-I error in both scenarios. Moreover, due to the small number of iterations in the gradient procedure, it is computationally efficient even in higher dimensions and with larger sample sizes. Future research will focus on a more thorough analysis on the control of the Type-I error and the power of the test in more complicated scenarios. It would also be of interest to better understand the theoretical and asymptotic properties of the proposed procedure.