Summary

Abstract

Controlled branching processes are appropriate probabilistic models for the description of population dynamics in which the number of individuals with reproductive capacity in each generation is controlled by a random control function. The probabilistic theory of these processes has been extensively developed, being an important issue to examine the inferential problems arising from them. We focus our attention on controlled branching processes with offspring distribution belonging to a general parametric family.

The purpose of this work is to consider minimum disparity estimators of the underlying offspring parameters and to study their asymptotic properties in the supercritical case, and their robustness against gross errors. The results obtained show that the mentioned procedure provides efficient estimators of the parameter of interest and also an effective treatment of anomalous data points.

Keywords: Controlled branching processes, minimum disparity estimation, robustness.

State of art

Branching processes are mathematical models for the description of the evolution of systems whose elements originate new ones according to probability laws. The simplest model arose at the end of XIX century from the study of the extinction of family lines of the European aristocracy. Since then it has been extensively studied owing to its applicability in a large variety of applied fields, for instance, growth and extinction of populations, Biology (gene amplification, clonal resistance theory of cancer cells, polymerase chain reactions, etc.), Epidemiology (the evolution of infectious diseases), Cell proliferation kinetics (stem cells, etc.), and algorithm and data structures. The complexity of some problems in such fields has required the introduction and study of new branching processes, among which the controlled branching process is.

In a controlled branching process, every individual reproduces independently of the others with the same probability law (the offspring distribution) and once the number of offspring is known, a control mechanism defines the number of progenitors that take part in the reproduction process in each generation. Thus, this process lets include several branching processes as particular cases.

From its emergence to the present, the probabilistic theory of this process has been extensively developed, especially the study of its extinction problem and its limiting behaviour. The comportment of these populations is strongly associated to the main parameters of the offspring and control distributions, as a consequence, nowadays, an important issue is to study the inferential problems arising from this model, which has become in the principal focus of research. From a frequentist outlook, results on maximum likelihood estimation, weighted conditional least squares estimation or using martingale theory have been established. In a Bayesian framework, the unique results in controlled branching processes correspond to Monte Carlo Markov Chain and Approximate Bayesian Computation methodologies in processes with deterministic control function.

In the context of branching processes, robust estimation has hardly studied. Robust inference constitutes the main goal of this work and it is developed by using the minimum disparity estimation. This approach has exclusively been studied for Hellinger distance in supercritical Bienaymé-Galton-Watson processes. The purpose of this work is to generalize these results to controlled branching processes, not only by considering the Hellinger distance but a general disparity measure.

Contributions of the work

Let us define mathematically a controlled branching process with random control function, denoted $\{Z_n\}_{n\geq 0}$:

$$Z_0 = N, \quad Z_{n+1} = \sum_{j=1}^{\phi_n(Z_n)} X_{nj}, \quad n = 0, 1, \ldots,$$

where N is a nonnegative integer, and $\{X_{nj} : \ n = 0, 1, \ldots; j = 1, 2, \ldots\}$ and $\{\phi_n(k) : n, k = 0, 1, \ldots\}$ are two independent families of nonnegative integer valued random variables. In addition, X_{nj}, $n = 0, 1, \ldots$, $j = 1, 2, \ldots$, are independent and identically distributed random variables, and for each $n = 0, 1, \ldots$, $\{\phi_n(k)\}_{k\geq 0}$, are independent stochastic processes with equal one-dimensional probability distributions. The common probability distribution of the random variables X_{nj} is denoted by $p = \{p_k\}_{k\geq 0}$, which is known as offspring distribution or reproduction law, and its mean and variance by m and σ^2 (assumed finite), respectively, and we refer to them as offspring mean and offspring variance.

Assuming that p belongs to a general parametric family $\mathcal{F}_\Theta$, that is $p = p(\theta_0)$, for $\theta_0 \in \Theta$, we study the minimum disparity estimation of the main parameters related to the offspring distribution of a controlled branching process with random control function. Given a nonparametric estimator of p, $\tilde{p}_n = \{\tilde{p}_{n,k}\}_{k\geq 0}$, the minimum disparity estimator of θ_0 for a certain disparity ρ based on $\tilde{p}_n$ is defined as:

$$\tilde{\theta}_n^\rho(\tilde{p}_n) = \arg\min_{\theta\in\Theta} \rho(\tilde{p}_n, \theta),$$

where

$$\rho(\tilde{p}_n, \theta) = \sum_{k=0}^{\infty} G\left(\frac{\tilde{p}_{n,k}}{p_k(\theta)} - 1\right) p_k(\theta),$$

and $G(\cdot)$ is a three times differentiable and strictly convex function on $[-1, \infty)$ with $G(0) = 0$ (see the extended paper for more details).

First of all, we present several interesting examples of disparities and we establish conditions for the existence and uniqueness of minimum disparity estimators (MDEs) in a general discrete model. To this end, we make use of the disparity functional T^ρ associated to ρ (see the extended paper) and we weaken the common assumption of the compactness of the parametric space in a similar way to that given for the Hellinger distance in Simpson (1987). We also show that one can obtain several MDEs of the offspring parameter by considering different disparity measures and nonparametric estimators of the offspring distribution based on several sample schemes.

The first sample that we consider is the one given by the entire family tree. For this sample, we determine the MDE of the offspring parameter defined by the nonparametric maximum likelihood estimator (MLE) of the offspring distribution, denoted by $\hat{p}_n = \{\hat{p}_{n,k}\}_{k\geq 0}$, and defined as:

$$\widehat{p}_{n,k} = \frac{Y_{n-1}(k)}{\Delta_{n-1}}, \qquad k \geq 0,$$

where $Y_l(k) = \sum_{j=0}^{l} Z_j(k)$, $Z_j(k) = \sum_{i=1}^{\phi_j(Z_j)} I_{\{X_{ji}=k\}}$, and $\Delta_l = \sum_{k=0}^{l} \phi_k(Z_k)$, $0 \leq l \leq n-1$, $k \geq 0$ (see González et al. (2015)). Using the properties of this estimator, we prove the consistency of the MDEs of the offspring parameter based on the family tree assuming certain hypotheses on the process (see the extended paper for details).

Theorem 1. *Suppose that* $\arg\min_{\theta\in\Theta} \rho(p, p(\theta))$ *is unique, where* $p = p(\theta_0)$ *is the true reproduction law. Under conditions which guarantee* $\hat{p}_{n,k}$ *is a strongly consistent estimator of* p_k*, for each* $k \geq 0$*, and the assumptions which guarantee the existence and continuity of the disparity functional, it is satisfied*

$$\hat{\theta}_n^\rho(\hat{p}_n) \to \theta_0 \quad \text{a.s. on } \{Z_n \to \infty\}.$$

As a consequence, under continuity conditions, the consistency of the estimators of the reproduction

law and offspring mean and variance defined by this MDE are obtained. The limiting normality of MDE of the offspring parameter suitably normalized is also established; to do this, we make use of the asymptotic normality of the associated MLE.

Theorem 2. *Let $p = p(\theta_0)$ be the true reproduction law. Under certain assumptions on the process, on the parametric family and on the disparity, it is verified*

$$\left(\sum_{l=0}^{n-1} \phi_l(Z_l)\right)^{1/2} (\tilde{\theta}_n^{\rho}(\hat{p}_n) - \theta_0) \to N\left(0, I(\theta_0)^{-1}\right),$$

with respect to the distribution $P[\cdot|Z_n \to \infty]$, being $I(\theta_0) = \sum_{k=0}^{\infty} \left(\frac{p_k'(\theta_0)}{p_k(\theta_0)}\right)^2 p_k(\theta_0)$.

Since these results are satisfied by a wide class of disparities to which the Hellinger distance belongs, and the controlled branching process includes the Bienaymè-Galton-Watson process as a particular case, the previous results are regarded as a generalization of those given for supercritical Bienaymè-Galton-Watson processes in Sriram & Vidyashankar (2000).

Owing to the difficulty observing the entire family tree, we also propose MDEs in two more realistic situations, one considering the sample defined by the total number of individuals and progenitors in each generation and the other one given by only generation sizes. This represents a significant leap forward the investigation of this methodology in branching processes although it entails the problem of determining a nonparametric estimator of the offspring distribution. Along the same line as when we observe the entire family tree, we opt for a nonparametric MLE of the offspring distribution. Using the Expectation-Maximization algorithm, we obtain the MLEs based on both samples, that given by the total number of individuals and progenitors in each generation and that given by only generation sizes (see González et al. (2015)). However, for these two samples, an explicit expression for the nonparametric estimator of the offspring distribution is not available, thus, the consistency of the MLEs can be only checked by an empirical way, so can the associated MDEs.

Having established the asymptotic properties of the MDEs, we also show their robustness properties under contaminated models. We examine their influence curves and α-influence curves, $\alpha \in (0,1)$, under mixture models for gross errors, which are defined as $p(\theta, \alpha, L) = (1-\alpha)p(\theta) + \alpha\delta_L$, $\alpha \in (0,1)$, $\theta \in \Theta$, $L \in \mathbb{N}_0$ and being η_L a point mass distribution at a nonnegative integer L.

Theorem 3. *Under certain conditions, it is satisfied*

(a) $\lim_{L\to\infty} T^{\rho}(p(\theta, \alpha, L)) = \theta$.

(b) $T^{\rho}(p(\theta, \alpha, L))$ *is a bounded and continuous function of* L.

(c) $\lim_{\alpha\to 0} \alpha^{-1}(T^{\rho}(p(\theta, \alpha, L)) - \theta) = p_L'(\theta)(I(\theta)p_L(\theta))^{-1}$.

We also obtain a lower bound for the asymptotic breakdown point for MDEs determined by a general strongly consistent estimator of the offspring distribution (see the extended paper). To this end, we adapt the results provided in Park & Basu (2004) for continuous models to discrete models. These robustness features make a wide class of MDEs into very suitable alternatives to previously studied estimators in the context of controlled branching processes.

Finally, to compare the MDEs based on different samples and disparities and to illustrate the methodology, we make a simulation-based study. We present two simulated examples. In the first one, we illustrate the consistency of the estimates based on the three aforementioned samples for the Hellinger distance and the negative exponential disparity and we compare these results with both nonparametric and parametric maximum likelihood estimates under a contaminated model. In the second one, we compare the accuracy of the MDEs based on the entire family tree, under an uncontaminated model and under different mixture models for gross errors, when we consider the likelihood disparity, the squared Hellinger distance and the negative exponential disparity. Both empirical studies show the accuracy of the estimates provided by the squared Hellinger distance and the negative exponential procedures, in contrast to the minimum likelihood disparity method.

References

GONZÁLEZ, M., MINUESA, C., & DEL PUERTO, I. (2015). Maximum likelihood estimation and Expectation-Maximization algorithm for controlled branching process. *Computational Statistics and Data Analysis* DOI: 10.1016/j.csda.2015.01.015.

PARK, C. & BASU, A. (2004). Minimum disparity estimation: asymptotic normality and breakdown point results. *Bulletin of Informatics and Cybernetics* **36**, 19–33.

SIMPSON, D. (1987). Minimum Hellinger distance estimation for the analysis of count data. *Journal of the American Statistical Association* **82**, 802–807.

SRIRAM, T. N. & VIDYASHANKAR, A. N. (2000). Minimum Hellinger distance estimation for supercritical Galton–Watson processes. *Statistics and Probability Letters* **50**, 331–342.