Austrian Journal of Statistics
April 2020, Volume 49, 99–105.
http://www.ajs.or.at/
doi:10.17713/ajs.v49i4.1135

A Consistent Estimator of Structural Distribution

Marijus Radavičius
Institute of Applied Mathematics, Vilnius University

Abstract

We consider sparse count data models with the sparsity rate $\tau = N/n = O(1)$ where $N = N(n)$ is the number of observations and $n \to \infty$ is the number of cells. In this case the plug-in estimator of the structural distribution of expected frequencies is inconsistent. If $\tau = O(n^{-\alpha})$ for some $\alpha > 0$, the nonparametric maximum likelihood estimator, in general, is also inconsistent. Assuming that some auxiliary information on the expected frequencies is available, we construct a consistent estimator of the structural distribution.

Keywords: structural distribution, Poisson mixture, sparsity rate, nonparametric estimator, consistency, weak convergence.

1. Introduction

Let us consider the multinomial sampling scheme

$$\mathbf{y} = (y_1, \dots, y_n), \quad \mathbf{y} \sim Multinomial_n(N, \mathbf{p}), \quad \mathbf{p} = (p_1, \dots, p_n) \in \mathcal{P}_n, \tag{1}$$

in case of sparse asymptotics (cf. Fienberg and Holland 1973; Bishop, Fienberg, and Holland 1975):

$$\mathbf{p} = \mathbf{p}(n), \quad N = N(n) \to \infty \quad \text{as} \quad n \to \infty.$$

Here $\mathcal{P}_n$ is the standard $(n-1)$-simplex of probabilities $\mathbf{p}$. Khmaladze (1988) proposed specifications of sparse asymptotics by introducing sampling schemes with *large number of rare events*. It assumes that

$$N = O(n), \quad \text{as} \quad n \to \infty.$$

In this case, a consistent estimator of probabilities $\mathbf{p}$ does not exist for any reasonable metric (Kolchin, Sevastyanov, and Chistyakov 1978; Khmaladze 1988; Klaassen and Mnatsakanov 2000; Radavičius 2019).

Sometimes it is natural to assume that problem under consideration possesses some invariance properties. For instance, palindromic Bernoulli distributions (Marchetti and Wermuth 2016) and palindromic Ising models (Marchetti and Wermuth 2017) are invariant with respect to palindromic transformations, DNA sequence symmetry models assume invariance of DNA sequence distributions with respect to various symmetric transformations including reverse, complement and their combination (Kong, Fan, Chen, Hsu, Zhou, Zheng, and Lee 2009; Radavičius, Rekašius, and Židanavičiūte 2019). Let us suppose that *cell numbering is irrelevant* for statistical inference. This means that components of the vector $\mathbf{y}$ are *exchangeable*, i.e. their distribution is invariant with respect to the coordinate permutations. In turn, this ensures that all useful information about the cell probabilities $\mathbf{p}$ is contained in their *structural distribution.*

Structural distribution Klaassen and Mnatsakanov (2000) (cf. Khmaladze and Chitshvili 1989; Khmaladze 1988) defined the (empirical) *structural distribution* G_n as the empirical distribution of the "observations" $\lambda = \lambda(n) := N \cdot \mathbf{p}$,

$$G_n := \frac{1}{n} \sum_{j=1}^{n} \delta_{\lambda_j}. \tag{2}$$

Here and in what follows δ_a denotes the Dirac measure centered at a. The basic assumption is that G_n (weakly) converges to a probability distribution G, i.e.,

$$G_n \overset{\mathcal{W}}{\to} G, \quad n \to \infty. \tag{3}$$

From the viewpoint of latent distribution modelling it is more natural to reserve the term *structural distribution* for the distribution G and to refer to G_n as the *empirical structural distribution.*

Structural distributions are widely used in *quantitative linguistics.* In this case, N is the size in words of a language corpus or some text and N is the size of its vocabulary (see, e.g. Khmaladze and Chitshvili 1989; Piaseckiene and Radavičius 2014).

Khmaladze (1988) has noticed that the natural (plug-in) estimator of G obtained by substituting y_j for λ_j ($j = 1, \ldots, n$) in (2) generally yields an inconsistent estimator (see also Klaassen and Mnatsakanov 2000; van Es, Klaassen, and Mnatsakanov 2003). Consistent estimators of structural distribution based on grouping or kernel smoothing are given in (Klaassen and Mnatsakanov 2000; van Es and Kolios 2002; van Es *et al.* 2003) under some smoothness conditions, see Example (a) below.

Poisson mixture Assuming that the probabilities $\mathbf{p}$ vanish as $n \to \infty$, a Poisson sampling scheme is commonly used as an approximation to that of multinomial (1) and can be obtained from the latter by taking the number of observations N to be a Poisson random variable. More subtle arguments based on poissonization show that, when dealing with the structural distribution problem, the multinomial scheme (1) can be replaced with that of Poisson (van Es *et al.* 2003).

Khmaladze (1988) pointed out that the *empirical* structural distribution can be treated as a latent mixing distribution in the empirical Bayes approach. Thus, the structural distribution can be interpreted as a *mixing distribution* in a Poisson mixture model. Nonparametric estimation of the mixing distribution is a well-known topic. We refer to the review (van de Geer 2003) and papers (Laird 1978; Redner and Walker 1984; Pfanzagl 1988; Lindsay 1983; Mnatsakanov and Klaassen 2003; Zhang 2008; Chen 2017) to mention few.

In this paper an extension of the results announced in (Radavičius 2019) is given. We consider a hierarchical Poisson sampling scheme with a certain sparsity rate. This enables us to cover the case of *very sparse data* where $N = o(n)$. Then the nonparametric maximum likelihood estimator, in general, is inconsistent, see Remark below. Assuming that some auxiliary information on expected frequencies is available, we construct a consistent estimator of the structural distribution.

In the next section we introduce the hierarchical Poisson sampling model and state our main result. The proofs are given in the last section.

2. Structural distribution estimator

We consider a *sparse hierarchical Poisson (independent) sampling* scheme with a sparsity rate τ:

$$[\mathbf{y}|\lambda] \sim Poisson_n(\tau\lambda), \quad \lambda \sim Q^{(n)}, \quad \lambda := (\lambda_1, \dots, \lambda_n),$$

where $\tau = \tau(n)$ is a positive convergent sequence, the components of $\mathbf{y} = (y_1, \dots, y_n)$ are mutually independent, the conditional distribution of y_j given λ is $Poisson(\tau\lambda_j)$, the components of λ are also mutually independent with $\lambda_j \sim Q_j = Q_j^{(n)}$, $j = 1, \dots, n$.

When $Q_j \equiv Q_1$ and $\tau \equiv 1$, we get a Poisson mixture model.

Similarly as in (2), define

$$G_n := \frac{1}{n}\sum_{j=1}^{n} Q_j^{(n)} \tag{4}$$

and assume (3), i.e., $G_n \overset{\mathcal{W}}{\to} G$ as $n \to \infty$. The limiting distribution G is called *structural distribution for the rate* τ. In the Poisson mixture model, $G = Q_1$.

Remark. Actually, we are interested in cases where $\tau \to 0$. Then the nonparametric maximum likelihood estimator of G, in general, is inconsistent.

Let us assume that the Poisson mixture model holds and, for some fixed integer $k > 1$,

$$\mu_k := \mathbf{E}\lambda_1^k = \int_0^\infty u^k dG(u) < \infty$$

and $\tau^k n = o(1)$. Note,

$$\mathbb{P}\{\max_{j=1,\dots,n} y_j \geq k\} \leq \sum_{j=1}^{n} \mathbb{P}\{y_j \geq k\} \leq \sum_{j=1}^{n} \frac{\tau^k \mathbb{E}\lambda_j^k}{k!} = \frac{n\tau^k\mu_k}{k!} = o(1).$$

Thus, with the probability tending to 1, the observations $\mathbf{y}$ take at most k different values and hence the cardinality of the support of the nonparametric maximum likelihood estimator $\widehat{G}_{NML}$ of G does not exceed k (Lindsay 1983, see also Zhang 2008). Consequently, $\widehat{G}_{NML}$ is inconsistent provided the support of G has more than k points.

Let $\Pi(G, F)$ denote the Levy-Prokhorov distance between distributions G and F.

Assumptions (P):

(P1) Let $\Delta(n) := \{\Delta_\ell, \ell = 0, 1, \dots, L\}$, $L = L(n)$, be a partition of $\{1, \dots, n\}$ such that $n_0 := |\Delta_0| = o(n)$ and, for some parametric family of distributions $\mathcal{F}(\Theta) := \{F_\theta, \theta \in \Theta\}$, $\Theta \subset \mathbf{R}^k$,

$$\Pi(G_{n,\ell}, F_{\theta_\ell}) \to 0, \quad \theta_\ell = \theta_\ell(n) \in \Theta,$$

as $n \to \infty$ uniformly with respect to $\ell = 1, \dots, L$. Here

$$G_{n,\ell} := \frac{1}{n_\ell}\sum_{j\in\Delta_\ell} Q_j^{(n)}, \quad n_\ell := |\Delta_\ell|, \quad \ell = 0, 1, \dots, L. \tag{5}$$

(P2) For some distribution H on Θ,

$$H_n := \sum_{\ell=0}^{L} \delta_{\theta_\ell}\frac{n_\ell}{n} \overset{\mathcal{W}}{\longrightarrow} H.$$

Here $\theta_0 \in \Theta$ can be chosen arbitrarily.

(P3) The family $F(\Theta)$ of distributions is uniformly continuous in the weak topology with respect to $\theta \in \Theta$.

(P4) There exist estimators $\hat\theta_\ell := \hat\theta(y_j, j \in \Delta_\ell)$ of θ_ℓ which are consistent *in average*, i.e., for any $\varepsilon > 0$,

$$\sum_{\ell=1}^{L} \mathbb{P}\{|\hat\theta_\ell - \theta_\ell| > \varepsilon\}\, \frac{n_\ell}{n} \to 0.$$

Theorem. *Let assumptions (P) be fulfilled. Then*

$$\widehat{G}_n := \sum_{\ell=0}^{L} F_{\hat\theta_\ell} \frac{n_\ell}{n} \tag{6}$$

is a consistent estimator of the structural distribution

$$G = \int_\Theta F_\theta \, H(\mathrm{d}\theta) \tag{7}$$

of the observations $\mathbf{y}$ *for the sparsity rate* τ*.*

Examples:

(a) *Smooth expected frequencies.* Let λ be non-random with

$$\lambda_j = g_{j,n} := g(j/n) + \varepsilon_{j,n} > 0,\ j = 1, \dots, n, \quad \max_j |\varepsilon_{j,n}| \to 0, \tag{8}$$

where $g(u), u \in [0,1]$ is a positive continuous function. The condition (8) (stated in a different form and assuming $\tau \equiv 1$) is basic in (Klaassen and Mnatsakanov 2000; van Es *et al.* 2003). In view of assumptions (P), the requirements for the function g can be relaxed to the requirement to be of finite variation.

Assumptions (P) are fulfilled for any partition $\Delta(n)$ of $\{1, \dots, n\}$ such that

$$\max_{j=1,\dots,L} \mathrm{diam}(\Delta_j) = o(n) \tag{9}$$

and

$$\tau \min_{j=1,\dots,L} n_j \to \infty.$$

When the error term $\varepsilon_{j,n}$ in (8) vanishes, one obtains a nonparametric Poisson regression model with respect to explanatory variable $x, x_j := j/n$, $j = 1, \dots, n$.

(b) *Negative binomial regression and related models.* Let $g_{j,n}$, $j = 1, \dots, n$, be given by (8) and $\lambda_j \sim \mathit{Gamma}(g_{j,n}, s)$, where $\mathit{Gamma}(a, s)$ denotes the *Gamma* distribution with the mean a and the shape parameter $s > 0$.

Assumptions (P) are fulfilled for any partition $\Delta(n)$ of $\{1, \dots, n\}$ such that (9) holds and

$$\tau^2 \min_{j=1,\dots,L} n_j \to \infty.$$

In (Radavičius and Samusenko 2012), this model was used for sparse data simulations. If the error term $\varepsilon_{j,n}$ when calculating $g_{j,n}$ in (8) is dropped, one has a nonparametric negative binomial regression model with respect to explanatory variable $x, x_j := j/n$, $j = 1, \dots, n$.

In (Piaseckiene and Radavičius 2014), zero inflated negative binomial regression model and the empirical Bayes method have been applied to estimate the structural distribution of words in Lithuanian texts.

3. Proofs

Lemma 1. *For a given collection of paired distributions* (G_ℓ, F_ℓ), $\ell = 0, 1, \dots, L$, *and probabilities* $\mathbf{q} = (q_0, q_1, \dots, q_L) \in \mathcal{P}_L$, *denote*

$$G := \sum_{\ell=0}^{L} G_\ell \, q_\ell, \quad F := \sum_{\ell=0}^{L} F_\ell \, q_\ell.$$

If $\Pi(G_\ell, F_\ell) \le \delta$, $\ell = 1, \dots, M$, *then* $\Pi(G, F) \le \delta + (1 - \delta)\, p_0$.

Proof. Proof follows directly from the definition of the Levy-Prokhorov distance. □

Lemma 2. *Suppose that conditions (P1)–(P3) hold. Then the structural distribution of* $\mathbf{y}$ *for the rate* τ *is given by (7).*

Proof. Take any $\theta_0 \in \Theta$ and set

$$F_n := \sum_{j=0}^{L} F_{\theta_j} \frac{n_j}{n}. \tag{10}$$

According to (2) and (5),

$$G_n = \sum_{j=0}^{L} G_{n,j} \frac{n_j}{n}. \tag{11}$$

Condition (P1) implies

$$\delta_n := \max_{\ell=1,\dots,L} \Pi(G_{n,\ell}, F_{\theta_\ell}) \to 0 \quad \text{as } n \to \infty. \tag{12}$$

From Lemma 1, (10), (11), (12) and condition (P1) it follows that

$$\Pi(G_n, F_n) \le \delta_n + (1 - \delta_n) \frac{n_0}{n} \to 0. \tag{13}$$

In view of condition (P3), the mapping $\theta \to F_\theta$ is bounded and continuous in weak topology. Thus, by the definition of weak convergence, assumption (P2) and (10)

$$F_n = \int_\Theta F_\theta \, dH_n \xrightarrow{\mathcal{W}} \int_\Theta F_\theta \, dH, \quad n \to \infty. \tag{14}$$

Combining (13) and (14) completes the proof. □

Lemma 3. *If assumptions (P) are valid then for the estimator* $\widehat{G}_n$ *defined by (6)*

$$\Pi(\widehat{G}_n, F_n) \to 0 \quad \textit{in probability.}$$

Proof. Fix some $\delta > 0$ and choose $\varepsilon > 0$ so that

$$\max_{|\theta - u| \le \varepsilon} \Pi(F_u, F_\theta) \le \delta. \tag{15}$$

Here we applied condition (P3). The maximum in (15) is taken over all pairs $(\theta, u) \in \Theta \times \Theta$ such that $|\theta - u| \le \varepsilon$. Let us introduce independent random variables

$$Z_i := \mathbf{1}(|\hat{\theta}_i - \theta_i| > \varepsilon), \quad i = 1, \dots, L, \tag{16}$$

(here $\mathbf{1}(\cdot)$ stands for an indicator) and denote by

$$\bar{Z}_\varepsilon := \sum_{j=1}^{L} Z_j \frac{n_j}{n} \tag{17}$$

the (weighted) proportion of cases where the parameter estimators $\hat{\theta}_\ell$ have deviations greater than ε. From (15) and Lemma 1 using (17) we obtain

$$\Pi(\widehat{G}_n, F_n) \le \delta + (1-\delta)(n_0/n + \bar{Z}_\varepsilon).$$

Since $\delta > 0$ is arbitrary and $n_0 = o(n)$ by condition (P1), it suffices to check that $\bar{Z}_\varepsilon \to 0$ in probability. Because of (17), (16) and condition (P4) we have

$$\mathbb{E}\,\bar{Z}_\varepsilon = \sum_{j=1}^{L} \mathbb{P}\{|\hat{\theta}_j - \theta_j| > \varepsilon\}\,\frac{n_j}{n} \to 0.$$

□

Proof of Theorem. Lemma 2 ensures that the structural distribution G of the observations $\mathbf{y}$ does exist and is given by (7). From Lemma 3, (14) and (7), it follows that $\widehat{G}_n$ weakly converges to G in probability. □

References

Bishop YM, Fienberg SE, Holland PW (1975). *Discrete Multivariate Analysis. Theory and Practice.* The MIT Press, Cambridge.

Chen J (2017). "Consistency of the MLE under Mixture Models." *Statistical Science*, **32**, 47–63.

Fienberg SE, Holland PW (1973). "Simultaneous Estimation of Multinomial Cell Probabilities." *Journal of the American Statistical Association*, **68**, 683–691.

Khmaladze EV (1988). "The Statistical Analysis of a Large Number of Rare Events." *Report MS-R8804*, CWI Report.

Khmaladze EV, Chitshvili RJ (1989). "The Statistical Analysis of a Large Number of Rare Events and Related Problems." *Proc. Tbilisi Mathematical Institute*, **92**, 196–245.

Klaassen CAJ, Mnatsakanov RM (2000). "Consistent Estimation of the Structural Distribution Function." *Scandinavian Journal of Statistics*, **27**(4), 733–746.

Kolchin VF, Sevastyanov B, Chistyakov V (1978). *Random Allocations.* Washington, Wiley.

Kong SG, Fan WL, Chen HD, Hsu ZT, Zhou N, Zheng B, Lee HC (2009). "Inverse Symmetry in Complete Genomes and Whole-Genome Inverse Duplication." *PLoS ONE*, **4**(11). URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0007553.

Laird NM (1978). "Nonparametric Maximum Likelihood Estimation of a Mixing Distribution." *Journal of the American Statistical Association*, **73**, 683–691.

Lindsay BG (1983). "The Geometry of Mixture Likelihoods: A General Theory." *Annals of Statistics*, **11**, 86–94.

Marchetti GM, Wermuth N (2016). "Palindromic Bernoulli Distributions." *Electronic Journal of Statistics*, **10**(2), 2435–2460.

Marchetti GM, Wermuth N (2017). "Explicit, Identical Maximum Likelihood Estimates: for Some Cyclic Gaussian and Cyclic Ising Models." *Stat*, **6**, 282–291.

Mnatsakanov RM, Klaassen CAJ (2003). "Estimation of the Mixing Distribution in the Poisson Mixture Models: Uncensored and Censored Samples." In *Proceedings of Hawaii International Conference on Statistics and Related Fields, Honolulu, Hawaii, June 4-8, 2003*, pp. 1–18. Honolulu, Hawaii. URL http://www.hicstatistics.org/2003StatsProceedings.

Pfanzagl J (1988). "Consistency of Maximum Likelihood Estimators for Certain Nonparametric Families, in Particular: Mixtures." *Journal of Statistical Planning and Inference*, **19**, 137–158.

Piaseckiene K, Radavičius M (2014). "Empirical Bayes Estimators of Structural Distribution of Words in Lithuanian Texts." *Nonlinear Analysis: Modelling and Control*, **19**(4), 611–625.

Radavičius M (2019). "Structural Distribution Estimation." In *Computer Data Analysis and Modeling: Stochastics and Data Science, Proceedings of the 12th International Conference, Minsk, September 18-221, 2019*, pp. 280–284. Publishing Center BSU, Minsk.

Radavičius M, Rekašius T, Židanavičiūte J (2019). "Local Symmetry of Non-Coding Genetic Sequences." *Informatica*, **30**(3), 553–571.

Radavičius M, Samusenko P (2012). "Goodness-of-Fit Tests for Sparse Nominal Data Based on Grouping." *Nonlinear Analysis: Modeling and Control*, **17**(4), 489–501.

Redner RA, Walker HF (1984). "Mixture Densities, Maximum Likelihood and the EM Algorithm." *SIAM Review*, **26**, 195–239.

van de Geer S (2003). "Asymptotic Theory for Maximum Likelihood in Nonparametric Mixture Models." *Computational Statistics and Data Analysis*, **41**, 453–464.

van Es B, Klaassen CAJ, Mnatsakanov RM (2003). "Estimating the Structural Distribution Function of Cell Probabilities." *Austrian Journal of Statistics*, **32**, 85–98.

van Es B, Kolios S (2002). "Estimating a Structural Distribution Function by Grouping." *Report PR/0203080*, Mathematics ArXiv.

Zhang Z (2008). "Estimation in Mixture Models." *Phd thesis*, School of Mathematics and Statistics, Carleton University, Ottawa, Ontario, Canada.

Affiliation:

Marijus Radavičius
Institute of Applied Mathematics
Vilnius University
LT-03225, Vilnius, Lithuania
E-mail: marijus.radavicius@mii.vu.lt
URL: https://www.mif.vu.lt/eka/katedra/radavicius.php

Austrian Journal of Statistics http://www.ajs.or.at/
published by the Austrian Society of Statistics http://www.osg.or.at/
Volume 49 *Submitted:* 2019-12-15
April 2020 *Accepted:* 2020-03-03