STRUCTURAL DISTRIBUTION ESTIMATION

M. Radavičius
Institute of Data Science and Digital Technologies, Vilnius University
Vilnius, LITHUANIA
e-mail: marijus.radavicius@mii.vu.lt

Abstract

We consider count data models in case of sparse asymptotics. Then a consistent estimator of expected frequencies does not exist for any reasonable metric. Moreover, a plug-in estimator of a structural distribution is also inconsistent. Assuming that some auxiliary information on expected frequencies is available, we construct a consistent estimator of the structural distribution.
Keywords: data science, count data, sparse asymptotics, structural distribution

1 Introduction

Let us consider multinomial sampling scheme

$$Y = (y_1, \dots, y_n), \quad Y \sim Multinomial_n(N, P), \quad P = (p_1, \dots, p_n) \in \mathcal{P}_n,$$

in case of sparse asymptotics: $n \to \infty$ and $P = P(n)$, $N = N(n) \to \infty$. Here $\mathcal{P}_n$ is the unit $(n-1)$-simplex of probabilities P.

Define *occupation statistics*:

$$V_m = V_m(n) := \sum_{j=1}^{n} I\{y_j = m\}, \quad m = 0, 1, \dots .$$

Here and in the sequel $I\{\cdot\}$ denotes an indicator function.

The statistic V_0 ($V^+ = V^+(n) := n - V_0$) is the number of *empty* (respectively, *nonempty*) *boxes*. In linguistics, V^+ (V_0) is the size of a vocabulary or the number of observed (respectively, unseen) word tokens.

Khmaladze (1988) [3] proposed specifications of sparse asymptotics by introducing sampling schemes with *large number of rare events* (LNRE). They are based on the following assumptions:

$$\lim_{n\to\infty} \frac{V_1(n)}{N(n)} > 0, \tag{1}$$

and

$$V^+(n) \to \infty, \quad \lim_{n\to\infty} \frac{V_1(n)}{V^+(n)} > 0. \tag{2}$$

Definition. ([3]) A multinomial sampling scheme with *large number of rare events* is said to be in zone (d1) (in zone (d2)) iff condition (1) (respectively, (2)) is satisfied.

Note that (1) implies (2).

In the LNRE model, a consistent estimator of probabilities P does not exist for any reasonable metric [3, 5, 2]. Sometimes much less informative characteristics of a model are sufficient for inference. For instance, if the cell numbering is irrelevant for statistical inference, all useful information about the cell probabilities P is contained in their *structural distribution*. Structural distributions are widely used in quantitative linguistics.

Klaassen and Mnatsakanov (2000) [5] (cf. Khmaladze & Chitashvili (1989) [4] and Khmaladze (1988) [3]) defined the (empirical) *structural distribution* G_n as the empirical distribution of the "observations" $N \cdot P$,

$$G_n := \frac{1}{n} \sum_{j=1}^{n} \delta_{Np_j}. \tag{3}$$

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.,

$$G_n \overset{\mathcal{W}}{\to} G, \quad n \to \infty. \tag{4}$$

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*.

Khmaladze (1988) [3] has noticed that a natural (plug-in) estimator of G obtained by substituting y_j for Np_j $(j = 1, \ldots, n)$ in (3) generally yields an inconsistent estimator. Consistent estimators of structural distribution based on grouping or kernel smoothing are provided by Klaassen & Mnatsakanov (2000) [5], van Es & Kolios (2003) [2] and van Es et al. (2003) [1] under some smoothness conditions, see assumption (U) below.

Assumption (U) ([5, 1]). The sequence of distribution densities

$$f_n(u) := \sum_{j=1}^{n} np_j \, I\left\{\frac{j-1}{n} < u \leq \frac{j}{n}\right\}, \quad u \in (0,1],$$

uniformly converges to a continuous distribution density f.

Assumption (U) implies an approximate *latent distribution model* with a latent variable $Z \sim f$:

$$p_j = \int_{(j-1)/n}^{j/n} f(u)du + \frac{\epsilon_j}{n}, \; j = 1, \ldots, n, \quad \max_j |\epsilon_j| \to 0.$$

In this study, we deal with a *Poisson sampling* scheme and construct a consistent estimator of structural distribution of expected cell frequencies.

2 Consistent estimator of structural distribution

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

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

where $\tau = \tau(n)$ is a positive convergent sequence, the components of $Y = (y_1, \dots, y_n)$ 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 $\lambda_j \sim Q_j = Q_j^{(n)},\ j = 1, \dots, n$, and $\lambda_+ = n$, $\lambda_+ := \sum_{j=1}^n \lambda_j$.

Actually, we are interested in cases where $\tau \to 0$.

The Poisson sampling scheme is used as an approximation to that of multinomial under the LNRE condition and can be obtained from the latter via Poissonization [2, 1]. When $Q_j \equiv Q_1$ and $\tau \equiv 1$, we get a Poisson mixture model considered in [6].

Similarly as in (3), define

$$G_n := \frac{1}{n}\sum_{j=1}^n Q_j^{(n)} \tag{5}$$

and assume (4), 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, $G = Q_1$.

Assumptions (P):

(P1) Let $\{\Delta_\ell, \ell = 1, \dots, L\}$ be a partition of $\{1, \dots, n\}$ such that $n_\ell := |\Delta_\ell| \geq n_{min}$ where $\tau\, n_{min} \to \infty$, and, for some parametric family of distributions $F(\Theta) := \{F_\theta,\ \theta \in \Theta\}$, $\Theta \subset {}^k$,

$$\frac{1}{n_\ell}\sum_{j\in\Delta_\ell} Q_j^{(n)} \overset{\mathcal{W}}{\to} F_{\theta_\ell}, \quad \theta_\ell \in \Theta,$$

as $n \to \infty$ uniformly with respect to $\ell = 1, \dots, L$. Moreover, for some distribution H on Θ,

$$\frac{1}{n}\sum_{\ell=1}^L n_\ell \delta_{\theta_\ell} \overset{\mathcal{W}}{\to} H.$$

(P2) Distributions of the family $F(\Theta)$ are uniformly continuous in weak topology with respect to $\theta \in \Theta$.

(P3) There exist estimators $\hat\theta_\ell := \hat\theta(y(j), j \in \Delta_\ell)$ of θ_ℓ which are consistent uniformly over $\{\ell = 1, \dots, L\}$, i.e., for each $\varepsilon > 0$,

$$\{\max_{\ell=1,\dots,L} |\hat\theta_\ell - \theta_\ell| > \varepsilon\} \to 0.$$

Proposition. Let assumptions (P) be satisfied. Then

$$\widehat{G} := \sum_{\ell=1}^L F_{\hat\theta_\ell}\frac{n_\ell}{n}$$

is a consistent estimator of the structural distribution (for the sparsity rate τ)

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

Examples:
(a) *Latent distribution model* (cf. assumption (U)):

$$\lambda_j = n \int_{(j-1)/n}^{j/n} f(u)du, \quad Q_j^{(n)} = \delta_{\lambda_j}, \quad j = 1, \dots, n,$$

f is a continuous probability density on $[0, 1]$.

(b) *Poisson regression and related models.* When $\lambda_j = \mu(j/n)$, $j = 1, \dots, n$, where $\mu(u), u \in [0, 1]$, is a nonnegative continuous function that integrate to 1, we have a nonparametric Poisson regression model with the explanatory variable $x, x_j := j/n$, $j = 1, \dots, n$. For a negative binomial regression model, one can take $Q_j \sim Gamma(\mu(j/n), \nu)$, where $Gamma(a, \nu)$ denotes *Gamma* distribution with the mean a and the shape parameter ν, and $\mu(u), u \in [0, 1]$, is the same as above (cf. [8]).

In [7], zero inflated negative binomial regression model and the empirical Bayes method have been applied to estimate the structural distribution of words in Lithuanian texts.

References

[1] van Es B., Klaassen C.A.J., Mnatsakanov R.M. (2003). Estimating the structural distribution function of cell probabilities. *Austrian Journal of Statistics*, **32**, pp. 85-98.

[2] van Es B., Kolios S. (2002). Estimating a structural distribution function by grouping. *Mathematics ArXiv* PR/0203080.

[3] Khmaladze E.V. (1988). The statistical analysis of a large number of rare events. *CWI Report MS-R8804*.

[4] Khmaladze E.V., Chitshvili R.J. (1989). The statistical analysis of a large number of rare events the related problem. *Proc. Tbilisi Mathematical Institute* **92**, pp. 196-245.

[5] Klaassen C.A.J., Mnatsakanov R.M. (2000). Consistent estimation of the structural distribution function, *Scand. J. Statist.*, **27** (4), pp. 733-746.

[6] Mnatsakanov R.M., Klaassen C.A.J. (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.
Available at http://www.hicstatistics.org/2003StatsProceedings.

[7] Piaseckienė K., Radavičius M. (2014). Empirical Bayes estimators of structural distribution of words in Lithuanian texts. *Nonlinear Analysis: Modelling and Control*, **19** (4), pp. 611-625.

[8] Radavičius M., Samusenko P. (2012). Nonparametric testing for sparse nominal data. *Nonlinear analysis: modeling and control.* **17** (4), pp. 489-501.