

Approximations to the Likelihood in Composite Samples with Fixed and Random Weights⁽¹⁾

Approssimazioni alla Verosimiglianza di Campioni Compositi con Pesi Fissi e Casuali

Marcello Chiodi

Dipartimento di Scienze Statistiche e Matematiche Silvio Vianelli

Università degli studi di Palermo

Viale delle Scienze, 90128- Palermo-Italy

chiodi@unipa.it; <http://dssm.unipa.it/chiodi>

Riassunto: Il campionamento composito è utile specialmente negli studi ambientali. Quando si hanno a disposizione campioni compositi, specialmente con pesi aleatori, la costruzione della verosimiglianza esatta può risultare impossibile. Sfruttando il fatto che le osservazioni sono delle medie ponderate di variabili i.i.d., abbiamo usato le tecniche basate sui punti di sella per approssimare le densità di combinazioni lineari di variabili gamma con pesi fissi ed anche il caso più interessante di medie di variabili gamma pesate con le componenti di una variabile con distribuzione di Dirichlet. Data l'elevata accuratezza delle approssimazioni di punto di sella, si ottengono buone stime della verosimiglianza, la cui bontà è stata verificata con simulazioni.

Questa approssimazione alla verosimiglianza ha permesso il calcolo degli stimatori di massima verosimiglianza per il caso con pesi fissi e con pesi aleatori.

Keywords: composite sampling, random weights, saddlepoint approximations, weighted averages.

1. Introduction

Composite sampling is a kind of sampling used in environmental studies, because it can decrease substantially the cost of the survey, without lack of sample information. Suppose the observable variable Y is obtained after a composition of the kind:

$$Y = \sum_{j=1}^k w_j X_j, \quad \text{with} \quad \sum_{j=1}^k w_j = 1, \quad (1)$$

where w_j are known and fixed coefficients, otherwise they are the determinations of a random variable with k components, with density $f_{\mathbf{w}}(\mathbf{w}; \boldsymbol{\theta})$, with $\boldsymbol{\theta}$ possibly unknown; the X_j are *ki.i.d.* random variables with density $f_X(x; \boldsymbol{\beta})$, whose unknown parameters $\boldsymbol{\beta}$ are the main goals of the inference.

The utility of composite sampling is that the process of sample unit selection (the X_j) is distinct from the process of physical measurements (the Y) (Boswell *et al.*, 1996). The number of units can be increased (and then sample information) without increasing the number of measurements: this reduces the costs of the analysis. This is particularly useful in environmental studies, where the cost of measurements is greater than the cost of sampling.

¹Work supported with MURST funds 60% 2000 and MURST PRIN funds 2002

To make parametric inference on the parameters β and θ , we need the likelihood of a sample of n independent observations y_i , drawn from the model (1):

$$L(\beta, \theta) \propto \prod_{i=1}^n f_Y(y_i; \beta, \theta)$$

. There is not a general exact solution, especially for the random weights situation, so that some kind of *approximation to* $f_Y(y_i; \beta, \theta)$ is needed. An exact solution, computationally hard, has been proposed by Di Salvo and Lovison (2000), for the fixed weights situation, based also on the results of Sim (1992) and Mathai (1982).

In this paper we will deal only with *i.i.d.* r.v. X_j following a Gamma distribution with two parameters $\gamma(\alpha, \lambda)$, using a saddlepoint approximations; the random weight situation is sketched in the fourth section, when weights follow a Dirichlet distribution. A comparison is made between simulated and approximated densities, to check the approximations; these approximations are then used to find the M.L.E. of the parameters. Only few results of the simulation are presented in this paper.

2. Saddlepoint approximations

Let X a random variable with moment generating function: $M_X(t) = E_X[\exp(tX)]$, and cumulant generating function (c.g.f.): $K_X(t) = \log M_X(t)$. We have a saddlepoint approximation $\tilde{f}(x_m)$ to the density of the distribution of the sum of m random variables *i.i.d.* X with the classical formula (Reid, 1991), (Daniels, 1954):

$$\tilde{f}(x_m) = \frac{\exp(m K_X(t_0) - x_m t_0)}{\sqrt{2\pi m K_X''(t_0)}}, \quad (2)$$

where t_0 is the *saddlepoint*, real solution of the equation: $m K_X'(t_0) = x_m$. The formula gives generally good local approximations to the desired density. Saddlepoint approximations are exact (or proportional to the exact density) only if X is *normal*, *gamma* or *inverse-gaussian* (Daniels, 1954).

3. Composition of gamma variables with k fixed weights

If each $Y_i = \sum_{j=1}^k w_j X_{ij}$ in a composite sample is a linear combination of *i.i.d.* gamma variables $X_{ij} \sim \gamma(\alpha, \lambda)$, the cumulant generating function (c.g.f.) of the generic r.v. Y is:

$$K_Y(t) = -\alpha \sum_{j=1}^k \log\left(1 - \frac{t w_j}{\lambda}\right) \quad \text{with} \quad t < \frac{\lambda}{\max_j w_j} \quad (3)$$

To approximate $f_Y(y_i; \beta, \theta)$, we use a saddlepoint technique, in a fashion which resembles that used by Daniels (1982) in a birth process problem.

Using the relationship (2) for any fixed value y , with the c.g.f. (3), the saddlepoint t_0 solves: $K_Y(t_0) = y$, and we obtain the approximation $\tilde{f}(y)$, using (2).

In order to check the results, we performed a preliminary simulation study to compare the saddlepoint approximations with the simulated densities with 100,000 replicates, for different values of the weights w , of k and α . The results are not reported here for brevity, but approximations revealed quite good.

4. Composite samples with k random weights from a Dirichlet distribution

Suppose that the vector of compositing weights $\mathbf{W}$ follows a Dirichlet distribution (independent from $\mathbf{X}$) (Di Salvo and Lovison, 2000), with $k - 1$ components and a common parameter δ , so that single components W_j have the same Beta marginal distribution, even if of course they are not independent. So the density of the weights is:

$$f_{\mathbf{W}}(\mathbf{w}; \delta) = \left[\left(\prod_{j=1}^s w_j \right) \left(1 - \sum_{j=1}^s w_j \right) \right]^{\delta-1} \frac{\Gamma((s+1)\delta)}{[\Gamma(\delta)]^{s+1}}$$

We can arrange the distribution of Y , as a conditioned sum. In fact, from elementary properties of the Dirichlet distribution, if Z_j are k i.i.d. r.v. with a Beta distribution of parameters: $(\delta, 1)$, then:

$$Y = \sum_{j=1}^k X_j W_j \sim \sum_{j=1}^k X_j Z_j \left| \left\{ \sum_{j=1}^k Z_j = 1 \right\} \sim \sum_{j=1}^k V_j \left| \left\{ \sum_{j=1}^k Z_j = 1 \right\} \right.$$

where we put $V_i = X_i Z_i$. In this way with we do not deal with non independent r.v. W_j , but with a conditioned distribution, suitable for a saddlepoint approximation: we approximated the conjoint bivariate density of $(\sum_{j=1}^k V_j, \sum_{j=1}^k Z_j)$ and then we easily computed an approximation of the distribution of $\sum_{j=1}^k V_j$ conditioned to $Z_j = 1$. We obtained again good approximation to the density of Y with such a saddlepoint technique. For brevity we omit here further details. In this paper we deal only with the case of uniform weights ($\delta = 1$).

In order to check the results also in this situation, we compared the saddlepoint approximations with the simulated densities, for different values of k and α . The results are not reported here for brevity, but approximations revealed quite good, apart for big values of k ($k > 5$) with small values of α ($\alpha < 0.5$).

We used this approximation to compute the log-likelihood in samples of n independent observations Y_i with random weights, to obtain the maximum likelihood estimates of α and λ , using the moment based estimates as a starting point.

5. Maximum likelihood estimation

We implemented in R the above mentioned algorithms, to obtain the maximum likelihood estimates $\{\hat{\alpha}, \hat{\lambda}\}$ of $\{\alpha, \lambda\}$, from a sample of n independent observations, with fixed and random uniform weights.

Given a sample of n independent observations $\mathbf{y} = \{y_1, y_2, \dots, y_i, \dots, y_n\}$, from a composite sampling of k components, for fixed $\{\alpha, \lambda\}$ and y_i , we can compute the saddlepoint approximation $\tilde{f}(y_i; \alpha, \lambda)$ and then the whole log-likelihood of the sample $l(\alpha, \lambda; \mathbf{y})$ with respect to $\{\alpha, \lambda\}$: $l(\alpha, \lambda; \mathbf{y}) = \sum_{i=1}^n \log \tilde{f}(y_i; \alpha, \lambda)$.

Indeed, we used the parametrization (μ, λ) , with $\mu = \frac{\alpha}{\lambda} = E[Y]$, and found the M.L.E. $\{\hat{\alpha}, \hat{\lambda}\}$ with a conjugate gradient method. Then we performed a simulation study to appreciate the properties of $\{\hat{\alpha}, \hat{\lambda}\}$, with different values of the parameters: $n = 10, 20, 50$, $\alpha = 1, 2, 4, 10$ and $\lambda = 1$. For brevity here we report only the results for $k = 3$ with

uniform random weights: for each of the 12 combination of values n, α , we simulated the drawing of 100 independent samples. On each sample the M.L. estimates $\{\hat{\mu}, \hat{\lambda}\}$ and the moments based estimates $\{\tilde{\mu}, \tilde{\lambda}\}$ have been computed.

In table 1 we report, for each pair of values of n, α , we report the values of simulated bias and the mean squared errors of $\{\hat{\mu}, \hat{\lambda}\}$ and $\{\tilde{\mu}, \tilde{\lambda}\}$ computed on 100 simulated samples. Results for the M.L.E. seem quite good, especially for $\hat{\lambda}$, and particularly for small values of α , judged from simulated sample variances and mean squared errors.

Table 1: Summary of the simulations on 100 samples for $n = 10, 20, 50, \alpha = 1, 2, 4, 10, \lambda = 1$ and $k = 3$ and uniform random weights: Bias and Mean Squared Errors of $\{\hat{\mu}, \hat{\lambda}\}$ and $\{\tilde{\mu}, \tilde{\lambda}\}$

			Bias= $M(\hat{\theta}) - \theta$				Mean Squared Errors			
	α	n	$\hat{\mu}$	$\hat{\lambda}$	$\tilde{\mu}$	$\tilde{\lambda}$	$\hat{\mu}$	$\hat{\lambda}$	$\tilde{\mu}$	$\tilde{\lambda}$
$k = 3$	1	10	0.038	0.385	-0.005	0.732	0.071	1.309	0.062	1.947
	1	20	0.046	0.128	0.010	0.303	0.028	0.273	0.024	0.460
	1	50	0.042	0.003	0.004	0.189	0.013	0.076	0.010	0.158
	2	10	0.038	0.371	0.011	0.505	0.072	1.097	0.071	1.381
	2	20	0.040	0.196	0.013	0.275	0.034	0.318	0.032	0.403
	2	50	0.004	0.046	-0.023	0.088	0.015	0.055	0.015	0.072
	4	10	0.022	0.365	-0.002	0.394	0.188	0.897	0.181	0.938
	4	20	-0.019	0.208	-0.043	0.240	0.079	0.298	0.081	0.317
	4	50	0.031	0.082	0.006	0.106	0.045	0.075	0.044	0.090
	10	10	0.051	0.440	0.027	0.464	0.405	1.008	0.401	1.048
	10	20	0.003	0.241	-0.019	0.246	0.236	0.321	0.240	0.316
	10	50	0.027	0.066	0.007	0.056	0.081	0.048	0.079	0.049

References

- Boswell M., Gore S., Lovison G. and Patil G. (1996) Annotated bibliography of composite sampling. part a 1936-92, *Environmental and Ecological Statistics*, 3, 1–50.
- Daniels H. (1954) Saddlepoint approximations in statistics, *The Annals of Mathematical Statistics*, 25, 631–650.
- Daniels H. (1982) The saddlepoint approximation for a general birth process, *Journal of Applied Probabilities*, 19, 20–28.
- Di Salvo F. and Lovison G. (2000) Parametric inference on composite samples with random weights, *Preprint*.
- Mathai A. (1982) Storage capacity of a dam with gamma type inputs, *Ann. Inst. Statist. Math.*, 34, A, 591–597.
- Reid N. (1991) Approximations and asymptotics, in: *Statistical theory and modelling. In honour of Sir David Cox, FRS*, 287–305.
- Sim C.H. (1992) Point processes with correlated gamma interarrival times, *Statistics and Probability Letters*, 15, 135–141.