PERGAMON

Applied Mathematics Letters 12 (1999) 31–36

Applied Mathematics Letters

Testing Stationary Distributions of Markov Chains Based on Rao's Divergence

M. C. PARDO
University School of Statistics
Complutense University of Madrid, 28040-Madrid, Spain

(Received and accepted January 1998)

Communicated by R. Ahlswede

Abstract—Statistical inference problems such as the estimation of parameters and testing composite hypothesis about stationary distributions in the set of states of Markov chains are solved. Both, the estimator and the statistic proposed are based on Rao's divergence. The asymptotic properties of the estimator and the critical values of asymptotically γ-level tests are obtained.

Keywords—Markov chain, Rao's divergence, Minimum R_ϕ-divergence estimate, Goodness-of-fit tests, R_ϕ-statistic.

1. INTRODUCTION

We consider a stationary irreducible aperiodic Markov chain $\mathbf{X}= (X_0, X_1, \dots)$ with a state space $\{1, \dots, m\}$. By $P = (p_{ij})_{i,j=1}^m$, we denote the matrix of transition probabilities of this chain and by $p = (p_1, \dots, p_m)^t$ the stationary distribution, i.e., solution of the equation $p = P^t p$. We assume that P is from the class $\mathbf{P}$ of irreducible aperiodic stochastic matrices with one ergodic class. The irreducibility means that there are no transient states, i.e., no $1 \leq i \leq m$ with $p_i = 0$. Therefore, p belongs to the set $\Delta_m = \{(p_1, \dots, p_m)^t \mid \sum_{i=1}^m p_i = 1, p_i \geq 0, i = 1, \dots, m\}$.

To solve the problem of testing the hypothesis $H_0 : p = \pi_0 \in T$ where $T \subset \Delta_m$ is simply an one point set, Tavaré and Altham [1] found the asymptotic distribution of the Pearson's chi-square statistic for dependent data. This statistic measures the discrepancy between the observed proportions and the hypothesized proportions. If the discrepancy is "too large", the null hypothesis is rejected. The key is the choice of a good test statistic to measure the discrepancy between the observed proportions and the hypothesized proportions. Every divergence measures this discrepancy. In fact, Menéndez *et al.* [2] proposed a family of statistics based on Csiszár divergence [3] to solve this problem and Pardo [4] a family based on Burbea and Rao's divergence [5]. This last family is defined as

$$R_\phi(\hat{p}_n, \pi_0) = \sum_{i=1}^m \phi\left(\frac{\hat{p}_{ni} + \pi_{0i}}{2}\right) - \frac{1}{2}\left\{\sum_{i=1}^m \phi(\hat{p}_{ni}) + \sum_{i=1}^m \phi(\pi_{0i})\right\}, \tag{1}$$

This work was supported by Grants DGICYT PB96-0635 and PR156/97-7159.

0893-9659/98/$—see front matter
PII: S0893-9659(98)00122-0

where $\phi : (0, \infty) \to R$ is a continuous concave function, $\phi(0) = \lim_{t \downarrow 0} \phi(t) \in (-\infty, \infty]$ and

$$\widehat{p}_n = \left(\frac{1}{n} \sum_{k=1}^{n} \mathbf{I}_{(1)}(X_k), \ldots, \frac{1}{n} \sum_{k=1}^{n} \mathbf{I}_{(m)}(X_k) \right)^t$$

is the observed cell frequencies in the data $\mathbf{X}_n = (X_1, \ldots, X_n)$ being $\mathbf{I}$ the indicator function.

This family has also been used by Pardo [6,7] for testing goodness-of-fit for independent observations under classical (fixed-cells) and sparseness assumptions, respectively. Some properties of this family of divergences can be seen in Burbea and Rao [5] and Pardo and Vajda [8].

Otherwise, the null hypothesis is called composite and specifies π to be a function of some fewer number of unknown parameters (i.e., π lies in the subset T of Δ_m) which need to be estimated from the experimental data $\mathbf{X}_n$. Menéndez *et al.* [9] study a family based on Csiszár's divergence under composite hypothesis.

In Section 3, we consider the composite hypothesis $H_0 : p = \pi_0$, where $\pi_0 = q(\theta) = (q_1(\theta), \ldots, q_m(\theta))^t \in T \subset \Delta_m$ being $\theta = (\theta_1, \ldots, \theta_s)^t \in \Theta \subseteq R^s$ the unspecified parameters vector. For every parameter $\theta \in \Theta$, we denote by $\mathbf{P}_\theta$ the sets of all matrices $P \in \mathbf{P}$ such that their stationary distribution p coincides with $q(\theta)$. This goodness-of-fit test requires us to estimate the unspecified parameters, i.e., to choose one value $q(\hat{\theta}) \in T$ that is "most consistent" with the data $\mathbf{X}_n = (X_1, \ldots, X_n)$ about the states. This last problem is solved in Section 2.

2. THE MINIMUM R_ϕ-DIVERGENCE ESTIMATOR

In this section, we study the estimation problem. Throughout, we assume that the true chain parameter is $\theta^0 \in \Theta$. This means that the true chain distribution is specified by the initial distribution $q(\theta^0)$ and by a transition matrix $P(\theta^0) \in \mathbf{P}_{\theta^0}$. The most well-known method to choose $q(\hat{\theta})$ consists of estimating θ by maximum likelihood, but another sensible way to estimate π_0 is to choose the $q(\hat{\theta}) \in T$ that it is closed to $\widehat{p}_n$ with respect to the measure $R_\phi(\widehat{p}_n, q(\theta))$. This leads to the minimum R_ϕ-divergence estimate defined as a $\widehat{\theta}_\phi \in \overline{\Theta}$ that verifies

$$R_\phi \left(\widehat{p}_n, q\left(\widehat{\theta}_\phi \right) \right) = \inf_{\theta \in \Theta} R_\phi \left(\widehat{p}_n, q(\theta) \right).$$

Let us introduce the following regularity conditions before studying the asymptotic properties of this estimator:

(A1) $q : \Theta \to \Delta_m$ is continuously differentiable in a neighborhood of θ^0 and

$$q(\theta) - q\left(\theta^0\right) = J_0 \left(\theta - \theta^0 \right) + o\left(\| \theta - \theta^0 \| \right), \qquad \text{for } \theta \to \theta^0,$$

where $J_0 = J(\theta^0) = (J_{jr}(\theta^0))$ is the Jacobian matrix being

$$J_{jr}(\theta) = \frac{\partial q_j(\theta)}{\partial \theta_r};$$

(A2) $A_0^t A_0$ is positive definite for

$$A_0 = \operatorname{diag}\left(\sqrt{-\phi''\left(q_1\left(\theta^0\right)\right)}, \ldots, \sqrt{-\phi''\left(q_m\left(\theta^0\right)\right)} \right) J_0.$$

Hereafter, we consider the matrix

$$B_0 = \operatorname{diag}\left(\sqrt{-\phi''\left(q\left(\theta^0\right)\right)} \right) \Omega_0 \operatorname{diag}\left(\sqrt{-\phi''\left(q\left(\theta^0\right)\right)} \right),$$

where $\Omega_0 = D_0C_0 + C_0^tD_0 - D_0 - q(\theta^0)q(\theta^0)^t$, being $D_0 = \text{diag}(q(\theta^0))$ and $C_0 = (\text{diag}(\mathbf{1}) - P(\theta^0) + \mathbf{1}q(\theta^0)^t)^{-1}$, with $\mathbf{1}$ the column vector of m units, is the asymptotic covariance matrix of the asymptotically normal zero mean random vector

$$\sqrt{n}\left(\widehat{p}_{n1} - q_1\left(\theta^0\right), \ldots, \widehat{p}_{nm} - q_m\left(\theta^0\right)\right)$$

(c.f. [10] or [1, equation (2.2)]). Put for brevity

$$\Delta_0 = A_0\left(A_0^tA_0\right)^{-1}, \qquad \Sigma_0 = \Delta_0A_0^t.$$

The following theorem summarizes the properties of minimum R_ϕ-divergence estimators of parameters of stationary distributions of Markov chains. Other similar results for maximum likelihood and other estimators with independent observations can be seen in [11–15].

THEOREM 1. *Let $\phi : (0, \infty) \to R$ be a twice continuously differentiable concave function. Under the above regularity conditions and assuming that the function $q : \Theta \to \Delta_m$ has continuous second partial derivatives in a neighborhood of θ^0, we have that*

$$\hat{\theta}_\phi = \theta^0 + \Delta_0^t \,\text{diag}\left(\sqrt{-\phi''\left(q\left(\theta^0\right)\right)}\right)\left(\widehat{p}_n - q\left(\theta^0\right)\right) + o\left(\left\|\widehat{p}_n - q\left(\theta^0\right)\right\|\right),$$

where $\hat{\theta}_\phi$ is unique in a neighborhood of θ^0.

PROOF. From the proof of Theorem 1 of [15], there exits a m-dimensional neighborhood U_0 of $q(\theta^0)$ in R^m and a unique, continuously differentiable function $\tilde{\theta} : U_0 \to R^s$ such that

$$\frac{\partial R_\phi\left(P, \tilde{\theta}(P)\right)}{\partial \theta_j} = 0, \qquad j = 1, \ldots, s$$

and

$$\tilde{\theta}(P) = \theta^0 + \left(A\left(\theta^0\right)^t A\left(\theta^0\right)\right)^{-1} A\left(\theta^0\right)^t \,\text{diag}\left(\sqrt{-\phi''\left(q\left(\theta^0\right)\right)}\right)\left(P - q\left(\theta^0\right)\right) + o\left(\left\|P - q\left(\theta^0\right)\right\|\right),$$

for all $P \in U_0$. Now then by the strong law of large numbers holding for the chains under consideration (cf. [10]) $\widehat{p} \xrightarrow[n\to\infty]{\text{c.s.}} q(\theta^0)$, so $\widehat{p}_n \in U_0$, and consequently, $\tilde{\theta}(\widehat{p}_n)$ is the minimum R_ϕ-divergence estimator, $\hat{\theta}_\phi$, that satisfies the following:

$$\hat{\theta}_\phi(\widehat{p}_n) = \theta^0 + \left(A\left(\theta^0\right)^t A\left(\theta^0\right)\right)^{-1} A\left(\theta^0\right)^t \,\text{diag}\left(\sqrt{-\phi''\left(q\left(\theta^0\right)\right)}\right)\left(\widehat{p}_n - q\left(\theta^0\right)\right) + o\left(\left\|\widehat{p}_n - q\left(\theta^0\right)\right\|\right). \quad \blacksquare$$

THEOREM 2. *Under Theorem 1 conditions, we have that*

(a) $\sqrt{n}(\hat{\theta}_\phi - \theta^0) \approx N(0, \Delta_0^tB_0\Delta_0)$;

(b) $\sqrt{n}(q(\hat{\theta}_\phi) - q(\theta^0)) \approx N(0, \text{diag}(\sqrt{-\phi''(q(\theta^0))})\Sigma_0^tB_0\Sigma_0\text{diag}(\sqrt{-\phi''(q(\theta^0))}))$.

PROOF.

(a) From above, we know that

$$\sqrt{n}\left(\widehat{p}_n - q\left(\theta^0\right)\right)^t \text{c.s.} N\left(0, \Omega_0\right),$$

and consequently,

$$\sqrt{n}\Delta_0^t \,\text{diag}\left(\sqrt{-\phi''\left(q\left(\theta^0\right)\right)}\right)\left(\widehat{p}_n - q\left(\theta^0\right)\right) \xrightarrow[n\to\infty]{L} N(0, \Sigma),$$

where

$$\Sigma = \Delta_0^t \text{diag}\left(\sqrt{-\phi''\left(q\left(\theta^0\right)\right)}\right) \Omega_0 \, \text{diag}\left(\sqrt{-\phi''\left(q\left(\theta^0\right)\right)}\right) \Delta_0.$$

So the result follows from Theorem 1.

(b) By (A1)

$$q\left(\hat{\theta}_\phi\right) - q\left(\theta^0\right) = J_0\left(\hat{\theta}_\phi - \theta^0\right) + o\left(\left\|\hat{\theta}_\phi - \theta^0\right\|\right).$$

Therefore,

$$\sqrt{n}\left(q\left(\hat{\theta}_\phi\right) - q\left(\theta^0\right)\right) \xrightarrow[n\to\infty]{L} N\left(0, J_0 \Delta_0^t B_0 \Delta_0 J_0^t\right). \qquad \blacksquare$$

REMARK 1. The matrix Ω_0, and consequently the matrices B_0, Δ_0, and Σ_0 figuring in Theorems 1 and 2, are known only if $P(\theta^0) \in \mathbf{P}_{\theta^0}$ is specified. If this is not the case and the values of these matrices are needed to obtain confidence intervals or critical regions of statistical tests, then we can estimate the matrices B_0, Δ_0, and Σ_0 consistently by replacing the unknown elements $p_{ij}(\theta^0)$ of $P(\theta^0)$ in Ω_0 by the relative frequencies

$$\widehat{p}_{nij} = \frac{\sum_{k=2}^{n} \mathbf{I}_{(i,j)}\left(X_{k-1}, X_k\right)}{\sum_{k=2}^{n} \mathbf{I}_{(i)}\left(X_{k-1}\right)}$$

as consistent estimates of elements $p_{ij}(\theta^0)$ of the matrix $P(\theta^0)$ (c.f. [10]).

3. COMPOSITE NULL HYPOTHESIS

In this section, we consider statistical tests of composite hypothesis introduced in Section 1 using the R_ϕ-divergence statistics (1). Assumptions (A1) and (A2) of Section 2 are supposed to be fulfilled.

First, it is necessary to obtain the asymptotic distribution of $R_\phi(\widehat{p}_n, q(\hat{\theta}))$ under H_0, where $\widehat{p}_n = (\hat{p}_{n1}, \ldots, \hat{p}_{nm})^t$ is the relative frequencies observed in the data $\mathbf{S}_n$ and $q(\hat{\theta}) = (q_1(\hat{\theta}), \ldots, q_m(\hat{\theta}))^t$ being $\hat{\theta}$ the maximum likelihood or minimum R_ϕ-divergence estimator.

THEOREM 3. Let $\phi : (0, \infty) \to R$ be a twice continuously differentiable concave function. Let $\widehat{p}_n$ be the relative frequencies vector, $q : \Theta \to \Delta_m$ a function with continuous second partial derivatives in a neighborhood of θ^0 and $\widehat{q}_{\phi^*} = q(\hat{\theta}_{\phi^*})$, then we have that

$$8nR_\phi(\widehat{p}_n, \widehat{q}_{\phi^*}) \xrightarrow[n\to\infty]{L} \sum_{i=1}^{m} \rho_i \chi_1^2,$$

where the χ_1^2 are independents and the ρ_i are the eigenvalues of the matrix

$$L_0 = \text{diag}\left(-\phi''\left(q\left(\theta^0\right)\right)\right) \Sigma_1$$

being

$$\Sigma_1 = \left(I - J_0 \Delta_0^t \text{diag}\left(\sqrt{-\phi^{*\prime\prime}\left(q\left(\theta^0\right)\right)}\right)\right) \Omega_0 \left(I - J_0 \Delta_0^t \text{diag}\left(\sqrt{-\phi^{*\prime\prime}\left(q\left(\theta^0\right)\right)}\right)\right)^t.$$

PROOF. By Lemma 1 in [8], on being $\widehat{p}_n$ and $\widehat{q}_{\phi^*}$ $\sqrt{n}$-consistent estimates, we have that

$$8nR_\phi(\widehat{p}_n, \widehat{q}_{\phi^*}) \approx n(\widehat{p}_n - \widehat{q}_{\phi^*})^t \text{diag}\left(-\phi''\left(q\left(\theta^0\right)\right)\right)(\widehat{p}_n - \widehat{q}_{\phi^*}).$$

From Theorem 2

$$\sqrt{n}\left(\widehat{q}_{\phi^*} - q(\theta^0)\right) \approx \sqrt{n} J_0 \Delta_0^t \, \text{diag}\left(\sqrt{-\phi^{*\prime\prime}\left(q\left(\theta^0\right)\right)}\right)\left(\widehat{p}_n - q(\theta^0)\right),$$

so

$$\sqrt{n}\left(\widehat{p}_n - \widehat{q}_{\phi^*}\right) = \sqrt{n}\left(\widehat{p}_n - q\left(\theta^0\right)\right) + \sqrt{n}\left(q\left(\theta^0\right) - \widehat{q}_{\phi^*}\right)$$
$$\approx \sqrt{n}\left(I - J_0\Delta_0^t \operatorname{diag}\left(\sqrt{-\phi''\left(q\left(\theta^0\right)\right)}\right)\right)\left(\widehat{p}_n - q\left(\theta^0\right)\right).$$

Consistently,

$$\sqrt{n}(\widehat{p}_n - \widehat{q}_{\phi^*}) \xrightarrow[n\to\infty]{L} N(0, \Sigma_1)$$

so $8nR_\phi(\widehat{p}_n, \widehat{q}_{\phi^*})$ is asymptotically distributed as $\sum_{i=1}^m \rho_i \chi_1^2$ where the χ_1^2 are independents and the ρ_i are the eigenvalues of the matrix L_0.

REMARK 2. Small values of $T = 8nR_\phi(\widehat{p}_n, \widehat{q}_{\phi^*})$ support H_0 but large values are not. Hence for large n, H_0 should be rejected at a level γ if $T > t_\gamma$ where t_γ is the upper γ-quantile of the distribution of $\sum_{i=1}^m \rho_i \chi_1^2$. This quantile can be approximated by the corresponding quantile of the distribution of $\lambda\chi_m^2$ where λ is determined so that $\sum_{i=1}^m \rho_i \chi_1^2$ and $\lambda\chi_m^2$ have the same expected values, that is $\lambda = (\sum_{i=1}^m \rho_i / m)$. See, for instance, Kotz *et al.* [16] and Rao and Scott [17]. In this case, t_γ is approximated by λ times the upper γ-quantile of the chi-square distribution with m degrees of freedom, that is $t_\gamma = \lambda\chi_{m,\gamma}^2$. However, the variance of $\lambda\chi_m^2$ is smaller than or equal to the variance of $\sum_{i=1}^m \rho_i \chi_1^2$ with equality if and only if all eigenvalues ρ_i are equal. Following Satterthwaite [18] or Scheffé [19], we can approximate the distribution of $\sum_{i=1}^m \rho_i \chi_1^2$ by the distribution $\lambda(1+a^2)\chi_\nu^2$ where a and ν are determined so that the two distributions have the same expected value and the same variance. This leads to

$$\nu = \frac{\left(\sum_{i=1}^m \rho_i\right)^2}{\sum_{i=1}^m \rho_i^2} \quad \text{and} \quad a^2 = \frac{\sum_{i=1}^m (\rho_i - \lambda)^2}{m\lambda^2},$$

or equivalently

$$\nu = \frac{\left(\operatorname{tr}\left(L_0\right)\right)^2}{\operatorname{tr}\left(L_0^2\right)} \quad \text{and} \quad \lambda\left(1 + a^2\right) = \frac{\operatorname{tr}\left(L_0^2\right)}{\operatorname{tr}\left(L_0\right)}.$$

In this case, we consider $t_\gamma = \lambda(1+a^2)\chi_{\nu,\gamma}^2$.

REMARK 3. The eigenvalues $\rho_1, \ldots, \rho_m$ depend not only on the unknown chain transition matrix $P(\theta^0)$, but also on the unknown stationary distribution $p(\theta^0)$. Replacing the matrix by the consistent estimate $\widehat{p}_{nij}$ defined in Remark 1 and $p(\theta^0)$ by the consistent estimate $\widehat{p}_n$, we obtain an estimate $\widehat{L}_n$ of the matrix L_0. Similarly as in Remark 1, we can argue that the eigenvalues $\rho_{n1}, \ldots, \rho_{nm}$ of $\widehat{L}_n$ are consistent estimates of the eigenvalues $\rho_1, \ldots, \rho_m$. Thus, the critical values are obtained by replacing the unknown eigenvalues $\rho_1, \ldots, \rho_m$ by their estimates $\rho_{n1}, \ldots, \rho_{nm}$.

REFERENCES

1. S. Tavaré and P.M.E. Altham, Serial dependence of observations leading to contingency tables, and corrections to chi-squared statistics, *Biometrika* **70**, 139–144, (1983).
2. M.L. Menéndez, D. Morales, L. Pardo and I. Vajda, Testing in stationary models based on f-divergences of observed and theoretical frequencies, *Kybernetika* **33** (5), 465–475, (1997).
3. I. Csiszár, Eine Informationtheoretische Ungleichung und ihre Anwendung auf den Bewis der Ergodizität von Markhoffschen Ketten, *Publ. Math. Inst. Hungar. Acad. Sci.* **Ser. A, 8**, 85–108, (1963).
4. M.C. Pardo, Goodness-of-fit tests for stationary distributions of Markov chains based on Rao's divergence, *Information Sciences*, (1998).
5. J. Burbea and C.R. Rao, On the convexity of some divergence measures based on entropy functions, *IEEE Transactions on Information Theory* **28**, 489–495, (1982).
6. M.C. Pardo, On Burbea-Rao divergences based goodness-of-fit tests for multinomial models (to appear).
7. M.C. Pardo, Goodness-of-fit tests based on Rao's divergence under sparseness assumptions (to appear).
8. M.C. Pardo and I. Vajda, About distances of discrete distributions satisfying the data proccesing theorem of information theory, *Trans. IEEE on Inform. Theory* **43** (4), 1288–1293, (1997).

9. M.L. Menéndez, D. Morales, L. Pardo and I. Vajda, Inference about stationary distributions of Markov chains based on divergences with observed frequencies, *Probability and Mathematical Statistics* (to appear).
10. P. Billingsley, Statistical methods in Markov chains, *Ann. Math. Statist.* **32**, 12–40, (1961).
11. M.W. Birch, A new proof of the Pearson-Fisher theorem, *Annals of Mathematical Statistics* **35**, 817–824, (1964).
12. Y.M.M. Bishop, S.E. Fienberg and P.W. Holland, *Discrete Multivariate Analysis Theory and Practice*, The MIT Press, Cambridge, MA, (1975).
13. T.R.C. Read and N. Cressie, *Goodness of Fit Statistics for Discrete Multivariate Data*, Springer, New York, (1988).
14. D. Morales, L. Pardo and I. Vajda, Asymptotic divergence of estimates of discrete distributions, *Journal of Statistical Planning and Inference* **48**, 347–369, (1995).
15. M.C. Pardo, Asymptotic behaviour of an estimator based on Rao's divergence, *Kybernetika* **33** (5), 489–504, (1997).
16. S. Kotz, N.M. Johnson and D.W. Boid, Series representation of quadratic forms in normal variables, I. Central case **AMS**, 823–837, (1967).
17. J.N.K. Rao and A.J. Scott, The analysis of categorical data from complex sample surveys: Chi-squared tests for goodness of fit and independence in two-way tables, *J. Amer. Stat. Assoc.* **76**, 221–230, (1981).
18. F.E. Satterthwaite, An aproximate distribution of estimates of variance components, *Biometrics* **2**, 110–114, (1946).
19. H. Scheffé, *The Analysis of Variance*, Wiley, (1959).