UDC 519.21

V. S. Korolyuk, acad. (Inst. Math. National Acad. Sci. Ukraine, Kiev),
D. Koroliuk, (Università di Roma II)

DIFFUSION APPROXIMATION OF STOCHASTIC MARKOV MODELS WITH PERSISTENT REGRESSION

ДИФУЗІЙНА АПРОКСИМАЦІЯ СТОХАСТИЧНИХ МАРКОВСЬКИХ МОДЕЛЕЙ З НЕЗНИКАЮЧОЮ РЕГРЕСІЄЮ

Sequences of sums of identically distributed random variables forming a homogeneous Markov chain are approximated by a time-discrete autoregression process of Ornstein–Uhlenbeck type.

Послідовність сум незалежних випадкових змінних з однаковим розподілом, що утворюють однорідний марковський ланцюг, апроксимована авторегресійним процесом з дискретним часом Орнштейна–Уленбека.

1. Introduction. Wide variety of stochastic systems can be considered possessing Markov property. The state of such systems are determined by influence of a great number of random factors. In many important cases, they can be formalized as stochastic systems which states are given by a normalized sum of independent and identically distributed (i.i.d.) random variables

$$\eta_r^{(N)} := \sum_{i=1}^{N} \frac{\xi_i(r)}{N}, \quad r \geq 0. \tag{1}$$

The main assumption is that, the distribution of terms in (1) depends only on the previous states of the sum,

$$P\{\xi_i(r+1) \in B \mid \eta_r^{(N)} = u\} = P(B \mid u). \tag{2}$$

For simplicity and almost without loss of generality in applications, we assume that the random variables $\xi_i(r)$ are bounded: $|\xi_i(r)| \leq 1$ hence, also $\left|\eta_r^{(N)}\right| \leq 1$.

The Markov chain $\eta_r^{(N)}$, $r \geq 0$, is determined by a rather complicated stochastic kernel

$$P^{(N)}(u, B) := P\{\eta_{r+1}^{(N)} \in B \mid \eta_r^{(N)} = u\} = P^{N*}(B \mid u),$$

where $P^{N*}(B \mid u)$ is the N-multiple convolution of probability distribution (2).

To extract some explicit information about behavior of the Markov chain (1) is an almost hopeless problem. The only way to get some almost explicit information is to use the Central Limit Theorem for correspondingly normalized sum (1). But the centering problem is appearing. That is why, the first step is to investigate equilibrium points of the Markov chain (1).

Different possibilities in this situation can be illustrated by a particular case of the discrete stochastic Wright–Fisher model in population genetics [1].

For simplicity we consider only one-locus two allele pure sampling model with a Markov regression [1]

$$v(p) = p(w_1 p + q)/w(p), \quad w(p) = w_1 p^2 + 2pq + w_2 q^2, \quad p + q = 1. \tag{3}$$

Here w_k, $k = 1, 2$, are the viability of $A_k A_k$ allele individuals and $w_{12} = w_{21} = 1$. The corresponding Markov chain fixing the frequency of the first allele, can be defined as (1) with Bernoulli distributed independent random variables taking two values:

$$P\{\xi_i(r+1) = 1 \mid \eta_r^{(N)} = p\} =$$

$$= 1 - P\{\xi_i(r+1) = 0 \mid \eta_r^{(N)} = p\} = c(p). \tag{4}$$

Therefore the regression function of the Markov chain in the model is $E\left[\eta_{r+1}^{(N)} / \eta_r^{(N)} = p\right] = c(p)$.

This regression function can be transformed into the following expression: $c(p) = c_0(p)/w(p) + p$, where

$$c_0(p) = p(1-p)(\sigma_2 - \sigma p), \quad \sigma_k := 1 - w_k, \quad k = 1, 2; \quad \sigma := \sigma_1 + \sigma_2. \tag{5}$$

The cubic parabola (5) has three real roots $p_0 = 0$, $p_1 = 1$, $\rho = \sigma_2/\sigma$. These roots contain all information about equilibrium points of frequencies of population zygotes.

The analysis of dynamic behavior of the frequency (1) can be formulated in the following behavior law (compare with [2] in the deterministic model).

I. For the values of viability coefficients $w_1 < 1$ and $w_2 < 1$ there exists the attractive equilibrium point $\rho = \sigma_2/\sigma$, $0 < \rho < 1$, so that the frequency of the first allele tends weakly to the equilibrium point:

$$\eta_r^{(N)} \Rightarrow \rho \quad \text{as} \quad r, N \to \infty;$$

II. For the values of viability coefficients $w_1 > 1$ and $w_2 > 1$ there exists repellent equilibrium point $\rho = \sigma_2/\sigma$, $0 < \rho < 1$, such that the behavior of the frequency of the first allele depends on the initial value:

$$\text{if } \eta_0^{(N)} < \rho, \text{ then } \eta_r^{(N)} \Rightarrow 0 \text{ as } r, N \to \infty;$$

$$\text{if } \eta_0^{(N)} > \rho, \text{ then } \eta_r^{(N)} \Rightarrow 1 \text{ as } r, N \to \infty.$$

III. For the values of viability coefficients $w_1 > 1 \geq w_2$ or $w_2 > 1 \geq w_1$, only one of two alleles survives, that is:

$$\eta_r^{(N)} \Rightarrow 1 \quad \text{or} \quad \eta_r^{(N)} \Rightarrow 0 \quad \text{as} \quad r, N \to \infty.$$

These three variants exhaust all possibilities. A more complete analysis of the Wright–Fisher discrete model including the discrete diffusion approximation will be done in our forthcoming papers.

Here we investigate the Markov chain (1) under some technically reasonable condition.

2. Equilibrium point. As usual, in order to get a diffusion approximation, we have to calculate the first two conditional moments.

The main suppositions are the following:

$$\boldsymbol{E}[\xi_i(r+1) \mid \eta_r^{(N)} = u] = C(u), \tag{6}$$

$$\boldsymbol{D}[\xi_i(r+1) \mid \eta_r^{(N)} = u] = \boldsymbol{E}[\xi_i^2(r+1) \mid \eta_r^{(N)} = u] - C^2(u) =: B^2(u), \tag{7}$$

where the functions $C(u)$ and $B(u)$ can be depend on N, but only in some smooth sense.

Notice that relations (6) and (7) imply

$$\boldsymbol{E}[\eta_{r+1}^{(N)} \mid \eta_r^{(N)} = u] = C(u). \tag{8}$$

$$D[\eta_{r+1}^{(N)} \mid \eta_r^{(N)} = u] = E[(\eta_{r+1}^{(N)})^2 \mid \eta_r^{(N)} = u] - C^2(u) = B^2(u)/N. \quad (9)$$

Theorem 1 (Equilibrium point). *Let the regression function $C(u)$ satisfy the Lipshitz condition*

$$|C(u) - C(v)| \le c\,|u - v| \quad \text{with} \quad c < 1 \quad (10)$$

and be convex (up or down), hence, there exists in the interval $(-1, +1)$ a unique solution of equation

$$C(\rho) = \rho \quad \text{with} \quad |\rho| < 1. \quad (11)$$

Let the conditional variance $B(u)$ be bounded: $|B(u)| \le B < \infty$.

Then the Markov chain $\eta_r^{(N)}$, $r \ge 0$, has an equilibrium point: under the condition $E_\pi \eta_0^{(N)} = \rho$, there exists the limit in probability as $N \to \infty$,

$$P - \lim_{N \to \infty} \eta_r^{(N)} \Rightarrow \rho \quad \text{for all} \quad r \ge 1, \quad (12)$$

or, in general case $\eta_0^{(N)} \ne \rho$, there exists the mean square limit

$$\lim_{N \to \infty} E_\pi\left[\eta_{r_N}^{(N)} - \rho\right]^2 = 0 \quad \text{for all} \quad r_N \to \infty. \quad (13)$$

If, in addition, the regression function $C(u)$ hasthe bounded second derivative, $|C''(u)| = C_1$, then the weak convergence of the processes

$$\sum_{r=1}^{N_t} \eta_r^{(N)}/N \Rightarrow \rho t \quad \text{as} \quad N \to \infty \quad (14)$$

takes place.

***Proof of Theorem* 1.** Considering the mean square convergence, we have

$$E_\pi\left[\eta_r^{(N)} - \rho\right]^2 = E_\pi\left[\eta_r^{(N)} - E\left[\eta_r^{(N)} \mid \eta_{r-1}^{(N)}\right]\right]^2 + $$
$$+ E_\pi\left[E\left[\eta_r^{(N)} \mid \eta_{r-1}^{(N)}\right] - \rho\right]^2. \quad (15)$$

By definition (9) of the regression function, we obtain

$$I_r^{(N)} := E_\pi\left[\eta_r^{(N)} - E\left[\eta_r^{(N)} \mid \eta_{r-1}^{(N)}\right]\right]^2 =$$
$$= E_\pi\left[\eta_r^{(N)} - C\left(\eta_{r-1}^{(N)}\right)\right]^2 = E_\pi B\left(\eta_{r-1}^{(N)}\right)/N$$

and, by definition of ρ, we have

$$H_r^{(N)} := E_\pi\left[E\left[\eta_r^{(N)} \mid \eta_{r-1}^{(N)}\right] - \rho\right]^2 =$$
$$= E_\pi\left[C\left(\eta_{r-1}^{(N)}\right) - \rho\right]^2 = E_\pi\left[C\left(\eta_{r-1}^{(N)}\right) - C(\rho)\right]^2.$$

Using Lipshitz condition (10), we can estimate $II_r^{(N)}$ as

$$II_r^{(N)} \le c^2 \, E_\pi\Big[\eta_{r-1}^{(N)} - \rho\Big]^2 .$$

By boundedness of the conditional variance, we can estimate $I_r^{(N)}$ as $I_r^{(N)} \le B/N$. So that we obtain the following recursive estimate for (15):

$$\Delta_r^{(N)} := E_\pi\Big[\eta_r^{(N)} - \rho\Big]^2 = I_r^{(N)} + II_r^{(N)} \le B/N + c^2 \Delta_{r-1}^{(N)}. \tag{16}$$

By induction, we obtain from (16):

$$\Delta_r^{(N)} \le \frac{1 - c^{2r}}{1 - c^2} B/N + c^{2r} \Delta_0^{(N)}, \quad \Delta_0^{(N)} := E_\pi\Big[\eta_0^{(N)} - \rho\Big]^2. \tag{17}$$

In the case $E\eta_0^{(N)} = \rho$, one has $\Delta_0^{(N)} = \rho(1-\rho)/N$. Hence, the mean square convergence

$$\lim_{N\to\infty} E_\pi\Big[\eta_r^{(N)} - \rho\Big]^2 = 0 \quad \text{for} \quad r \ge 1 \tag{18}$$

follows from (17).

By Chebyshev's inequality, the convergence (12) in probability follows from the mean square convergence (18).

In the case $E_\pi \eta_0^{(N)} \ne \rho$ we use the fact that $c < 1$, hence $c^{2r_N} \to 0$ as $r_N \to \infty$. So that the mean square convergence (13) follows from inequality (17).

We prove convergence (14) by employing the martingale approach. Consider the sum of the martingale-differences

$$\mu_t^{(N)} := \sum_{r=0}^{N_{t-1}} \Big[\eta_{r+1}^{(N)} - E\big[\eta_{r+1}^{(N)} \,|\, \eta_r^{(N)}\big]\Big] \Big/ N =$$

$$= \sum_{r=0}^{N_{t-1}} \Big[\eta_{r+1}^{(N)} - C\big(\eta_r^{(N)}\big)\Big] \Big/ N . \tag{19}$$

The square characteristics of the martingale (19) is the following:

$$\Big\langle \mu^{(N)} \Big\rangle_t = \sum_{r=0}^{N_{t-1}} E\Big[\Big(\eta_{r+1}^{(N)} - C\big(\eta_r^{(N)}\big)\Big)^2 \,\Big|\, \eta_r^{(N)}\Big] \Big/ N^2 = \sum_{r=0}^{N_{t-1}} B\big(\eta_r^{(N)}\big) \Big/ N^2 . \tag{20}$$

So that we have the following asymptotic estimate:

$$\Big\langle \mu^{(N)} \Big\rangle \le Bt/N \to 0 \quad \text{as} \quad N \to \infty \tag{21}$$

hence, by the limit theorem for martingales [3], we obtain

$$\mu_t^{(N)} \Rightarrow 0 \quad \text{as} \quad N \to \infty \tag{22}$$

in the weak sense.

Now we transform martingale (19) in the following way:

$$\mu_t^{(N)} = \sum_{r=0}^{N_{t-1}} \eta_{r+1}^{(N)}/N - \sum_{r=0}^{N_{t-1}} C\big(\eta_r^{(N)}\big) \Big/ N =$$

$$= \sum_{r=1}^{N_t} \eta_r^{(N)}/N - \rho t - \sum_{r=0}^{N_{t-1}} \left[C\left(\eta_r^{(N)}\right) - \rho \right] / N. \tag{23}$$

Taking into account (11), we obtain by using the Taylor formula that

$$C\left(\eta_r^{(N)}\right) - \rho = b\left[\eta_r^{(N)} - \rho\right] + \left[\eta_r^{(N)} - \rho\right]^2 c_r^{(N)}/2, \tag{24}$$

where

$$b := c'(\rho), \ c_r^{(N)} := c''\left(\theta_r^{(N)}\right).$$

Notice that, by assumption of Theorem 1, $|b| < 1$ and $c_r^{(N)} \le C_1 < \infty$. So that martingale (23) transforms into the following form:

$$\mu_t^{(N)} = (1-b)\left[\sum_{r=1}^{N_t} \eta_r^{(N)}/N - pt\right] - \sum_{r=0}^{N_{t-1}} \left[\eta_r^{(N)} - \rho\right]^2 c_r^{(N)}/2N + $$

$$+ b\left(\eta_{N_t}^{(N)} - \eta_0^{(N)}\right)^2 / N. \tag{25}$$

The stationary absolute mean value of the second term in (25) admits the following estimate:

$$E_\pi \left| \sum_{r=0}^{N_{t-1}} \left[\eta_r^{(N)} - \rho\right]^2 c_r^{(N)}/2N \right| \le$$

$$\le C_1 \sum_{r=0}^{N_{t-1}} E_\pi\left[\eta_r^{(N)} - \rho\right]^2 / N = C_1 \sum_{r=0}^{N_{t-1}} \Delta_r^{(N)}/N,$$

and, by inequality (16), we obtain

$$\sum_{r=1}^{N_t} \Delta_r^{(N)} \le Bt + c^2 \sum_{r=1}^{N_{t-1}} \Delta_{r-1}^{(N)}.$$

So that

$$\sum_{r=1}^{N_t} \Delta_r^{(N)} \le \left[Bt + c^2\left(\Delta_0^{(N)} + \Delta_{N_t}^{(N)}\right)\right] / \left(1 - c^2\right). \tag{26}$$

Hence, the second term in (25) admits the following estimate:

$$E_\pi \left| \sum_{r=0}^{N_{t-1}} \left[\eta_r^{(N)} - \rho\right]^2 c_r^{(N)}/2N \right| \le (A_0 + A_1 t)/N \to 0 \ \text{as} \ N \to \infty \tag{27}$$

with some constants A_0, A_1.

By the limit theorem for martingales in a series scheme [3], we conclude that the weak convergence (14) takes place which completes the proof of Theorem 1.

3. Diffusion approximation. Introduce the normalized process

$$\zeta_r^{(N)} := \left(\eta_r^{(N)} - \rho\right)\sqrt{N}, \quad r \ge 0. \tag{28}$$

Theorem 2 (diffusion approximation). *Let the conditions of Theorem 1 hold and the variance $B(u)$ by a continuous function. Then the following weak convergence takes place:*

$$\zeta_r^{(N)} \Rightarrow \zeta_r, \quad r \geq 1, \quad as \quad N \to \infty, \tag{29}$$

where the limit process ζ_r, $r \geq 1$, is determined by the following autoregression:

$$\zeta_{r+1} = b\zeta_r + \sigma w_{r+1}, \quad r \geq 0, \tag{30}$$

where w_r, $r \geq 1$, is a sequence of i.i.d. standard normally distributed random variables and

$$b = C'(\rho), \quad \sigma^2 = B(\rho). \tag{31}$$

The initial condition ζ_0 is defined by the relation $\zeta_0^{(N)} \Rightarrow \zeta_0$, as $N \to \infty$.

Proof. Introduce the sum of martingale-differences

$$\mu_t^{(N)} := \sum_{r=0}^{t-1} \left[\zeta_{r+1}^{(N)} - E\left[\zeta_{r+1}^{(N)} \mid \zeta_r^{(N)}\right]\right]. \tag{32}$$

Notice that

$$E\left[\zeta_{r+1}^{(N)} \mid \zeta_r^{(N)}\right] = \sqrt{N}\left[E\left[\eta_{r+1}^{(N)} \mid \eta_r^{(N)}\right] - \rho\right] = \sqrt{N}\left[C\left(\eta_r^{(N)}\right) - \rho\right].$$

By using the Taylor formula in (24), we obtain

$$E\left[\zeta_{r+1}^{(N)} \mid \zeta_r^{(N)}\right] = b\,\zeta_r^{(N)} + (\zeta_r^{(N)})^2 c_r^{(N)} / 2\sqrt{N}.$$

So that the martingales (32) has the following representation:

$$\mu_t^{(N)} := \sum_{r=0}^{t-1} \left[\zeta_{r+1}^{(N)} - b\zeta_r^{(N)}\right] + \delta_t^{(N)}, \tag{33}$$

where

$$\delta_t^{(N)} := \frac{1}{2\sqrt{N}} \sum_{r=0}^{t-1} \left(\zeta_r^{(N)}\right)^2 c_r^{(N)} = \frac{\sqrt{N}}{2} \sum_{r=0}^{t-1} \left[\eta_r^{(N)} - \rho\right]^2 c_r^{(N)}.$$

By using (2.12), we obtain the following estimate:

$$E_\pi\left|\delta_r^{(N)}\right| = \frac{\sqrt{N}}{2} \sum_{r=0}^{t-1} E_\pi\left|\left[\eta_r^{(N)} - \rho\right]^2 c_r^{(N)}\right| \leq$$

$$\leq \frac{C_1\sqrt{N}}{2} \sum_{r=0}^{t-1} \Delta_r^{(N)} \leq At/\sqrt{N}.$$

So the following mean value convergence takes place:

$$\sup_{0 \leq t \leq T} E_\pi\left|\delta_r^{(N)}\right| \to 0 \quad \text{as} \quad N \to \infty. \tag{34}$$

Square characteristic of martingale (32), according to formulas (9) and (28), has the following representation:

$$\left\langle \mu^{(N)} \right\rangle_t = \sum_{r=0}^{t-1} B\left(\eta_r^{(N)}\right).$$

By Average Theorem 1, we have

$$B\left(\eta_r^{(N)}\right) \underset{N\to\infty}{\Rightarrow} B(\rho) := \sigma^2.$$

So that

$$\left\langle \mu^{(N)} \right\rangle_t \underset{N\to\infty}{\Rightarrow} \sigma^2 t. \tag{35}$$

By limit theorem for martingales in the series scheme [3], we conclude that the following limit convergence takes place:

$$W_t^{(N)} := \sum_{r=0}^{t-1} \left[\zeta_{r+1}^{(N)} - b\zeta_r^{(N)}\right] \Rightarrow W_t \quad \text{as} \quad N\to\infty. \tag{36}$$

An important property of the limit process ζ_t gives the following lemma.

Lemma 1. *The process* W_t, $t \ge 0$, *defined in* (36), *has the following representation:*

$$W_t = \sum_{r=1}^{t} w_r, \quad t \ge 1, \tag{37}$$

where w_r, $t \ge 1$, *are independent normally distributed random variables with variance* $\boldsymbol{E} w_r^2 = \sigma^2 = B(\rho)$.

Proof. By Central Limit Theorem for sums of i.i.d. random variables, the following weak convergence takes place:

$$w_r^{(N)} := \zeta_{r+1}^{(N)} - b\zeta_r^{(N)} \Rightarrow w_r, \quad t \ge 1, \quad \text{as} \quad N\to\infty. \tag{38}$$

Now let us calculate the first two stationary moments of $w_r^{(N)}$:

$$\boldsymbol{E}_\pi w_r^{(N)} = \boldsymbol{E}_\pi\left[\zeta_{r+1}^{(N)} - b\zeta_r^{(N)}\right] = \boldsymbol{E}_\pi\left[\boldsymbol{E}\left[\zeta_{r+1}^{(N)} \mid \zeta_r^{(N)}\right] - b\zeta_r^{(N)}\right].$$

By virtue of (33), we obtain

$$\boldsymbol{E}_\pi w_r^{(N)} = \boldsymbol{E}_\pi\left[\left(\zeta_r^{(N)}\right)^2 c_r^{(N)}\right] \Big/ 2\sqrt{N} \to 0 \quad \text{as} \quad N\to\infty.$$

Therefore

$$\boldsymbol{E}_\pi w_r = 0, \quad r \ge 1. \tag{39}$$

By using (33), we obtain the following representation for the second stationary moment of $w_r^{(N)}$:

$$\boldsymbol{E}_\pi\left(w_r^{(N)}\right)^2 = \boldsymbol{E}_\pi\left[\zeta_{r+1}^{(N)} - b\zeta_r^{(N)}\right]^2 =$$

$$= \boldsymbol{E}_\pi\left[\zeta_{r+1}^{(N)} - \boldsymbol{E}\left[\zeta_{r+1}^{(N)} \mid \zeta_r^{(N)}\right] + \left(\zeta_r^{(N)}\right)^2 c_r^{(N)} / 2\sqrt{N}\right]^2 =$$

$$= E_\pi\left[\zeta_{r+1}^{(N)} - E\left[\zeta_{r+1}^{(N)} \mid \zeta_r^{(N)}\right]\right]^2 + E_\pi\left[\left(\zeta_r^{(N)}\right)^2 c_r^{(N)}\right]^2 \Big/ 4N.$$

By definition (9) of conditional variance, we have

$$E_\pi\left(w_r^{(N)}\right)^2 = B\left(\eta_r^{(N)}\right) + E_\pi\left[\left(\zeta_r^{(N)}\right)^2 c_r^{(N)}\right]^2 \Big/ 4N. \tag{40}$$

The second term in (40) admits the following estimate:

$$E_\pi\left[\left(\zeta_r^{(N)}\right)^2 c_r^{(N)}\right]^2 \Big/ 4N \le$$

$$\le C_1 E_\pi\left[\eta_r^{(N)} - \rho\right]^4 \Big/ 4N \to 0 \quad \text{as} \quad N \to \infty. \tag{41}$$

By Theorem 1, we conclude that

$$\lim_{N \to \infty} E_\pi\left(w_r^{(N)}\right)^2 = B(\rho) = \sigma^2,$$

and we finally obtain

$$E_\pi w_r^2 = \sigma^2. \tag{42}$$

The limit martingale (37), by formula (36), has the variance equal to the sum of variance of terms in (37). Hence, the random variables w_r, $r \ge 1$, are independent. The proof of Theorem 1 is complete.

Remark. Diffusion approximation (29) can be extended to the process

$$\zeta_{r_N+r}^{(N)} = \left(\eta_{r_N+r}^{(N)} - \rho\right)\sqrt{N}, \quad r \ge 0,$$

for $r_N \to \infty$ such that $E_\pi \eta_{r_N}^{(N)} = \rho + c/\sqrt{N}$.

The considered discrete diffusion approximation scheme can be extended to multidimensional stochastic Markov models with persistent regression.

1. *Wright S.* Evolution in Mendelian populations // Genetics. – 1931. – **16**. – P. 97–159.
2. *Charlesworth B.* Selection in population with overlapping generations. III. Conditions for genetic equilibrium // Theor. Pop. Biol. – 1972. – **4**. – P. 377–395.
3. *Hall P., Heyde C. C.* Martingale Limit Theory and its Application // Annals of Probability. – 1980. – 308 p.
4. *Korolyuk D., Korolyuk V. S.* Discrete diffusion model of cooperativity in proteins. – Preprint Volterra, 170, 1994 (submitted to: Appl. Stoch. Models and Data Analysis).

Received 22.02.95