Information Geometry and the Wright-Fisher model of mathematical population genetics

Summary of the thesis submitted by Tat Dat Tran

1. Introduction

My Ph. D. thesis addresses a systematic approach to stochastic models in population genetics, in particular, the Wright-Fisher models affected only by the random genetic drift. I used various mathematical methods such as Probability, PDE, and Geometry to answer an important question: "How do genetic change factors (random genetic drift, selection, mutation, migration, random environment, etc.) affect the behavior of gene frequencies or genotype frequencies in generations?".

In a Hardy-Weinberg model, the Mendelian population model of a very large number of individuals without genetic change factors, the answer is simple by the Hardy-Weinberg principle (see [7, 12]):

Theorem 1. *In a Hardy-Weinberg model, gene frequencies remain unchanged from generation to generation, and genotype frequencies from the second generation onward remain also unchanged from generation to generation.*

With directional genetic change factors (selection, mutation, migration), we will have a deterministic dynamics of gene frequencies, which has been studied rather in detail (see [7, 12]). With non-directional genetic change factors (random genetic drift, random environment), we will have a stochastic dynamics of gene frequencies, which has been studied with much more interests. A combination of these factors has also been considered (see [8, 9]).

2. Problem

We consider a monoecious diploid population of fixed size N with $n+1$ possible alleles $A_1,\dots,A_{n+1}$ at a given locus A, and assume that the evolution of population was only affected by the random genetic drift. Denote by X^i_τ the relative frequency of alleles A_i $(i=1,\dots,n+1)$ in population at generation τ, we shall obtain a discrete time discrete space Markov chain $\{\mathbf{X}_t\}_{t\in\frac{1}{2N}\mathbb{N}_0}$ (where $t=\frac{\tau}{2N}$) in

$$S_n^{(2N)}=\left\{\mathbf{x}=(x^1,\dots,x^n)\Big|x^i\in\left\{0,\frac{1}{2N},\dots,1\right\},\sum_{i=1}^n x^i\le 1\right\},$$

with the transition probability function

$$\begin{aligned} p_{\boldsymbol{\alpha},\boldsymbol{\beta}} =&\mathbb{P}\left(\mathbf{X}_{t+\delta t}=\boldsymbol{\beta}|\mathbf{X}_t=\boldsymbol{\alpha}\right)\\ =&\frac{2N!}{(2N\beta^1)!\dots(2N\beta^{n+1})!}\left(\alpha^1\right)^{2N\beta^1}\cdots\left(\alpha^{n+1}\right)^{2N\beta^{n+1}}\quad\text{for } \boldsymbol{\alpha},\boldsymbol{\beta}\in S_n^{(2N)}. \end{aligned}$$

where $\alpha^{n+1}=1-\alpha^1-\ldots-\alpha^n;\beta^{n+1}=1-\beta^1-\ldots-\beta^n,\delta t=\frac{1}{2N}$.

The question is that what is the behavior of $\{\mathbf{X}_t\}$ in time t and its stochastic quantities.

3. Mathematical methods

When N is large enough, we can approximate this discrete Markov chain to a continuous Markov process in

$$S_n=\left\{\mathbf{x}=(x^1,\dots,x^n)\Big|x^i\ge 0,\sum_{i=1}^n x^i\le 1\right\},$$

with the same characteristics. In 1931, Kolmogorov (see [16]) first introduced a nice relation between a continuous Markov process and diffusion equations. These equations called the (backward/forward) Kolmogorov equations which have been first applied in population genetics in 1945 by Wright (see [18]). In our model, the forward Kolmogorov equation (known to physicists under the name Fokker-Planck equation) is

$$\frac{\partial u(\mathbf{x},\mathbf{p},t)}{\partial t}=\sum_{i,j=1}^n\frac{1}{2}\frac{\partial^2(b^{ij}(\mathbf{x})u(\mathbf{x},\mathbf{p},t))}{\partial x^i\partial x^j},\quad\forall\mathbf{x},\mathbf{p}\in\mathrm{int}S_n,$$

and the Kolmogorov backward equations is

$$\frac{\partial u(\mathbf{x},\mathbf{p},t)}{\partial t}=\sum_{i,j=1}^{n}\frac{1}{2}b^{ij}(\mathbf{p})\frac{\partial^2 u(\mathbf{x},\mathbf{p},t)}{\partial p^i\partial p^j},\quad \forall \mathbf{x},\mathbf{p}\in \mathrm{int}S_n,$$

where $\mathbf{p}$ is the initial state, $\mathbf{x}$ is the present state, $b^{ij}(\mathbf{x})=x^i(\delta_{ij}-x^j)$, and $u(\mathbf{x},\mathbf{p},t)$ is the "local" probability density function (local solution) of the continuous Markov process up to the first exit time.

Note that these equations are singular parabolic equations (diffusion coefficients vanish on boundary). To solve them, we use generalized hypergeometric functions (see [4, 5]). To know more about what will happen after the first exit time, or more general, the behavior of whole process, in joint work with J. Hofrichter, we define the global solution by moment conditions, calculate the component solutions by boundary flux method and combinatorics method (also see [13]).

One interesting property is that some statistical quantities of interest are solutions of a singular elliptic second order linear equation with discontinuous (or incomplete) boundary values. A lot of papers, textbooks have used this property to find those quantities. However, the uniqueness of these problems have not been proved. Littler, in his PhD thesis in 1975 (see [17]), took up the uniqueness problem but his proof, in my view, is not rigorous. In joint work with J. Hofrichter, we showed two different ways to prove the uniqueness rigorously. The first way is the approximation method. The second way is the blow-up method which is conducted by J. Hofrichter (see [13]).

By applying the Information Geometry, which was first introduced by Amari in 1985 (see [1]), we see that the local state space is an Einstein space, and also a dually flat manifold with the Fisher metric; the differential operator of the Kolmogorov equation is the affine Laplacian which can be represented in various coordinates and on various spaces. Dynamics on the whole state space explains some biological phenomena (see [2, 3]).

References

[1] S. Amari, *Differential geometrical methods in statistics*, Lecture Notes in Statistics, Springer-Verlag, Berlin, 1985.

[2] P. L. Antonelli, C. Strobeck, *The Geometry of random drift I. Stochastic Distance and Diffusion*, Advances in Applied Probability, Vol. **9**, No. **2**, (1977), 238-249.

[3] P. L. Antonelli, J. Chapin, G. M. Lathrop, K. Morgan, *The Geometry of Random Drift II. The Symmetry of Random Genetic Drift*, Advances in Applied Probability, Vol. **9**, No. **2** (1977), 250-259.

[4] P. Appell, *Sur les series hypergeometriques de deux variables, et sur des equations differentielles lineaires aux derivees partielles*, C. R. Acad. Sci. Paris **90** (1880), 296-299.

[5] P. Appell, *Sur les series hypergeometriques de deux variables, et sur des equations differentielles lineaires simultanees aux derivees partielles*, C. R. Acad. Sci. Paris **90** (1880), 731-734.

[6] D. G. Aronson, H. F. Weinberger, *Multidimensional nonlinear diffusion arising in population genetics*, Adv. Math. **30** (1978) 33-76.

[7] R. Bürger, *The mathematical theory of selection, recombination, and mutation*, Willey Series in Mathematical and Computational Biology, John Willey & Sons, LTD, 2000.

[8] J. F. Crow, M. Kimura, *An introduction to population genetics theory*, Harper and Row, N.Y., 1970.

[9] Warren J. Ewens, *Mathematical Population Genetics I. Theoretical Introduction*, Springer-Verlag New York Inc., Interdisciplinary Applied Mathematics, 2nd ed., 2004.

[10] W. Feller, *The parabolic differential equations and the associated semi-groups of transformations*, Ann. Math., **Vol 55**, No. 3 (1952) 468-519.

[11] W. Feller, *Diffusion processes in one dimension*, Trans. Amer. Math. Soc., **77** (1954) 1-31.

[12] J. Hofbauer, K. Sigmund, *The theory of evolution and dynamical systems*, London Mathematical Society Student Texts, Cambridge University Press, Vol. **7**, 1988.

[13] J. Hofrichter, *On the diffusion approximation of Wrigh-Fisher models with several alleles and loci and its geometry*, PhD thesis, to be published.

[14] J. Jost, *Information geometry*, Lecture Notes, IMPRS, 2007.

[15] J. Jost, *Riemannian geometry and geometric analysis*, Universitext, Springer, 5th ed., 2008.

[16] Kolmogorov, *Über die analytischen Methoden in der Wahrscheinlichkeitsrechnung*, Mathematische Annalen, Vol. 104, Nr. 1, (1931) 415-458.

[17] R. A. Littler, *Multidimensional stochastic models in genetics*, PhD Thesis, Monash University, 1975.

[18] S. Wright, *The differential equation of the distribution of gene frequencies*, Proc. Nat. Acad. Sci., **31** (1945), 382-389.