Bayesian countable representation of some population genetics diffusions

Matteo Ruggiero[1]

University of Turin

BNR, Cambridge, August 6-10, 2007

[1]Joint work with Stephen G. Walker (University of Kent)

Abstract

The present work

- Provides an explicit Bayesian construction of some Fleming-Viot (F-V) measure-valued diffusions, by means of urn schemes
- Highlights new connections between population genetics and Bayesian nonparametrics
- Elicits a previously unknown stationary distribution.

Outline

Outline

Fleming-Viot Processes and Population Genetics
Generalities

Consider a population whose types belong to the set $\{1, \dots, k\}$.

The state of the population with k different types can be represented at a single time point by a probability distribution on the simplex

$$\Delta_k = \left\{ (p_1, \dots, p_k) : \ p_1, \dots, p_k \geq 0, \ \sum_{i=1}^{k} p_i = 1 \right\}$$

where p_i is the relative frequency of the i-th type.

Example

k-dimensional Wright-Fisher diffusion.

Fleming-Viot Processes and Population Genetics

Generalities

Consider now a more general framework, by substituting

- $\{1, \ldots, k\}$ with a complete separable metric space $\mathcal{X}$
- the simplex Δ_k with the set of Borel probability measure $\mathscr{P}(\mathcal{X})$, endowed with the topology of weak convergence

Fleming-Viot process

Fleming and Viot (1979) introduced a class of probability-measure-valued diffusion processes in this setting.

The Fleming-Viot process has sample paths in the *càdlàg* space $D_{\mathscr{P}(\mathcal{X})}[0, \infty)$.

The *càdlàg* space is needed to deal with finite-population approximations.

Fleming-Viot Processes: applications and related fields

1. Population genetics

   Donnelly and Kurtz (1993,1996,1999a,1999b), Ethier (1990), Ethier and Kurtz (1981,1987,1993a), Ethier and Shiga (2000)

2. Measure-valued Markov processes / Infinite-dimensional diffusions

   Dynkin (1989), Ethier and Kurtz (1986,1992,1994,1998), Ethier and Griffiths (1987,1993b), Barbour, Ethier and Griffiths (2000), Dawson and Hochberg (1982), Handa (2002), Shiga (1981), Dawson (1993)

3. Combinatorial stochastic processes / Coalescent theory

   Birkner *et al.* (2005), Donnelly and Kurtz (1996,1999), Stephens (2001)

4. Ancestral inference / Genealogical processes (1+3)

   Tavarè (1984), Stephens (2000,2001), Stephens and Donnelly (2000,2002), Krone and Neuhauser (1997)

Fleming-Viot Processes and Population Genetics

Representations

The Fleming-Viot process admits

- *Characterisation*

  as the unique solution of a martingale problem for a certain infinitesimal generator $\mathbb{A}\varphi(\mu)$, with domain

  $$\mathscr{D}(\mathbb{A}) = \{\varphi \in B(\mathscr{P}(\mathcal{X})) : \varphi(\mu) = F(\langle f_1, \mu\rangle, \ldots, \langle f_m, \mu\rangle)\}.$$

  e.g. Fleming and Viot (1979), Ethier and Kurtz (1993,1994)

- *Representation*

  as the weak limit of (a transformation of) a particle process in $D_{\mathcal{X}^n}[0, \infty)$, which describes the dynamics of a countable number of individuals.

  e.g. Donnelly and Kurtz (1996,1999)

Fleming-Viot Processes and Population Genetics

Evolutionary Mechanisms

The Fleming-Viot model describes the following genetic mechanisms:

- Mutation

  Introduces *novel variants* into the population according to its own random process.

- Genetic drift

  The intrinsic statistical drift of frequencies over time, due to random *resampling* effects.

- Fertility selection

  Makes some genotypes more likely than others to *generate offspring*

- Viability selection

  Makes some genotypes more likely than others to *survive longer*

Fleming-Viot Processes and Bayesian Nonparametrics
The Neutral Diffusion Model

In the simplest case, the Fleming-Viot process has

- parent-independent mutation operator:

$$\frac{1}{2}\theta \int_{\mathcal{X}} \left[f(y) - f(x)\right] \nu_0(\mathrm{d}y), \qquad \theta > 0$$

  where ν_0 is a diffuse distribution
- genetic drift: uniform resampling of individuals
- no selection

and it is called neutral diffusion model.

Ethier and Kurtz (1994)

The stationary distribution of the neutral diffusion model is the Dirichlet process with parameters (θ, ν_0).

Fleming-Viot Processes and Bayesian Nonparametrics cont'd

Ethier and Griffiths (1993)

The transition function of the neutral diffusion model is

$$P(t,\mu,\mathrm{d}\nu)=\sum_{m=0}^{\infty}d_m(t)\int_{\mathcal{X}^m}\mathscr{D}ir\left(\mathrm{d}\nu\,\middle|\,\theta\nu_0+\sum_{i=1}^{m}\delta_{X_i}\right)\mu(\mathrm{d}X_1)\ldots\mu(\mathrm{d}X_m) \qquad (1)$$

where $d_m(t)=P(D_t=m)$ with D_t a death process.

See also Walker, Hatjispyros and Nicoleris (2007).

Ethier and Kurtz (1994)

The stationary distribution of the F-V process with diploid fertility selection is

$$C\,e^{\langle\sigma,\mu^2\rangle}\,\mathscr{D}ir(\mathrm{d}\mu|\theta\nu_0)$$

where $\langle\sigma,\mu^2\rangle=\iint\sigma(x,y)\mu(\mathrm{d}x)\mu(\mathrm{d}y)$ and C is a constant.

Outline

Neutral Diffusion Model

Construction via Blackwell-MacQueen Urn Scheme

Given the vector of individuals $(x_1, \dots, x_n)$, define a Gibbs sampler based Markov jump process with the following transitions:

1. Select x_i from $\{x_1, \dots, x_n\}$ with probability $1/n$
2. Sample a holding time from $\text{Exp}(\lambda_n)$, where

$$\lambda_n = \frac{1}{2} n(\theta + n - 1)$$

3. At the next renewal, replace x_i with a sample from the Blackwell-MacQueen urn predictive

$$p_n(\mathrm{d}x_i' \,|\, x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n) = \frac{\theta \nu_0(\mathrm{d}x_i') + \sum_{k \neq i}^{n} \delta_{x_k}(\mathrm{d}x_i')}{\theta + n - 1}$$

4. The other components stay constant, so that the arrival state is

$$(x_1, \dots, x_{i-1}, x_i', x_{i+1}, \dots, x_n)$$

Note: x_i' is sampled from the full-conditional of x_i.

Neutral Diffusion Model

Convergence of Measure-Valued Process

Define now the process of empirical measures with sample paths in $D_{\mathscr{P}(\mathcal{X})}[0,\infty)$

$$\mu_n(t) = \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i(t)}.$$

By means of uniform convergence of the generators, namely

$$\sup_{x^n\in\mathcal{X}^n} |\langle A^n f, \mu_n^{(m)}\rangle - \mathbb{A}\varphi(\mu_n)| \to 0, \qquad f \in B(\mathcal{X}^m)$$

it can be shown that the process of empirical measures weakly converges to the neutral diffusion model

$$\{\mu_n(t),\, t\geq 0\} \Rightarrow \{\mu_t,\, t\geq 0\}$$

in the Skorohod topology of $D_{\mathscr{P}(\mathcal{X})}[0,\infty)$.

In the limit for n, the stationary distribution of $\{\mu_n(t),\, t\geq 0\}$ is $\mathscr{D}ir(\,\cdot\,|\theta\nu_0)$.

(from de Finetti theorem + martingale problem properties)

Outline

Fleming-Viot Process with Fertility Selection

Generalisation of Blackwell-MacQueen Urn

Consider the Dirichlet Process Mixture Model in hierarchical framework:

$$\begin{aligned} y_1, \dots, y_n \,|\, x_1, \dots, x_n &\sim \prod_{i=1}^{n} K_n(\cdot \,|\, x_i) \\ x_1, \dots, x_n \,|\, \mu &\stackrel{iid}{\sim} \mu \\ \mu &\sim \mathscr{D}ir(\theta \nu_0) \end{aligned}$$

and assume the density K_n is such that $K_n(1 \,|\, x) = \beta_n(x)$ bounded. Then we can write

$$p_n(x_1, \dots, x_n \,|\, y_1 = 1, \dots, y_n = 1) \propto p_n(x_1, \dots, x_n) \prod_{i=1}^{n} \beta_n(x_i).$$

Removing x_i, the predictive is

$$q_n(\mathrm{d}x_i|\mathbf{x}_{-i}) = \frac{\beta_n(x_i)\, p_n(\mathrm{d}x_i|\mathbf{x}_{-i})}{\int \beta_n(x_i)\, p_n(\mathrm{d}x_i|\mathbf{x}_{-i})}.$$

Since the marginal law of the x's is Pólya urn, we have

$$q_n(\mathrm{d}x_i|\mathbf{x}_{-i}) = \frac{\theta \beta_n(x_i)\, \nu_0(\mathrm{d}x_i) + \sum_{l \neq i}^{n} \beta_n(x_l)\, \delta_{x_l}(\mathrm{d}x_i)}{\theta \int \beta_n(x_i)\, \nu_0(\mathrm{d}x_i) + \sum_{l \neq i}^{n} \beta_n(x_l)}$$

Fleming-Viot Process with Fertility Selection

Definition of Particle Process

Given $(x_1, \ldots, x_n)$, define the following transition:

1. A particle x_i is selected with probability $1/n$
2. A random holding time is sampled from

$$\mathrm{Exp}\left[\frac{1}{2}\,n\left(\theta\int\beta_n(x_i)\,\nu_0(\mathrm{d}x_i)+\sum_{k\neq i}\beta_n(x_k)\right)\right]$$

3. At the next renewal, x_i is replaced with a sample from

$$q_n(\mathrm{d}x_i'|\mathbf{x}_{-i})=\frac{\theta\beta_n(x_i')\,\nu_0(\mathrm{d}x_i')+\sum_{l\neq i}^n\beta_n(x_l)\,\delta_{x_l}(\mathrm{d}x_i')}{\theta\int\beta_n(x_i')\,\nu_0(\mathrm{d}x_i')+\sum_{l\neq i}^n\beta_n(x_l)}$$

4. The other components remain constant.

Note that:

- the process is still Markov
- β_n acts as fertility selection.

Fleming-Viot process with Fertility Selection

Convergence of Measure-Valued Process

Take the process of empirical measures

$$\mu_n^\sigma(t) = \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i(t)}$$

and choose, for $\sigma \in B(\mathcal{X})$,

$$\beta_n(x) = 1 + \frac{2}{n}\sigma(x).$$

The convergence of generators

$$\sup_{x^n \in \mathcal{X}^n} |\langle A^n f, \mu_n^{\sigma^{(m)}}\rangle - \mathbb{A}^\sigma \varphi(\mu_n)| \longrightarrow 0$$

implies the weak convergence to the F-V process with fertility selection

$$\{\mu_n^\sigma(t),\ t \geq 0\} \Rightarrow \{\mu^\sigma(t),\ t \geq 0\}.$$

in the Skorohod topology of $D_{\mathscr{P}(\mathcal{X})}[0,\infty)$.

Extension to Diploid Fertility Selection

Construction

An analogous construction can be done for the bivariate selection case, with the Poisson rate

$$\lambda_{n,i} = \frac{1}{2}\,n\left(\theta \int \sum_{j\neq i}^{n} \beta_n(x_i,x_j)\,\nu_0(\mathrm{d}x_i) + \sum_{k\neq i}^{n}\sum_{j\neq i}^{n} \beta_n(x_k,x_j)\right)$$

and the generalised Blackwell-MacQueen predictive

$$q_n(\mathrm{d}x_i|\mathbf{x}_{-i}) = \frac{\theta_n\,\nu_n(\mathrm{d}x_i) + \sum_{k\neq i}^{n}\sum_{j\neq i}^{n}\beta_n(x_i,x_j)\,\delta_{x_k}(\mathrm{d}x_i)}{\theta_n + \sum_{k\neq i}^{n}\sum_{j\neq i}^{n}\beta_n(x_k,x_j)}$$

where

$$\theta_n = \theta \int \sum_{j\neq i}^{n} \beta_n(x_i,x_j)\,\nu_0(\mathrm{d}x_i) \qquad \text{and} \qquad \nu_n(\mathrm{d}x_i) = \frac{\sum_{j\neq i}^{n}\beta_n(x_i,x_j)\,\nu_0(\mathrm{d}x_i)}{\int \sum_{j\neq i}^{n}\beta_n(x_i,x_j)\,\nu_0(\mathrm{d}x_i)}.$$

Extension to Diploid Fertility Selection

Convergence of Measure-Valued Process

Take again

$$\mu_n^{\sigma^2}(t) = \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i(t)}$$

and choose, for $\sigma \in B_{\text{sym}}(\mathcal{X}^2)$,

$$\beta_n(x,y) = \frac{1}{n}\left(1 + \frac{2}{n}\sigma(x,y)\right).$$

The convergence of the generators implies the weak convergence to the F-V process with diploid fertility selection

$$\{\mu_n^{\sigma^2}(t),\ t \geq 0\} \Rightarrow \{\mu^{\sigma^2}(t),\ t \geq 0\}.$$

in $D_{\mathscr{P}(\mathcal{X})}[0,\infty)$.

Outline

Fleming-Viot process with Viability Selection

The Two-Parameter Poisson-Dirichlet Process

Denoting with $\mathscr{PD}(\sigma, \theta)$ the two-parameter Poisson-Dirichlet process, if

$$x_1, \dots, x_n \mid \mu \overset{iid}{\sim} \mu$$
$$\mu \sim \mathscr{PD}(\sigma, \theta) \qquad \sigma \in [0, 1), \quad \theta > -\sigma$$

then the following characterisation holds (Pitman, 1996)

$$x_n | x_1, \dots, x_{n-1} \sim \frac{\theta + \sigma k}{\theta + n - 1}\, \nu_0 + \frac{1}{\theta + n - 1} \sum_{j=1}^{k} (n_j - \sigma)\delta_{x_j^*}.$$

where k = number of unique values x_j^*, and n_j = multeplicity of x_j^*.

Fertility Transition

Markov Attempt

Constructing a Markov jump process as for *Fertility Selection* does not yield a well-behaved process in the limit for n.

Fleming-Viot Process with Viability Selection

Definition of Particle Process

Define a semi-Markov jump process as follows:

1. The outgoing particle x_i is selected with probability

$$\frac{1}{n}\left(1+\frac{2\,\sigma\,\alpha(x_i)}{n}\right)$$

2. The incoming particle x_i' is sampled from the Pitman urn predictive

$$\frac{\theta+\sigma k}{\theta+n-1}\;\nu_0(\mathrm{d}x_i')+\frac{1}{\theta+n-1}\sum_{j=1}^{k}(n_j-\sigma)\delta_{x_j^*}(\mathrm{d}x_i')$$

3. A holding time is sampled from

$$\mathrm{Exp}\left(\left(\frac{\theta}{\theta+\sigma k}\right)^{-1}\frac{n(\theta+n-1)}{2}\right)\qquad \text{if}\qquad x_i' \text{ is from } \nu_0$$

$$\mathrm{Exp}\left(\left(1-\frac{\sigma}{n_j}\right)^{-1}\frac{n(\theta+n-1)}{2}\right)\qquad \text{if}\qquad x_i' = x_j^*$$

4. The other components remain constant.

Fertility Transition

Fleming-Viot Process with Viability Selection

Convergence and Stationary Distribution

Define

$$\mu_n^*(t) = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i(t)}.$$

- The convergence of the generators implies the weak convergence to the F-V process with viability selection

  $$\{\mu_n^*(t),\, t \geq 0\} \Rightarrow \{\mu^*(t),\, t \geq 0\}.$$

  in $D_{\mathscr{P}(\mathcal{X})}[0, \infty)$.

- The stationary distribution of the F-V process with viability selection is the two-parameter Poisson-Dirichlet process $\mathscr{PD}(\sigma, \theta)$.

  (de Finetti theorem + martingale problem properties)

Summary

Particle process	Transition density	Fleming-Viot process	Stationary distribution
Markov	Blackwell-MacQueen Urn	Neutral	$\mathscr{D}(\theta\nu_0)$ (known)
Markov	Generalisation 1 of Blackwell-MacQueen	Fertility Selection	$Ce^{\int \sigma d\mu}\mathscr{D}(\theta\nu_0)$ (known)
Markov	Generalisation 2 of Blackwell-MacQueen	Diploid Fertility Selection	$Ce^{\int\int \tilde{\sigma} d\mu^2}\mathscr{D}(\theta\nu_0)$ (known)
Semi-Markov	Pitman Urn	Viability Selection	$\mathscr{P}\mathscr{D}(\sigma, \theta)$

References

Ethier S.N. and Griffiths R.C. (1993)
The transition function of a Fleming-Viot process.
Ann. Probab. 21.

Donnelly P. and Kurtz T.G. (1996)
A countable representation of the Fleming-Viot measure-valued diffusion.
Ann. Probab. 24.

Donnelly P. and Kurtz T.G. (1999)
Genealogical processes for Fleming-Viot models with selection and recombination.
Ann. Appl. Probab. 9.

Ethier S.N. and Kurtz T.G. (1994)
Convergence to the Fleming-Viot process in the weak atomic topology.
Stoch. Proc. Appl. 54.

Fleming W. H. and Viot M. (1979)
Some measure-valued processes in population genetics theory.
Indiana University Mathematics J. 28.

Lo A. Y. (1984)
On a class of Bayesian nonparametric estimates I: density estimates
Ann. Statist., 12

Pitman J. (1996)
Some development of the Blackwell-MacQueen urn scheme.
IMS Lecture Notes 30.

Generator of F-V with selection

$$\mathbb{A}\varphi(\mu) = \sum_{j=1}^{m} \langle B_j f, \mu^m \rangle \qquad \text{mutation}$$

$$+ \frac{1}{2} \sum_{1 \leq i \neq j \leq m} \Big(\langle \Phi_{ij} f, \mu^{m-1} \rangle - \langle f, \mu^m \rangle \Big) \qquad \text{genetic drift}$$

$$+ \sum_{j=1}^{m} \Big(\langle \sigma_{j\cdot}(\cdot,\cdot) f, \mu^{m+1} \rangle - \langle \sigma(\cdot,\cdot) \otimes f, \mu^{m+2} \rangle \Big) \qquad \text{selection}$$

where:

- $\langle f, \mu \rangle = \int_{\mathcal{X}} f \mathrm{d}\mu$
- $Bg(x) = \frac{1}{2}\theta \int_{\mathcal{X}} \big[g(y) - g(x) \big] \nu_0(\mathrm{d}y)$
- $\Phi_{ij} f$ sets the i-th and j-th variable in f equal.
- $\sigma \in B_{\mathrm{sym}}(\mathcal{X}^2)$ such that

  $\sigma_{j\cdot}(\cdot,\cdot)$ denotes $\sigma(x_j, x_{m+1})$ and $\sigma(\cdot,\cdot) \otimes f$ denotes $\sigma(x_{m+1}, x_{m+2}) f(x_1, \ldots, x_m)$.