

# Canonical representations for dependent Dirichlet populations.

D. Spanò

University of Oxford

(joint work with R.C. Griffiths, Oxford)

# The two-type Wright-Fisher diffusion in Genetics.

Transition density:  $\alpha > 0, \beta > 0$ . For each  $x, y \in [0, 1]$ ,

$$p_t^{(\alpha, \beta)}(x, y) dy = \pi_{\alpha, \beta}(y) dy \left\{ 1 + \sum_{n=1}^{\infty} \rho_n^{(\alpha + \beta)}(t) P_n^{(\alpha, \beta)}(x) P_n^{(\alpha, \beta)}(y) \right\}, \quad t > 0.$$

- $\pi_{\alpha, \beta}(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)} \mathbb{I}(x \in (0, 1))$ : stationary distribution;
- $\{P_n^{(\alpha, \beta)}(x)\}$ : Jacobi polynomials, orthonormal w.r.t.  $\pi_{\alpha, \beta}$ ;
- $\rho_n^{\alpha + \beta}(t) = e^{-\frac{1}{2}tn(n + \alpha + \beta - 1)}$

Generator:  $AP_n^{(\alpha, \beta)}(x) = -\frac{1}{2}n(n + \alpha + \beta - 1)P_n^{(\alpha, \beta)}(x).$

# A classical problem.

1. (*Lancaster problem*) Consider:

$$p^{(\alpha,\beta)}(x, y) dy = \pi_{\alpha,\beta}(y) dy \left\{ 1 + \sum_{n=1}^{\infty} a_n P_n^{(\alpha,\beta)}(x) P_n^{(\alpha,\beta)}(y) \right\}.$$

For which  $a_n$  is  $p(x, y)$  a density for every  $x$ ?

With  $a_0 \equiv 1$ , every such solution implies

$$a_n P_n^{(\alpha,\beta)}(x) = E(P_n^{(\alpha,\beta)}(Y) | x), \quad n \geq 1.$$

Regression for  $L_2$  functions (Fourier expansion):

$$f(x) \sim \sum_{n=0}^{\infty} c_n P_n^{(\alpha,\beta)}(x) \Rightarrow E(f(Y) | x) \sim \sum_{n=0}^{\infty} a_n c_n P_n^{(\alpha,\beta)}(x).$$

# A classical problem.

2. (Bochner problem) Consider

$$p_t^{(\alpha,\beta)}(x, y) dy = \pi_{\alpha,\beta}(y) dy \left\{ 1 + \sum_{n=1}^{\infty} a_n(t) P_n^{(\alpha,\beta)}(x) P_n^{(\alpha,\beta)}(y) \right\} \quad t > 0.$$

For which  $a_n(t) = e^{-\Lambda_n t}$  is  $p_t(x, y)$  the transition function of a Markov Process  $X = (X_t : t \geq 0)$ ?

Every such solution implies

$$a_n(t) P_n^{(\alpha,\beta)}(x) = E(P_n^{(\alpha,\beta)}(X_t) \mid X_0 = x).$$

Semigroup for  $L_2$  functions (Fourier expansion):

$$f(x) \sim \sum_{n=0}^{\infty} c_n P_n^{(\alpha,\beta)}(x) \Rightarrow P_t f(x) := E(f(X_t) \mid X_0 = x) \sim \sum_{n=0}^{\infty} a_n(t) c_n P_n^{(\alpha,\beta)}(x).$$

(Remember, generator:  $Af = \frac{d}{dt} P_t f$ ).

## A (less) classical problem.

3. Solve Lancaster (and Bochner) problem for Dirichlet measures on  $d \leq \infty$  points.

$$\alpha = (\alpha_1, \dots, \alpha_d) \in \mathbb{R}_+^d, \quad |\alpha| := \sum_{i=1}^d \alpha_i ;$$
$$\pi_\alpha = \text{Dirichlet on } \Delta_{(d-1)} := \{x \in [0, 1]^{d-1} : |x| \leq 1\}.$$

$$p_t^{(\alpha, \beta)}(x, dy) = \pi_\alpha(dy) \left\{ 1 + \sum_{|n|=1}^{\infty} \sum_{m \in \mathbb{N}^d: |m|=|n|} a_m(t) P_m^{(\alpha)}(x) P_m^{(\alpha)}(y) \right\} \quad t > 0.$$

$\{P_m(x)\}_{m \in \mathbb{N}^d} = \text{multivariate OP's w.r.t. } \pi_\alpha$

(for  $d = \infty$ ,  $\alpha = \text{measure on } \mathbb{R}$  and  $\pi_\alpha = \text{PD}(|\alpha|)$  or  $\text{GEM}(|\alpha|)$  or  $FD(\alpha)$ ).

Bochner (1954) answers to problems 1 and 2 for  $\alpha = \beta > 1/2$ . Gasper (1974) generalizes to  $\alpha < \beta$  with  $1/2 \leq \alpha$  or  $\alpha + \beta \geq 2$ . No answer for more general  $\alpha, \beta$  (only sufficient conditions)!!

(i) *N+S condition for  $p(x, y)$  to be a conditional density  $\forall x$  is that for some positive ( $\sigma$ -additive) measure  $H$ ,*

$$a_n = \int_0^1 R_n^{(\alpha, \beta)}(x) H(dx)$$

where  $R_n^{(\alpha, \beta)}(x) := P_n^{(\alpha, \beta)}(x) / P_n^{(\alpha, \beta)}(1)$ .

(ii) *N+S condition for  $p_t(x, y)$  to be a transition density is that  $a_n = e^{-\frac{1}{2}\Lambda_n t}$  with*

$$\Lambda_n = \sigma n(n + \alpha + \beta + 1) + \int_0^{1-0} \frac{1 - R_n^{\alpha, \beta}(x)}{1 - x} dH(x)$$

Key property for the proof for  $d = 2$ :

$$R_n^{(\alpha,\beta)}(x)R_n^{(\alpha,\beta)}(y) = \int_0^1 R_n^{(\alpha,\beta)}(z)m_{\alpha,\beta}(z)dz,$$

for a nonnegative measure  $m_{\alpha,\beta} \ll \pi_{\alpha,\beta}$  (Koornwinder 1972, Gasper 1973).

This guarantees hypergroup structure hence convolution.

For  $d > 2$ , Koornwinder and Schwartz (1991): Product formula for one choice of multivariate Jacobi  $\{P_m^\alpha\}_{m \in \mathbb{N}^d}$ , ( $\alpha \in \mathbb{R}^d$ ) with mixing measure  $m_\alpha$  explicitly described. BUT:

- *Multivariate OP are not unique!*
- *K+S product formula does not give N+S conditions;*
- *K+S product formula depends heavily on dimension  $d$  !!!*

# Polynomial kernels

Alternative approach for  $d \geq 2$  (Griffiths and S, 2007): work with

$$Q_{|n|}^\alpha(x, y) = \sum_{|m|=|n|} P_m^\alpha(x) P_m^\alpha(y).$$

Fourier expansion analogue:  $f(x) \sim \sum_{|n|} \mathbb{E}(f(Y) Q_{|n|}^\alpha(x, Y))$ .

- $Q_{|n|}^\alpha(x, y)$  *unique!*
- *Lead to N+S condition for all  $a_m = a_{|m|}$  for the positivity of*

$$p(x, dy) = \pi_\alpha(dy) \left\{ 1 + \sum_{|n| \geq 1} a_{|n|} \sum_{|m|=|n|} Q_{|n|}^\alpha(x, y) \right\};$$

- *Characterization independent of  $d \rightarrow$  possible extension to  $d = \infty$  (measure-valued processes);*
- *Explicit description leads to probabilistic interpretation (cf. Walker et al. 2006).*



# Polynomial kernels.

**Proposition** (Griffiths and S, 2007).

$$Q_{|n|}^{\alpha}(x, y) = (|\alpha| + 2|n| - 1) \sum_{|m|=0}^{|n|} (-1)^{|n|-m} \frac{(|\alpha| + m)_{(|n|-1)}}{m!(|n| - m)!} \xi_{|m|}^{\alpha}(x, y),$$

where

$$\xi_{|m|}^{\alpha}(x, y) = \sum_{|l|=|m|} \binom{|m|}{l} \frac{(|\alpha|)_{(|m|)}}{\prod_1^d (\alpha_i)_{(l_i)}} \prod_1^d (x_i y_i)^{l_i}$$

with  $\binom{|m|}{l} = |m|! / (l_1! \cdots l_d!)$ .

$$\Rightarrow \xi_{|m|}^{\alpha}(x, y) \pi_{\alpha}(dy) = \sum_{|l|=|m|} Mn(l|x) \pi_{\alpha+l}(dy) = \mathbb{E} (\pi_{\alpha+L}(dy) \mid X = x, |L| = |m|)$$

→ Walker and Muliere (2003) Bivariate DP as  $d \rightarrow \infty$ .

# Product formula and Lancaster problem.

Remember:  $R_{|n|}^{(\alpha,\beta)}(x)R_{|n|}^{(\alpha,\beta)}(y) = \int_0^1 R_{|n|}^{(\alpha,\beta)}(z)m_{\alpha,\beta}(z)dz.$

**Proposition** (Griffiths and S, 2007). For every  $d \geq 2$ , let  $\alpha \in \mathbb{R}_+^d$  be such that, for every  $j = 1, \dots, d$ ,  $\alpha_j \leq \sum_{i=1}^{j-1} \alpha_i$  and  $1/2 \leq \alpha_j$ , or  $\sum_{i=1}^j \alpha_i \geq 2$ .

$$Q_{|n|}^\alpha(x, y) = h_{|n|}^{\alpha_d, |\alpha| - \alpha_d} \int R_{|n|}^{\alpha_d, |\alpha| - \alpha_d}(z) m_{x,y;\alpha}(dz)$$

for some positive measure  $m_{x,y,\alpha}$  on  $[0, 1]$  ( $h_{|n|}^{(\alpha,\beta)}$  normaliz. constant).

**Corollary.** Same constraints on  $\alpha$ . A sequence  $\{a_{|n|} : |n| \in \mathbb{N}\}$  solve Lancaster's problem for the Dirichlet( $\alpha$ ) distribution if and only if, for at least a subset  $I$  of  $\{1, \dots, d\}$ ,  $a_{|n|}$  is a solution for the Beta( $\alpha_I, |\alpha| - \alpha_I$ ) distribution, where

$$\alpha_I := \sum_{j \in I} \alpha_j.$$

# Bivariate Dirichlet measures.

**Remark 1.** Extension to  $d \rightarrow \infty$  possible for GEM, PD, FD process with total mass  $\theta > 2$ .

**Remark 2.** Bayesian interpretation:

$$\begin{aligned} p(x, dy) &= \sum_{|n|=0}^{\infty} a_{|n|} Q_{|n|}^{\alpha}(x, y) \pi_{\alpha}(dy) \\ &= \sum_{|m|=0}^{\infty} \mathbb{P}(|L| = |m|) \mathbb{E}(\pi_{\alpha+L}(dy) \mid X = x, |L| = |m|) \end{aligned}$$

where

$$\mathbb{P}(|L| = |m|) \propto \int_0^1 \sum_{|l|=0}^{\infty} \frac{(|\alpha| + 2|l + m| - 1)(|\alpha| + 2|m|)_{(|l|)} (-1)^{|l|}}{|l|!} R_{|m+l|}^{(\alpha_d, |\alpha| - \alpha_d)}(z) H(dz).$$

for some positive measure  $H$ .

**Remark 3.** For  $d \rightarrow \infty$  solution to Bochner's problem (suitable  $H_t$ ) satisfies conditions of Walker *et al.* (2006) !!!

# Dirichlet measure-valued Markov processes.

$$\mathbb{P}(|L_t| = |m|) \propto \sum_{|l|=0}^{\infty} \frac{(|\alpha| + 2|l + m| - 1)(|\alpha| + 2|m|)_{(|l|)} (-1)^{|l|}}{|l|!} e^{-t\Lambda_{|m|}}.$$

$$\Lambda_{|m|} = \sigma|m|(|m| + |\alpha| - 1) - \int_0^{1-} \frac{1 - R_{|m|}^{(\alpha_d, |\alpha| - \alpha_d)}(z)}{1 - z} H(dz)$$

Examples:

1.  $\Lambda_{|m|} = 2^{-1}|m|(|m| + |\alpha| - 1)$ : Kingman's binary coalescent.
2.  $\Lambda_{|m|}^* = |m|$ : coalescent with simultaneous binary collisions.

**Proposition.** (Griffiths and S. 2007).

$$(X_{\Lambda^*}(t) : t \geq 0) = (X_{\Lambda}(Z_t) : t \geq 0)$$

for a stable subordinator  $(Z_t : t \geq 0)$ , independent of  $(X_{\Lambda}(t) : t \geq 0)$ .

# The d-type Moran B&D process in Genetics.

Countable representation for Wright-Fisher diffusion.

Transition density:  $\alpha \in \mathbb{R}^d$ . For every  $m, r \in \mathbb{N}^d : |m| = |r|$ ,

$$q_t^{(\alpha, |n|)}(m, r) = M_{(\alpha, \beta, |r|)}(r) \left\{ 1 + \sum_{|n|=1}^{\infty} \rho_{|n|}^{|\alpha|}(t) h_n^{(\alpha, |r|)}(m) h_n^{(\alpha, |r|)}(r) \right\}.$$

- $M_{(\alpha, |r|)}(r) = \int_{\Delta_{(d-1)}} \binom{|r|}{r} x^r \pi_{\alpha}(dx) = \binom{|r|}{r} \frac{\prod_{i=1}^d (\alpha_i)_{(r_i)}}{(|\alpha|)_{(|r|)}}$ ;
- $h_{|n|}^{(\alpha, |m|)}(r)$ : Multivariate Hahn polynomials, Karlin-McGregor (1978);
- $\rho_{|n|}^{\alpha+\beta}(t) = e^{-\frac{1}{2}t|n|(|n|+\alpha+\beta-1)}$  same as Wright-Fisher diffusion.

# Solving Lancaster/Bochner problem for $M_{(\alpha,|r|)}$ .

**Proposition.** (Griffiths and S. 2007)

(i) *Multivariate Hahn (non-unique) are given by:*

$$h_n^{(\alpha,|m|)}(r) = \int_{\Delta_{(d-1)}} P_n^\alpha(x) \pi_{\alpha+r}(dx)$$

where  $P_{|n|}^\alpha$  are multivariate Jacobi.

(ii) *Polynomial kernel in  $M_{(\alpha,|r|)}$  uniquely determined by*

$$k_n^{(\alpha,|m|)}(m, r) = \int_{\Delta_{(d-1)}^2} Q_{|n|}^\alpha(x, y) \pi_{\alpha+m}(dx) \pi_{\alpha+r}(dy).$$

**Corollary.**  $M_{(\alpha,|r|)}$  and  $\pi_\alpha$  share the same set of solution for Bochner/Lancaster's problem.

## Current & future directions.

- Study tree-structure for other eigenvalues.
- Characterize general positive-definite multivariate sequences (extend Koornwinder's product formula).
- Kernel for Pitman-Yor, Beta-Stacy, NTR, NTL distributions and their sampling formulae.

## Bonus: Kernel for Poisson-Dirichlet point process.

$n$ -Kernel polynomials on the  $d$  unlabelled points ordered by size  $X_{(1)} > X_{(2)} > \dots > X_{(d)}$  are

$$Q_{|n|}^* = (d!)^{-1} \sum_{\pi} Q_{|n|}(\pi(x), y),$$

where  $\pi(x) = (x_{\pi(1)}, \dots, x_{\pi(d)})$ . Take limit as  $d \rightarrow \infty$ . Same structure:

$$Q_{|n|}^{*\infty} = \sum_{|m| \leq |n|} a_{|n||m|} \xi_{|m|}^{*\infty}$$

where

$$\xi_{|m|}^{*\infty}(x, y) = |\epsilon|_{(m)} \sum \frac{m! \alpha(1)! \cdots \alpha(k)! [x; \alpha] [y; \alpha]}{|\epsilon|^k [0!1!]^{\alpha(1)} \cdots [(k-1)!k!]^{\alpha(k)}}$$

and

$$[x; \alpha] = \sum x_{(i_1)}^{l_1} \cdots x_{(i_k)}^{l_k}.$$



## Bonus 2: Orthogonal polynomials in the GEM distribution.

For  $d < \infty$ ,  $\alpha > 0$ , let  $\pi_{\alpha,d}$  denote Dirichlet  $(\alpha, \alpha, \dots, \alpha)$ : Increments

$$B_j = \frac{X_j}{1 - \sum_{i=1}^{j-1} X_i}, \quad j = 1, \dots, d-1$$

are independent Beta $(\alpha, (d-j)\alpha)$ , respectively.

OP's are of the form:

$$R_n^\alpha(x) = \prod_{j=1}^{d-1} \left[ R_{n_j}^{\alpha, (d-j)\alpha + 2N_j}(B_j) \right] (1 - B_j)^{N_j}$$

where  $N_j = n_{j+1} + \dots + n_{d-1}$ .

Size-biased permutation

$$SBP \pi_{\alpha,d}(\sigma x) dx = \prod_{j=1}^d \frac{X_{\sigma(j)}}{1 - \sum_{i=1}^{j-1} X_{\sigma(i)}} \pi_{\alpha,d}(x) dx$$

The new increments  $B_j^{SBP}$  are now independent Beta $(1+\alpha, (d-j)\alpha)$ . *Same structure for OP !!* Let  $d \rightarrow \infty$  while  $d\alpha \rightarrow \theta$ . The limit is GEM $(\theta)$ .