A NOVEL SPECTRAL METHOD FOR INFERRING GENERAL DIPLOID SELECTION FROM TIME SERIES GENETIC DATA

By Matthias Steinrücken*, Anand Bhaskar* and Yun S. Song

University of California, Berkeley

SUPPLEMENTARY MATERIAL

A. Proofs. In this section, we give proofs of the results stated in Section 2 of the main text.

Proof of Proposition 1. We obtain the coefficients $b_{0,n}$ of the initial vector $\boldsymbol{b}_0$ by projecting $\rho(y)$ onto the basis functions $\big\{\pi(y)B_n(y)\big\}_{n\in\mathbb{N}_0}$ via the integral

$$b_{0,n} = \frac{1}{c_n}\int_0^1 \rho(y)\pi(y)B_n(y)\frac{1}{\pi(y)}dy, \tag{A.1}$$

with c_n given in (2.22) of the main text. Substituting the initial density for the allele frequency when selection arises, $\rho(y) = \delta(x-y)$, into (A.1) yields

$$b_{0,n} = \frac{B_n(x)}{c_n},$$

thus proving the statement of the proposition. □

Proof of Theorem 2. The spectral decomposition of the transition density function of the Wright-Fisher diffusion generator given in (2.16) can be written in matrix-vector notation as

$$p_\Theta(\tau;x,y) = \sum_{n=0}^{\infty} e^{-\lambda_n\tau}\pi(y)\frac{B_n(x)B_n(y)}{\langle B_n,B_n\rangle_\pi} = \boldsymbol{B}^T(x)\boldsymbol{D}^{-1}\exp\big\{-\boldsymbol{\Lambda}\tau\big\}\pi(y)\boldsymbol{B}(y), \tag{A.2}$$

where $\boldsymbol{B}(y)$ is the vector notation for the eigenfunctions $B_n(y)$ defined in (2.21) of the main text, $\boldsymbol{D} = \mathrm{diag}\big(c_0, c_1, \dots\big)$ is the diagonal matrix of the squared norms of the eigenfunctions $B_n(y)$ with entries given in (2.22),

*These authors contributed equally to this work.

and $\boldsymbol{\Lambda}$ is the diagonal matrix of eigenvalues of the Wright-Fisher diffusion generator $\mathcal{L}$.

Substituting (2.17), (2.18) and (A.2) into the recurrence (2.7) relating g_k and f_{k-1} yields

$$\begin{aligned}\boldsymbol{a}_k \pi(y_k)\boldsymbol{B}(y_k) &= \int_0^1 \boldsymbol{b}_{k-1}\pi(y_{k-1})\boldsymbol{B}(y_{k-1})\boldsymbol{B}^T(y_{k-1})\boldsymbol{D}^{-1}dy_{k-1}\times \\ &\qquad \exp\big\{-\boldsymbol{\Lambda}(\tau_k-\tau_{k-1})\big\}\pi(y_k)\boldsymbol{B}(y_k) \\ &= \boldsymbol{b}_{k-1}\exp\big\{-\boldsymbol{\Lambda}(\tau_k-\tau_{k-1})\big\}\pi(y_k)\boldsymbol{B}(y_k),\end{aligned}$$

where in the second equality, we used the fact that

$$\int_0^1 \pi(y)\boldsymbol{B}(y)\boldsymbol{B}^T(y)dy = \boldsymbol{D}. \tag{A.3}$$

Equation (A.3) holds because the eigenfunctions $B_n(y)$ form an orthogonal basis with respect to $\pi(y)$. This proves equation (2.23) in Theorem 2.

Letting $\boldsymbol{H}^{(\Theta)}(y) := (H_0^{(\Theta)}(y), H_1^{(\Theta)}(y), \ldots)^T$ denote the column vector of the functions $\big\{H_m^{(\Theta)}(y)\big\}_{m\in\mathbb{N}_0}$ given in (2.12), the representation of the eigenfunctions $B_n(y)$ given in (2.15) can be written in matrix-vector notation as

$$\boldsymbol{B}(y) = \boldsymbol{W}\boldsymbol{H}^{(\Theta)}(y), \tag{A.4}$$

where $\boldsymbol{W}$ is the matrix whose rows are formed from the eigenvectors of the matrix $\boldsymbol{M}$ given in (2.14). Substituting (A.4), (2.17), and (2.18) into the recurrence (2.8) relating f_k and g_k, we have

$$\begin{aligned}\boldsymbol{b}_k\pi(y_k)\boldsymbol{B}(y_k) &= \boldsymbol{a}_k\pi(y_k)\boldsymbol{B}(y_k)y_k^{d_k}(1-y_k)^{n_k-d_k} \\ &= \boldsymbol{a}_k\boldsymbol{W}\boldsymbol{H}^{(\Theta)}(y_k)\pi(y_k)y_k^{d_k}(1-y_k)^{n_k-d_k} \\ &= \boldsymbol{a}_k\boldsymbol{W}\mathrm{diag}\big(y_k^{d_k}(1-y_k)^{n_k-d_k}\big)\boldsymbol{H}^{(\Theta)}(y_k)\pi(y_k) \\ &= \boldsymbol{a}_k\boldsymbol{W}\boldsymbol{G}^{d_k}(\mathbb{1}-\boldsymbol{G})^{n_k-d_k}\boldsymbol{H}^{(\Theta)}(y_k)\pi(y_k) \\ &= \boldsymbol{a}_k\boldsymbol{W}\boldsymbol{G}^{d_k}(\mathbb{1}-\boldsymbol{G})^{n_k-d_k}\boldsymbol{W}^{-1}\boldsymbol{B}(y)\pi(y_k). \qquad \text{(A.5)}\end{aligned}$$

In (A.5), diag(y) denotes the matrix $\big(y\cdot\delta_{n,m}\big)_{n,m\in\mathbb{N}_0}$, where the Kronecker-delta $\delta_{n,m}$ is 1 if $n=m$ and 0 otherwise. Furthermore, we used the fact that the three-term recurrence relation for the Jacobi polynomials in (B.4) implies the matrix-vector identity $\mathrm{diag}(y)\boldsymbol{H}^{(\Theta)}(y) = \boldsymbol{G}\boldsymbol{H}^{(\Theta)}(y)$ with $\boldsymbol{G} = \big(G_{n,m}^{(\alpha,\beta)}\big)_{n,m\in\mathbb{N}_0}$ denoting the matrix of coefficients of the three-term recurrence. This proves (2.24) in Theorem 2.

The matrix $\boldsymbol{W}$ is an orthogonal matrix up to scaling of its rows and columns. In particular, substituting (A.4) into the orthogonality relation (A.3) for the eigenfunctions $B_n(y)$, we have

$$\begin{aligned}
\boldsymbol{D} &= \int_0^1 \pi(y)\boldsymbol{B}(y)\boldsymbol{B}^T(y)dy \\
&= \int_0^1 \pi(y)\boldsymbol{W}\boldsymbol{H}^{(\Theta)}(y){\boldsymbol{H}^{(\Theta)}}^T(y)\boldsymbol{W}^T dy \\
&= \boldsymbol{W}\left(\int_0^1 \pi(y)\boldsymbol{H}^{(\Theta)}(y){\boldsymbol{H}^{(\Theta)}}^T(y)dy\right)\boldsymbol{W}^T \\
&= \boldsymbol{W}\left(\int_0^1 \pi(y)e^{-\bar{\sigma}(y)}\boldsymbol{R}^{(\alpha,\beta)}(y){\boldsymbol{R}^{(\alpha,\beta)}}^T(y)dy\right)\boldsymbol{W}^T \\
&= \boldsymbol{W}\left(\int_0^1 y^{\alpha-1}(1-y)^{\beta-1}\boldsymbol{R}^{(\alpha,\beta)}(y){\boldsymbol{R}^{(\alpha,\beta)}}^T(y)dy\right)\boldsymbol{W}^T \\
&= \boldsymbol{W}\boldsymbol{C}\boldsymbol{W}^T, \qquad \text{(A.6)}
\end{aligned}$$

where $\boldsymbol{R}^{(\alpha,\beta)}(y) := \left(R_0^{(\alpha,\beta)}(y), R_1^{(\alpha,\beta)}(y), \dots\right)$, the fourth equality follows from definition (2.12), the fifth follows from definition (2.13), and the last follows from the definition of $\boldsymbol{C}$. Equation (A.6) implies (2.25) of Theorem 2. □

Proof of Proposition 3. Substituting the representation for the densities f_k in (2.17) into (2.9) yields

$$\begin{aligned}
\mathbb{P}_\Theta\{O_{[1:K]}\} &= \int_0^1 \sum_{n=0}^{\infty} b_{K,n}B_n(y)\pi(y)dy \\
&= \frac{1}{B_0(0)}\int_0^1 \sum_{n=0}^{\infty} b_{K,n}B_0(y)B_n(y)\pi(y)dy \\
&= \frac{c_0}{B_0(0)}b_{K,0},
\end{aligned}$$

where we have used $B_0(y) \equiv B_0(0)$ (see Section D) and the orthogonality of the eigenfunctions $B_n(y)$ with respect to $\pi(y)$. Using (2.15) along with the fact that $R_m^{(\alpha,\beta)}(0) = (-1)^m \frac{\Gamma(m+\alpha)}{\Gamma(m+1)\Gamma(\alpha)}$, we have the following expression for $B_0(0)$,

$$B_0(0) = \sum_{m=0}^{\infty}(-1)^m w_{0,m}\frac{\Gamma(m+\alpha)}{\Gamma(m+1)\Gamma(\alpha)}, \qquad \text{(A.7)}$$

which completes the proof. □

B. Jacobi polynomials. We briefly list some facts about our modified Jacobi polynomials and their relationship to the classical Jacobi polynomials [1, Chapter 22]. For $\alpha, \beta > 0$, we define the modified Jacobi polynomials $R_n^{(\alpha,\beta)}(x)$ by

$$R_n^{(\alpha,\beta)}(x) := p_n^{(\beta-1,\alpha-1)}(2x-1),$$

where $p_n^{(a,b)}(x)$ are the classical Jacobi polynomials. The polynomials $R_n^{(\alpha,\beta)}(x)$ form an orthogonal basis of the Hilbert space $L^2\big([0,1], x^{\alpha-1}(1-x)^{\beta-1}\big)$ with the weight function $x^{\alpha-1}(1-x)^{\beta-1}$. In particular

$$\text{(B.1)} \qquad \int_0^1 R_n^{(\alpha,\beta)}(x) R_m^{(\alpha,\beta)}(x) x^{\alpha-1}(1-x)^{\beta-1} dx = c_n^{(\alpha,\beta)} \delta_{n,m}$$

where $c_n^{(\alpha,\beta)}$ is given by

$$\text{(B.2)} \qquad c_n^{(\alpha,\beta)} = \frac{\Gamma(n+\alpha)\Gamma(n+\beta)}{(2n+\alpha+\beta-1)\Gamma(n+\alpha+\beta-1)\Gamma(n+1)}.$$

Further, $R_n^{(\alpha,\beta)}$ are the eigenfunctions of the neutral Wright-Fisher diffusion generator $\mathcal{L}_0$ given in (2.11). Thus

$$\mathcal{L}_0 R_n^{(\alpha,\beta)}(x) = -\lambda_n^{(\alpha,\beta)} R_n^{(\alpha,\beta)}(x),$$

where $\lambda_n^{(\alpha,\beta)}$ is the eigenvalue for the eigenfunction $R_n^{(\alpha,\beta)}$ and is given by

$$\text{(B.3)} \qquad \lambda_n^{(\alpha,\beta)} = \frac{1}{2} n(n+\alpha+\beta-1).$$

Finally, the Jacobi polynomials satisfy the three-term recurrence relation

$$\text{(B.4)} \quad x\ R_n^{(\alpha,\beta)}(x) = G_{n,n-1}^{(\alpha,\beta)} R_{n-1}^{(\alpha,\beta)}(x) + G_{n,n}^{(\alpha,\beta)} R_n^{(\alpha,\beta)}(x) + G_{n,n+1}^{(\alpha,\beta)} R_{n+1}^{(\alpha,\beta)}(x),$$

where the coefficients $G_{n,m}^{(\alpha,\beta)}$ are given by

$$\text{(B.5)} \qquad G_{n,m}^{(\alpha,\beta)} = \begin{cases} \frac{(n+\alpha-1)(n+\beta-1)}{(2n+\alpha+\beta-1)(2n+\alpha+\beta-2)}, & \text{if } m = n-1 \text{ and } n > 0, \\ \frac{1}{2} - \frac{\beta^2-\alpha^2-2(\beta-\alpha)}{2(2n+\alpha+\beta)(2n+\alpha+\beta-2)}, & \text{if } m = n \text{ and } n \geq 0, \\ \frac{(n+1)(n+\alpha+\beta-1)}{2(2n+\alpha+\beta)(2n+\alpha+\beta-1)}, & \text{if } m = n+1 \text{ and } n \geq 0, \\ 0, & \text{otherwise.} \end{cases}$$

C. Coefficients to compute the matrix $\boldsymbol{M}$. With the parameters $\Theta = (\sigma_1, \sigma_2, \alpha, \beta, \tau_0, N_e)$, the coefficients in the definition of the matrix $\boldsymbol{M}$ in equation (2.14) are given by

$$\begin{aligned}
q_0^{(\Theta)} &= \alpha\sigma_1, \\
q_1^{(\Theta)} &= -(2 + 3\alpha + \beta - 2\sigma_1)\sigma_1 + (1+\alpha)\sigma_2, \\
q_2^{(\Theta)} &= -10\sigma_1^2 - (1 + \alpha + \beta)\sigma_2 + (2 + 2\alpha + 2\beta + 4\sigma_2)\sigma_1, \\
q_3^{(\Theta)} &= 16\sigma_1^2 - 12\sigma_1\sigma_2 + 2\sigma_2^2, \\
q_4^{(\Theta)} &= -2(\sigma_2 - 2\sigma_1)^2.
\end{aligned}$$

D. Initial allele frequency density. As mentioned in Section 2, it is also possible to use other density functions for the allele frequency at the time when selection arises. For example, suppose this initial density function is the stationary distribution describing mutation-selection balance. In order to compute the coefficients in Proposition 1, we will need to compute the normalizing constant to make $\pi(y)$ a probability density function. This constant is given by

$$C_\pi = \int_0^1 \pi(y)dy = \frac{1}{\big(B_0(0)\big)^2}\int_0^1 \big(B_0(y)\big)^2 \pi(y)dy = \frac{c_0}{\big(B_0(0)\big)^2}.$$

Here we used the fact that the eigenfunction associated to the eigenvalue $\lambda_0 = 0$ is a constant function, and thus its value $B_0(y) \equiv B_0(0)$ is independent of y. This holds since it is straightforward to show that the differential operator $\mathcal{L}$ annihilates any constant C. We can then substitute $\rho(y) = C_\pi^{-1}\pi(y)$ into the projection integral (A.1) to get

$$\begin{aligned}
b_{0,n} &= \frac{1}{c_n}\int_0^1 C_\pi^{-1}\pi(y)\pi(y)B_n(y)\frac{1}{\pi(y)}dy \\
&= \frac{B_0(0)}{c_n c_0}\int_0^1 B_0(y)B_n(y)\pi(y)dy \\
&= \frac{B_0(0)}{c_0}\delta_{n,0}.
\end{aligned}$$

Thus, for this initial distribution, all $b_{0,n}$ are zero, except the coefficient for $n = 0$ is equal to $B_0(0)/c_0$. The value of $B_0(0)$ is given in (A.7).

Another initial density function for the allele frequency is the case of mutation-drift balance, which describes the stationary distribution of the allele frequency in the case of neutral evolution. In particular,

$$\pi_0(y) = \frac{1}{B(\alpha,\beta)} y^{\alpha-1}(1-y)^{\beta-1}, \tag{D.1}$$

where $B(\alpha, \beta)$ is the Beta function. Again, we obtain the coefficients $b_{0,n}$ of the initial vector of coefficients in Proposition 1 by projecting the initial density $\rho(y) = \pi_0(y)$ onto the basis functions $\{\pi(y)B_n(y)\}_{n\in\mathbb{N}_0}$. Substituting (D.1) into (A.1) and using the basis representation of the eigenfunctions $B_n(y)$ given in (2.15) yields

$$b_{0,n} = \frac{1}{c_n B(\alpha,\beta)} \int_0^1 y^{\alpha-1}(1-y)^{\beta-1} \sum_{m=0}^{\infty} w_{n,m} e^{-\bar{\sigma}(y)/2} R_m^{(\alpha,\beta)}(y)dy. \tag{D.2}$$

Further, from (2.15) we have

$$e^{-\bar{\sigma}(y)/2} = \frac{1}{B_0^-(0)} \sum_{m=0}^{\infty} w_{0,m}^- R_m^{(\alpha,\beta)}(y), \tag{D.3}$$

where $w_{0,m}^-$ denote the entries of the eigenvector $\boldsymbol{w}_0^-$ obtained from the matrix in (2.14) with σ_1 and σ_2 replaced by $-\sigma_1$ and $-\sigma_2$ respectively, $B_0^-(y)$ denotes the corresponding eigenfunction, and $\bar{\sigma}^-(y) = -\bar{\sigma}(y)$. Substituting (D.3) into (D.2) and using the orthogonality of the Jacobi polynomials given in (B.1) yields

$$b_{0,n} = \frac{1}{c_n B(\alpha,\beta) B_0^-(0)} \sum_{m=0}^{\infty} w_{n,m} w_{0,m}^- c_m^{(\alpha,\beta)}.$$

References.

[1] M. Abramowitz and I. A. Stegun, editors. *Handbook of Mathematical Functions.* Dover Publications, 1965.

M. Steinrücken
Department of Statistics and
Computer Science Division
University of California, Berkeley
Berkeley, CA 94720
USA
E-mail: steinrue@stat.berkeley.edu

A. Bhaskar
Computer Science Division
University of California, Berkeley
Berkeley, CA 94720
USA
E-mail: bhaskar@eecs.berkeley.edu

Y. S. Song
Department of Statistics,
Computer Science Division, and
Department of Integrative Biology
University of California, Berkeley
Berkeley, CA 94720
USA
E-mail: yss@stat.berkeley.edu