Supplementary Materials: Bézier interpolation improves the inference of dynamical models from data

Kai Shimagaki[1,*]
[1] *Department of Physics and Astronomy, University of California, Riverside, USA.*

John P. Barton[1,2,†]
[2] *Department of Computational and Systems Biology, University of Pittsburgh School of Medicine, USA.*

CONTENTS

Methods

Data and code

Raw data and code used in our analysis is available in the GitHub repository located at https://github.com/bartonlab/paper-Bezier-interpolation. This repository also contains Jupyter notebooks that can be run to reproduce the results presented here.

Optimization of control points for Bézier curves

For simplicity, we will discuss a one-dimensional case, but the following discussion can easily be extended to arbitrary dimensions. The control points of Bézier curves are obtained by solving an optimization problem that is derived from properties we want the Bézier curve to satisfy. In this study, we impose the C^2 smoothness condition, which is that up to the second derivative of the curve exist. Formally, we can represent these conditions as follows,

$$\partial_\tau x_B^{(k-1)}(\tau=1) = \partial_\tau x_B^{(k)}(\tau=0) \ , \tag{1}$$

and,

$$\partial_\tau^2 x_B^{(k-1)}(\tau=1) = \partial_\tau^2 x_B^{(k)}(\tau=0) \ , \tag{2}$$

Where, $x_B^{(k)}(\tau)$ is the interpolated function between successive discrete time points t_k and t_{k+1} and defined in the main section. Since these constraints are defined at each junction of adjacent segments, the number of conditions is $2(K-1)$. On the other hand, the number of control points is $2K$, so we will introduce two more constraints to make the problem solvable:

$$\begin{aligned} \partial_\tau^2 x_B^{(0)}(\tau=0) &= 0 \\ \partial_\tau^2 x_B^{(K-1)}(\tau=1) &= 0 \end{aligned} \ .$$

By rearranging (1) and (2), we can reduce them to the following difference equations.

$$\phi_1^{(k)} - 2x^{(k)} = \phi_2^{(k-1)} \ , \tag{3}$$

and

$$-2\phi_1^{(k)} + \phi_2^{(k)} = \phi_1^{(k-1)} - 2\phi_2^{(k-1)} \ .$$

Also, the additional boundary constraints lead to

$$\begin{aligned} x^{(0)} - 2\phi_1^{(0)} + \phi_2^{(0)} &= 0 \ , \\ \phi_1^{(k-1)} - 2\phi_2^{(k-1)} + x^{(k)} &= 0 \end{aligned} \ .$$

These difference equations are summarized as the following single linear equation by assuming that $\{\phi_2^{(k)}\}_{k=0}^{K-1}$

is a function of $\{\phi_1^{(k)}, x^{(k)}\}_{k=0}^{K-1}$, then by marginalizing $\{\phi_2^{(k)}\}_{k=0}^{K-1}$ from the difference equations,

$$\mathrm{M}^{\mathrm{Bez,K}}\boldsymbol{\phi}_1 = \boldsymbol{\psi}((x^{(k)})_{k=0}^{K+1}) \ , \tag{4}$$

where $\boldsymbol{\phi}_1 = (\phi_1^{(0)}, \ldots, \phi_1^{(K)})^T$, and let

$$\boldsymbol{\psi}((x^{(k)})_{k=0}^{K+1}) = \begin{pmatrix} x^{(0)} + 2x^{(1)} \\ 2(2x^{(1)} + x^{(2)}) \\ \vdots \\ 2(2x^{(K-1)} + x^{(K)}) \\ 8x^{(K)} + x^{(K+1)} \end{pmatrix} \ , \tag{5}$$

and the matrix $\mathrm{M}_B^{(K)}$ is defined as

$$\mathrm{M}_B^{(K)} = \begin{pmatrix} 2 & 1 & 0 & \ldots & \ldots & \ldots & \ldots & 0 \\ 1 & 4 & 1 & 0 & \ldots & \ldots & \ldots & \vdots \\ 0 & 1 & 4 & 1 & 0 & \ldots & \ldots & \vdots \\ \vdots & \ldots & \ddots & \ddots & \ddots & \ldots & \ldots & \vdots \\ \vdots & \ldots & \ldots & \ddots & \ddots & \ddots & \ldots & \vdots \\ \vdots & \ldots & \ldots & 0 & 1 & 4 & 1 & 0 \\ \vdots & \ldots & \ldots & \ldots & 0 & 1 & 4 & 1 \\ 0 & \ldots & \ldots & \ldots & \ldots & 0 & 2 & 7 \end{pmatrix} \ . \tag{6}$$

By solving (4), we get a set of control points, hence we get a Bézier curve. Interestingly, instead of the C^2 smoothness constraint, assuming a C^1 smoothness condition and imposing a constraint that minimizes the Euclidean distance of the total trajectory leads to almost the same linear equation in (4) depending on (6) and (5).

For multivariate frequencies, the Bézier curve can be obtained by solving each linear equation individually. Practically, the control points are obtained by operating the inverse of $\mathrm{M}_B^{(K)}$ to $\boldsymbol{\psi}((x^{(k)})_{k=0}^{K+1})$ vectors on each site $i \in \{1, \ldots, L\}$. Thus, we can efficiently perform the operation and its computational time is fast. Also, the above arguments are held for the arbitrary $q > 1$ dimension case, which is relevant, for example, when considering the frequency of individuals with multiple possible nucleotides or amino acids at each site in a genetic sequence. Replacing scalar variables with vector variables leads to exactly the same linear equation in (4).

Integrated frequency and covariance using Bézier interpolation

In this section, we will show explicit representations of the integrated mutant frequencies and covariances from the WF model using Bézier interpolation.

To derive it, we apply the following useful properties of P-th order Bernstein basis ($P = 3$ for quadratic Bézier interpolation), for $\forall n \in \{0, 1, \ldots, P\}$,

$$m_n^{(P)} := \int_0^1 \beta_n(\tau)\mathrm{d}\tau = \frac{1}{P+1} \ , \tag{7}$$

and, for $\forall n, m \in \{0, 1, \ldots, P\}$,

$$\mathrm{Q}_{nm}^{(P)} := \int_0^1 \beta_n(\tau)\beta_m(\tau)\mathrm{d}\tau = \frac{\binom{P}{n}\binom{P}{m}}{(2P+1)\binom{2P}{n+m}} \ . \tag{8}$$

More general properties of the Bernstein basis can be found in refs. [1, 2].

First, we will get the integrated single mutant frequency at site i, which is shown below,

$$\begin{aligned} \Delta x_{B,i}^{(k)} &:= \Delta t_k \int_0^1 x_{B,i}^{(k)}(\tau)\mathrm{d}\tau \\ &= \Delta t_k \sum_{n=0}^{P} \left(\int_0^1 \beta_n(\tau)\mathrm{d}\tau\right) \phi_{i,n}^{(k)} \\ &= \Delta t_k \sum_{n=0}^{P} m_n^{(P)}\phi_{i,n}^{(k)} \\ &= \frac{1}{(P+1)} \sum_{n=0}^{P} \phi_{i,n}^{(k)} \ , \end{aligned} \tag{9}$$

we used the property of Bernstein in (7).

Next, we will get the integrated covariance for different sites at i and j,

$$\Delta C_{ij}^{(k)} := \Delta t_k \int_0^1 \left(x_{B,ij}^{(k)}(\tau) - \ x_{B,i}^{(k)}(\tau)x_{B,j}^{(k)}(\tau)\right)\mathrm{d}\tau \ , \tag{10}$$

the first term in (10) is the same as in (9) but we replaced a single interpolated mutant frequency by a matrix that contains the entire interpolated pairwise mutant frequencies as its elements.

The second term of the covariance in (10) is also straightforward,

$$\begin{aligned} &\int_0^1 x_{B,i}^{(k)}(\tau)x_{B,j}^{(k)}(\tau)\mathrm{d}\tau \\ &= \int_0^1 \left(\sum_{n=0}^{P} \beta_n(\tau)\phi_{i,n}^{(k)}\right)\left(\sum_{m=0}^{P} \beta_m(\tau)\phi_{j,m}^{(k)}\right)\mathrm{d}\tau \\ &= \sum_{n=0}^{P}\sum_{m=0}^{P} \left(\int_0^1 \beta_n(\tau)\beta_m(\tau)\mathrm{d}\tau\right)\phi_{i,n}^{(k)}\phi_{j,m}^{(k)} \\ &= \sum_{n=0}^{P}\sum_{m=0}^{P} Q_{nm}^{(P)}\phi_{i,n}^{(k)}\phi_{i,m}^{(k)} \ . \end{aligned}$$

Here we used the property of Bernstein (8) in the last equality.

In the case of the $P = 3$, which is the cubic Bézier, $\mathrm{Q}^{(3)}$ matrix will be

$$\mathrm{Q}^{(3)} = \begin{pmatrix} \alpha & \beta & \gamma & \delta \\ \beta & \gamma & \delta & \gamma \\ \gamma & \delta & \gamma & \beta \\ \delta & \gamma & \beta & \alpha \end{pmatrix} \ ,$$

where $\alpha = 1/7, \beta = 1/14, \gamma = 1/35, \delta = 1/140$.

Normalization of probabilities

We will show that the interpolation of probability trajectories using the Bézier interpolation is always normalized. We refer to this property as normalizability, hereafter.

First, we will discuss the normalizability of the interpolated probability distribution for a categorical distribution depending on an arbitrary number of states $q > 1$. Next, we denote a probability distribution depending on the data points k and index i as $\boldsymbol{x}_i^{(k)} = (x_{i,1}^{(k)}, \dots, x_{i,q}^{(k)})^T$, and a sum of the all states is normalized, that is $\sum_{a=1}^q x_{i,a}^{(k)} = 1$ for all k, i.

Then, we can prove that when probability distributions are interpolated using Bézier's method, any interpolated function $\boldsymbol{x}_{B,i}(k) = (x_{B,i,1}^{(k)}, \dots, (x_{B,i,q}^{(k)})^\top$ is also normalized in arbitrary point $\tau \in [0,1]$:

$$\sum_{a=1}^q x_{B,i,a}^{(k)} = 1 \ .$$

For the sake of simplicity, we will omit the site index hereafter, that is $x_{i,a}^{(k)} \to x_a^{(k)}$. To see the proof, we will start by showing the normalizability of the control points $\sum_{a=1}^q \phi_{1,a}^{(k)} = 1, \forall k$ because this condition immediately leads to $\sum_{a=1}^q \phi_{2,a}^{(k)} = 1$ by plugging it into the (3), and the following part is straightforward as shown below,

$$\begin{aligned}
\phi_{2,a}^{(k-1)} &= 2x_a^{(k)} - \phi_{1,a}^{(k)} \\
&= 2(1 - \sum_{b=1|\neq a}^q x_b^{(k)}) - (1 - \sum_{b=1|\neq a}^q \phi_{1,b}^{(k)}) \\
&= 1 - (2 \sum_{b=1|\neq a}^q x_b^{(k)} - \sum_{b=1|\neq a}^q \phi_{1,b}^{(k)}) \\
&= 1 - \sum_{b=1|\neq a}^q \phi_{2,b}^{(k-1)} \ ,
\end{aligned}$$

so $\sum_{a=1}^q \phi_{2,a}^{(k)} = 1$ and it is normalized when $\boldsymbol{\phi}_1^{(k)}$ is normalized for $k \in \{1, \dots, K-1\}$. In the case of boundaries, time points at $k = 0, K$, exactly the same argument holds, which is almost trivial, so we omit to repeat the same kind of proof.

Therefore, we will show the normalizability of $\boldsymbol{\phi}_1^{(k)}$ as follows. First, we consider a sum of all the states on the left hand side in (5),

$$l.h.s. = \mathrm{M}_B^K \begin{pmatrix} \sum_{a=1}^q \phi_{1,a}^{(0)} \\ \vdots \\ \sum_{a=1}^q \phi_{1,a}^{(K)} \end{pmatrix} \ .$$

Next, we also perform a sum of all the states on the right hand side in (5),

$$r.h.s. = \sum_{a=1}^q \begin{pmatrix} x^{(0)} + 2x^{(1)} \\ 2(2x^{(1)} + x^{(2)}) \\ \vdots \\ 2(2x^{(K-1)} + x^{(K)}) \\ 8x^{(K)} + x^{(K+1)} \end{pmatrix} = \begin{pmatrix} 3 \\ 6 \\ \vdots \\ 6 \\ 9 \end{pmatrix} \ .$$

Then, we immediately notice that

$$\mathrm{M}_B^K \mathbf{1} = (3, 6, \dots, 6, 9)^\top \ .$$

Therefore, we find the normalization of the control points $\sum_{a=1}^q \phi_{1,a}^{(k)} = 1, \ \forall k \in \{0, 1, \dots, K\}$.

Finally, we sum the interpolated function using Bézier's method at arbitrary τ while considering the normalizability conditions for the control points we have seen earlier. A sum of the interpolated functions for the all states $a \in \{1, \dots, q\}$ at any position $\tau \in [0,1]$ is:

$$\sum_{a=1}^q x_{B,i,a}^{(k)} = \sum_{l=0}^P \beta_l(\tau) = 1 \ ,$$

for the first equality, we used the fact that all the control points are normalized. For the second equality, we used the nature of the Bernstein polynomial, a sum of all the Bernstein bases is one.

Treatment for negative interpolated frequencies and negative eigenvalues in real data

The sum of q categorical variables using Bézier interpolation is conserved, guaranteeing the conservation of probability density. However, interpolated probabilities can occasionally exceed the boundaries at 0 and 1, and eigenvalues of the integrated covariance matrix can become negative. This issue can occur when frequency trajectories are close to the boundaries, variables take one of the multiple possible states ($q > 0$), and sampling points are heterogeneously and sparsely distributed.

To alleviate this problem, we employed the following treatment: if the time interval $\Delta t_k = t_{k+1} - t_k$ is greater than a threshold value (set to 50 days for the analysis of HIV-1 sequence data), then we insert mean frequency points at the middle time points $(t_{k+1} + t_k)/2$ such that $\boldsymbol{x}(t_k) + \boldsymbol{x}(t_{k+1})/2$. In addition, for each frequency individually, we insert mean frequency points at middle time points when the frequency changes sharply within one time interval (more than 70% change in the case of HIV-1 data).

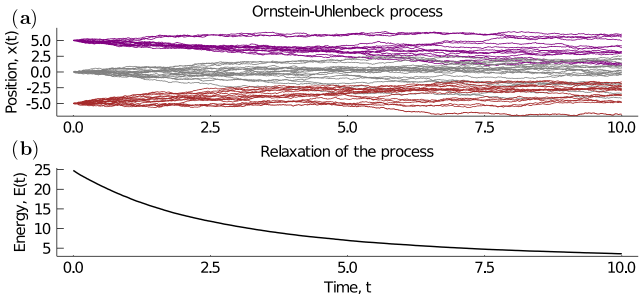

FIG. 1. **Typical Ornstein-Uhlenbeck dynamics. (a)** We generated trajectories using the Euler-Maruyama (EM) scheme 5000 times with a small increment time step of $dt = 10^{-3}$. Each trajectory shows evolution of one of the elements of a multivariate variable. **(b)** Evolution of the average effective energy of the OU process. The process relaxed from initial states randomly chosen from $\{-5, 0, 5\}^L$ to low energy states. We obtained the energy function, $E(t) = -\boldsymbol{x}(t)^\top \mathrm{J}\boldsymbol{x}(t)/2$ by running EM simulations 100 times and averaging the results.

Generate data sets from Ornstein Uhlenbeck processes.

To generate test data, we simulated the OU process using negative definite interaction matrices parameterized as $\mathrm{J} = -\frac{\alpha}{\sqrt{P}} \sum_{\nu=1}^{P} \boldsymbol{\xi}_\nu \boldsymbol{\xi}_\nu^\top$, where $\boldsymbol{\xi}_\nu$ is a pattern generated from the multivariate normal distribution, $\boldsymbol{\xi}_\nu \sim \mathcal{N}(0,1)^L$, $\alpha = \mathcal{O}(1/L)$ is a small parameter, and P is the number of embedded patterns. This construction ensures that the OU process does not diverge. We used the Euler-Maruyama (EM) scheme [3] to simulate the OU process defined in the main text (**Fig. 1a**). We simulated 1000 trajectories each for 10 randomly generated interaction matrices, as described above. We chose $L = 50$, and $\alpha = 1/L = 0.02$ in our simulations. For inference, we sampled data from the simulations every $\Delta t = 1.0$ units of time.

Maximum path-likelihood estimation for the Ornstein-Uhlenbeck process

Based on the stochastic differential equation (STD) of OU process defined in the main text, we can get the following Fokker-Planck equation [4], which is characterized by the drift and diffusion terms,

$$\frac{\partial}{\partial t} p(\boldsymbol{x}(t), t) = \mathcal{L}\; p(\boldsymbol{x}(t), t)$$
$$\mathcal{L} = -\sum_{i=1}^{L}\sum_{j=1}^{i} J_{ij} x_j \frac{\partial}{\partial x_i} + \sum_{i,j=1}^{L} \Sigma_{ij} \frac{\partial^2}{\partial x_i \partial x_j}\,. \tag{11}$$

The first term corresponds to the drift due to the pairwise interaction, and the second term corresponds to the diffusion due to the white noise.

The FP equation in (11) is effectively a diffusion equation for probability measures, and the general solution of the diffusion equation characterized by the drift and diffusion terms is known and defined as a transition probability between time points t_k and $t_{k+1} = t_k + \Delta t_k$,

$$\begin{aligned} &p\left(\boldsymbol{x}(t_{k+1}), t_{k+1} \mid \boldsymbol{x}(t_k), t_k\right) \\ &= \frac{1}{\sqrt{\|2\pi\Sigma\|\Delta t_k}} \exp\Big(-\frac{1}{2\Delta t_k}\left(\Delta\boldsymbol{x}(t_k) - \Delta t_k \mathrm{J}\boldsymbol{x}(t_k)\right)^\top \\ &\qquad \times \Sigma_t^{-1}\left(\Delta\boldsymbol{x}(t_k) - \Delta t_k \mathrm{J}\boldsymbol{x}(t_k)\right)\Big)\,, \end{aligned}$$

where $\Delta\boldsymbol{x}(t_k) = \boldsymbol{x}(t_{k+1}) - \boldsymbol{x}(t_k)$. The solution of the FP equation tells that as the time interval approaches zero, the transition probability goes to the *Kronecker delta* like distribution having a finite probability density around the previous time step. As the time interval increase, the variance increase as the square root of time, which is the nature of Brownian diffusion.

The likelihood path function for the OU model can be defined as a product of the transition probability because of the independence of the increments of the Wiener processes. Hence the log path-likelihood can be written as

$$\begin{aligned} &\mathcal{S}(\mathrm{J}|\Gamma((\boldsymbol{x}(t_k))_{k=0}^{K-1})) \\ &= \sum_{k=0}^{K-1}\left(-\frac{1}{2\Delta t_k}\left(\Delta\boldsymbol{x}(t_k) - \mathrm{J}\boldsymbol{x}(t_k)\right)^\top \Sigma_t^{-1}\left(\Delta\boldsymbol{x}(t_k) - \mathrm{J}\boldsymbol{x}(t_k)\right)\right) \\ &\quad + \text{const.} \end{aligned} \tag{12}$$

The log-likelihood corresponds to the *action* in statistical physics, where $\Gamma((\boldsymbol{x}(t_k))_{k=0}^{K-1}) = (\boldsymbol{x}(t_0), \ldots, \boldsymbol{x}(t_{K-1}))$ is a single trajectory of the stochastic variable.

Since the action in (12) is a convex function of the coupling matrix, the most probable coupling matrix (i.e., the one that maximizes the likelihood of the observed path) can be obtained by computing the derivative of the action with respect to the coupling matrix, setting it to zero, and solving for the coupling matrix.

The derivative of the log-path-likelihood function with respect to the coupling matrix can be factorized by the noise covariance because of its time-independence, giving the following closed-form solution

$$\begin{aligned} \hat{\mathrm{J}} = &\left(\sum_{k=0}^{K-1} \Delta\boldsymbol{x}(t_k)\boldsymbol{x}(t_k)^\top\right) \\ &\times \left(\sum_{k=0}^{K-1} \Delta t_k \boldsymbol{x}(t_k)\boldsymbol{x}(t_k)^\top\right)^{-1}, \end{aligned} \tag{13}$$

The single trajectory maximum path likelihood estimate (MPLE) in (13) can be easily generalized to the case of multiple trajectories or paths by replacing the action in (12) to an ensemble-averaged action $\langle \mathcal{S}(\mathrm{J}|\Gamma)\rangle_{\Gamma \in \text{ensemble}}$ (or, equivalently, by observing that the likelihood of a set of independent paths is equal to the product of the likelihoods for each individual path). The corresponding MPLE solution after ensemble averaging is

$$\hat{\mathrm{J}} = \left(\sum_{m=1}^{M} \sum_{k=0}^{K^m-1} \Delta \boldsymbol{x}^m(t_k) \boldsymbol{x}^m(t_k)^\top \right) \times \left(\sum_{m=1}^{M} \sum_{k=0}^{K^m-1} \Delta t_k \boldsymbol{x}^m(t_k) \boldsymbol{x}^m(t_k)^\top \right)^{-1} ,$$

where $m = 1, \ldots, M$ is the ensemble index.

In fact, by assuming the discretization of the OU process defined in the main text, we can estimate sample size dependence on the MPLE, and it is an *unbiased estimator*, as shown in below,

$$\begin{aligned}\hat{\mathrm{J}} &= \left(\sum_{m=1}^{M} \sum_{k=0}^{K^m-1} \Delta t_k \left(\hat{\mathrm{J}}^* \boldsymbol{x}^m(t_k) + \boldsymbol{W}(t_k) \right) \boldsymbol{x}^m(t_k)^\top \right) \\ &\quad \times \left(\sum_{m=1}^{M} \sum_{k=0}^{K^m-1} \Delta t_k \boldsymbol{x}^m(t_k) \boldsymbol{x}^m(t_k)^\top \right)^{-1} \\ &\sim \hat{\mathrm{J}}^* + \tilde{\boldsymbol{W}}/\sqrt{M} \quad \xrightarrow{M\to\infty} \quad \hat{\mathrm{J}}^* .\end{aligned}$$

To derive the scaling of the estimation bias, we used the assumption of the independence of the white noise.

Cameron-Martin-Girsanov theorem and application for Ornstein-Uhlenbeck process inference

In this section, we will show that the inference problem of the OU model can be solved by maximizing the *Radon-Nikodym* (RN) derivative or *likelihood ratio*, which is facilitated by the Cameron-Martin-Girsanov (CMG) theorem [4–7]. Since the aim of this section is only to rationalize the inference approach based on the CMG theory, we will discuss minimal ingredients of the CMG theory. A more general and comprehensive description can be found in refs. [4, 7].

First, let us define the RN derivative. If two probability measures $\mathbb{P}$ and $\mathbb{Q}$ satisfy the following conditions, then the $\mathbb{P}$ and $\mathbb{Q}$ are said to be *mutually absolutely continuous*,

$$\begin{aligned}\mathbb{E}_{\mathbb{Q}}[Y] &= \mathbb{E}_{\mathbb{P}}[YZ] \\ \mathbb{E}_{\mathbb{P}}[Y] &= \mathbb{E}_{\mathbb{Q}}[Y/Z] ,\end{aligned}$$

where $\forall Y > 0$. Z is some random variable, and if it satisfies the condition, $\mathbb{E}_{\mathbb{P}}[Z] = 1$, then Z is called Radon-Nikodym derivative (or likelihood ratio). In fact, it is nothing more than the changing of the probability measures

$$\mathbb{E}_{\mathbb{Q}}[Y] = \int Y d\mathbb{Q} = \int Y \frac{d\mathbb{Q}}{d\mathbb{P}} d\mathbb{P} = \mathbb{E}_{\mathbb{P}} \left[Y \frac{d\mathbb{Q}}{d\mathbb{P}} \right] .$$

Therefore, such a random variable Z is denoted as $\frac{d\mathbb{Q}}{d\mathbb{P}} := Z$ in general. Since the RN derivative gives transformation of a probability measure to another probability measure without obtaining (or even knowing explicit form of) the probability measure $\mathbb{Q}$, it enables us to estimate some statistics under the probability measure $\mathbb{Q}$ that are unobtainable directly. For example, *importance sampling* falls in this class of problems and is widely used in computational studies.

Informally speaking, the CMG theorem states that under some transformation of the drift term in a Wiener process, a probability measure after the transformation exists and can represent its explicit RN derivative. So, the CMG theorem provides a way to estimate the statistics under a probability density after a general transformation of the drift of the Wiener process.

More formally, the statement of the Cameron-Martin-Girsanov theorem is that for a Brownian motion $\{W_t\}_{t\geq 0}$ that follows a probability measure $\mathbb{P}$ and observable process γ_t that satisfies the following *Nikodym condition*

$$\mathbb{E}_{\mathbb{P}} \left[\exp \left(\frac{1}{2} \int_0^t \gamma_s^2 ds \right) \right] < \infty \ , \quad \forall t \geq 0 \ ,$$

the probability measure $\mathbb{Q}$ that corresponds to the stochastic process $dX_t = -\gamma_t dt + dW_t$ exists and the $\mathbb{Q}$-process is equivalent to $\mathbb{P}$-Brownian motion by modifying the Wiener process such that

$$\tilde{W}_t = W_t + \int_0^t \gamma_s ds \quad .$$

We can transform most stochastic processes to this type of stochastic process. For example, a stochastic process given by

$$dX_t = \gamma_t dt + \sigma_t(X_t) dW_t \ ,$$

Here, $\sigma_t(X_t)$ is a covariance that can depend not only on time but also on random variables, so it becomes a multiplicative noise [7]. Then we transform the stochastic process and drift such that $\tilde{X}_t = \sigma_t(X_t)^{-1} X_t$ and $\tilde{\gamma}_t(X_t) = \sigma_t(X_t)^{-1} \gamma_t$, then we can get the following stochastic process

$$d\tilde{X}_t = \tilde{\gamma}_t dt + dW_t \ .$$

These probability measures $\mathbb{P}$ and $\mathbb{Q}$ are related by the Radon-Nikodym derivative, which is defined as follows,

$$\frac{d\mathbb{Q}}{d\mathbb{P}} = \exp \left(- \int_0^t \gamma_s dW_s - \frac{1}{2} \int_0^t \gamma_s^2 ds \right) \ .$$

Using the CMG theorem, we can estimate statistical quantities under a more general probability measure $\mathbb{Q}$.

Since the CMG theorem provides explicit transformation of probability measures, the maximization of the likelihood ratio can be a substitution of the maximum likelihood,

$$\begin{aligned}\max_\theta \mathbb{Q}_\theta(A) &= \max_\theta \int_A \frac{d\mathbb{Q}_\theta}{d\mathbb{P}}(x)\ d\mathbb{P}(x) \\ &\leq \int_A \max_\theta \left\{\frac{d\mathbb{Q}_\theta}{d\mathbb{P}}(x)\right\}\ d\mathbb{P}(x)\ .\end{aligned}$$

Thus, we can estimate the most probable parameters by maximizing the likelihood ratio.

Now, we can apply the CMG theorem to the inference problem of the OU model. The CMG theorem lets the SDE (**??**) transform into the following

$$d\tilde{\boldsymbol{X}}_t = -\tilde{\boldsymbol{\gamma}}_t dt + d\boldsymbol{W}_t$$

where $\tilde{\boldsymbol{X}}_t = \Sigma^{-1/2}\boldsymbol{X}_t$ and $\tilde{\boldsymbol{\gamma}}_t = -\Sigma^{-1/2}\mathrm{J}\boldsymbol{X}_t$. More general transformation can be done by the *Lamperti transformation* that provides a systematic variable transformation rule so that a given SDE with multiplicative noise transforms to another SDE with an additive noise [8].

Therefore, the likelihood ratio of the OU model becomes as follows,

$$\begin{aligned}\frac{d\mathbb{Q}_\mathrm{J}}{d\mathbb{P}} &= \exp\left(-\int_0^t \tilde{\boldsymbol{\gamma}}_s^\top d\tilde{\boldsymbol{X}}_s - \frac{1}{2}\int_0^t \tilde{\boldsymbol{\gamma}}_s^\top \tilde{\boldsymbol{\gamma}}_s ds\right) \\ &= \exp\Bigg(\int_0^t (\mathrm{J}\boldsymbol{X}_s)^\top \Sigma^{-1} d\boldsymbol{X}_s \qquad (14) \\ &\qquad - \frac{1}{2}\int_0^t (\mathrm{J}\boldsymbol{X}_s)^\top \Sigma^{-1}(\mathrm{J}\boldsymbol{X}_s)ds\Bigg)\ ,\end{aligned}$$

where we used the symmetry of the covariance matrix and definition of the square matrix, $\Sigma^{1/2}\Sigma^{1/2} = \Sigma$.

Since the likelihood ratio (14) is a convex function of the coupling matrix, its derivative with respect to the coupling matrix gives the equation to solve the maximum likelihood estimator. So the derivative of the likelihood ratio is

$$\begin{aligned}\frac{\partial}{\partial \mathrm{J}} \log \frac{d\mathbb{Q}_\mathrm{J}}{d\mathbb{P}} &= -\int_0^t \Sigma^{-1} d\boldsymbol{X}_s \boldsymbol{X}_s^\top - \int_0^t \Sigma^{-1}\mathrm{J}\boldsymbol{X}_s\boldsymbol{X}_s^\top ds \\ &\xrightarrow{\mathrm{J}\to\mathrm{J}^*}\ 0\ .\end{aligned}$$

This immediately leads the maximum likelihood ratio estimator

$$\hat{\mathrm{J}} = \left(\int_0^t d\boldsymbol{X}_s\boldsymbol{X}_s^\top\right)\left(\int_0^t \boldsymbol{X}_s\boldsymbol{X}_s^\top ds\right)^{-1}\ .$$

To derive this solution, we used the fact that the inverse of the covariance is independent from the time and stochastic process.

The important consequence is that the maximum likelihood ratio based on the CMG theorem gives exactly the same solution as in the case of the path-likelihood maximization shown in (13).

Another derivation of optimal Wright-Fisher selection coefficients via Cameron-Martin-Girsanov theorem

In this section, we will rederive the maximum path likelihood solution of the selection in the WF model using the CMG theorem.

We can write the Langevin equation for the Wright-Fisher diffusion as

$$d\boldsymbol{X}_t = \left(\mathrm{C}(\boldsymbol{X}_t)\boldsymbol{s} + \boldsymbol{\mu}(\boldsymbol{X}_t)\right)dt + \sqrt{\mathrm{C}(\boldsymbol{X}_t)}d\boldsymbol{W}_t\ .$$

Applying the formulation of the Radon-Nikodym derivative to this Langevin equation, we obtain

$$\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}} = \exp\left(\int_0^t (\mathrm{C}(\boldsymbol{x}_s)\boldsymbol{s} + \boldsymbol{\mu}(\boldsymbol{x}_s))^\top \mathrm{C}(\boldsymbol{x}_s)^{-1} d\boldsymbol{x}_s - \frac{1}{2}\int_0^t (\mathrm{C}(\boldsymbol{x}_s)\boldsymbol{s} + \boldsymbol{\mu}(\boldsymbol{x}_s))^\top \mathrm{C}(\boldsymbol{x}_s)^{-1}(\mathrm{C}(\boldsymbol{x}_s)\boldsymbol{s} + \boldsymbol{\mu}(\boldsymbol{x}_s))ds\right)\ .$$

Since the logarithm of the RN derivative is a convex function, its derivative gives the solution that maximizes the RN derivative,

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{s}} \log \frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}} &= \int_0^t d\boldsymbol{x}_s - \int_0^t (\mathrm{C}(\boldsymbol{x}_s)\boldsymbol{s} + \boldsymbol{\mu}(\boldsymbol{x}_s))ds \\ &\xrightarrow{\boldsymbol{s}\to\boldsymbol{s}^*}\ \boldsymbol{0}\ .\end{aligned}$$

Therefore, the solution equivalent to the maximum path-likelihood solution is obtained.

$$\hat{\boldsymbol{s}} = \left(\int_0^t \mathrm{C}(\boldsymbol{x}_s)ds\right)^{-1}\int_0^t (d\boldsymbol{x}_s - \boldsymbol{\mu}(\boldsymbol{x}_s)ds)$$

Supplements

Performance of Bézier interpolation on real data

To apply Bézier interpolation to biological sequence data, we extended the approach described in the main sections binary variables to multivariates. This is necessary because DNA or RNA sequences have five possible states at each site, including four nucleotides and a "gap" symbol, which represents the absence of a nucleotide at a site that is present in other related sequences.

We applied multivariate Bézier interpolation to study human immunodeficiency virus (HIV-1) evolution in a set of 13 individuals [9] (see Methods for details). The distribution of selection coefficients inferred using Bézier interpolation is highly correlated with previous analysis using linear interpolation [10], indicating broad consistency with past results (**Fig. 2**). However, as we observed in simulations, inference using Bézier interpolation tends to result in slightly larger selection coefficients.

Consistent with past analyses [10], we found that the largest inferred selection coefficients are overwhelmingly associated with potentially functional mutations. Among the largest 1% of selection coefficients inferred across these 13 individuals, around 40% correspond to mutations that help the virus to escape from the host immune system. This represents a more than 20-fold enrichment in immune escape mutations among the most highly selected mutations, compared to chance expectations.

In summary, Bézier interpolation applied to real data leads to the inference of selection coefficients that are stronger than, but broadly consistent with, those that are found using linear interpolation. Large inferred selection coefficients also have clear biological interpretations. For HIV-1, many highly beneficial mutations correspond to ones that the virus uses to escape from the immune system.

Bézier interpolation also has the advantage that it conserves sums of categorical variables, which is not typically guaranteed for standard stochastic regression methods such as Gaussian process regression/Kriging [12, 13] or nonlinear approaches such as kernel regression or least squares [13, 14]. This property is especially useful for interpolating quantities that can be interpreted as probabilities (e.g., frequency vectors, as we considered above) or other conserved parameters. A few studies have applied regression methods to probabilities using logarithmic transformations. However, in such cases, regions around the 0 and 1 boundaries in the probability space tend to dominate regression results due to the coordinate transformation [15].

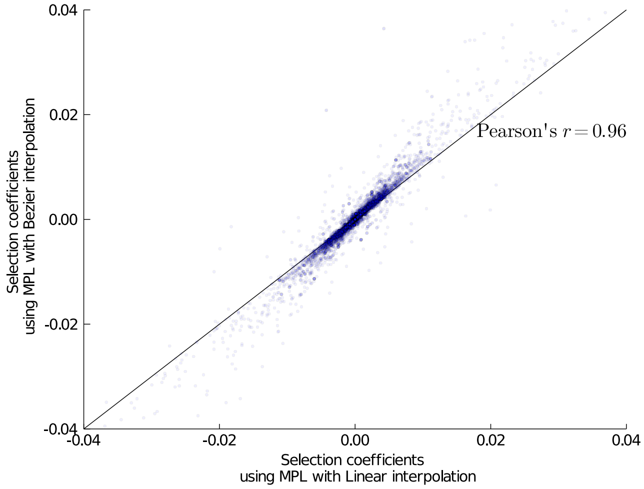

FIG. 2. **HIV-1 selection coefficients estimated by MPL with Bézier interpolation are strongly correlated with those estimated with linear interpolation.** Consistent with simulation results in **Fig. 3** in the main text, Bézier interpolation typically yields larger estimated selection coefficients. Selection coefficients were obtained for roughly 50 to 900 mutations per individual and sequencing region. Samples were obtained from 3-9 times per individual, with 7-40 sequences per time point. Sequences were collected frequently early in infection with $\Delta t \sim 10$ days, stretching to 100-200 days late in infection. Mutation rates from past studies [11] were used to estimate selection coefficients. The regularization strength is $\gamma = 10$ in both linear and Bézier cases.

Effect of regularization strength γ

We report the influence of the regularization on the precision of the selection coefficients based on positive predictive value (PPV) curves. In this test, we chose the following different regularization values $\gamma \in \{10^{-3}, 0.1, 1, 5, 10\}$. Through the all tests, we fixed the sampling interval as $\Delta t = 75$. For the other parameters, we use the same parameters that are used in the main section.

Supplementary Fig. 3 shows how inference accuracy depends on the regularization strength for MPL using different interpolation methods: piece-wise constant, linear, and Bézier interpolation.

In the case of the small to medium regularization values ($\gamma = 10^{-3}, 0.1, 1$), PPV curves using the Bézier interpolation are significantly higher than the PPV curves using other interpolation methods. As the regularization value increases, the difference between the PPV curves for linear and Bézier interpolations becomes smaller.

Supplementary Fig. 4 shows that MPL with Bézier interpolation outperforms MPL with linear interpolation for any regularization strength γ. The best PPV curves of MPL with linear interpolation are still lower than the majority of PPV curves for MPL using Bézier interpolation. Moreover, although a large regularization improves the PPV curves of MPL with linear interpolation, due to the strong regularization effect, the estimated selection coefficients are strongly biased and are underestimated as shown in **Supplementary Fig. 5**.

Effect of sampling interval Δt

Here, we discuss the effects of the sampling interval Δt on the different interpolation methods in detail. In this study, the model parameters for the population size and mutation rate are the same as in the main text, and the regularization coefficient is fixed as $\gamma = 0.1$.

Supplementary Fig. 6 shows PPV curves for estimated selection coefficients using MPL with piece-wise constant, linear, and Bézier interpolation depending on various sampling intervals $\Delta t \in \{1, 10, 30, 75, 100\}$.

For $\Delta t = 1, 10$, there is no difference among these methods. However, when $\Delta t = 30$, the PPV curves for the piecewise constant case deteriorate compared with the other methods and the ones for the linear and Bézier interpolations are indistinguishable. This is consistent with the argument in the main section: the characteristic time scale, $\gamma\Delta t$, is not so large that nonlinear effects are noticeable, hence PPV curves for the linear and Bézier interpolation are indistinguishable.

In the $\Delta t = 75$ case, the PPV curves of the MPL with Bézier interpolation are systematically higher than the cases of MPL with linear interpolation, hence MPL with Bézier interpolation outperforms other approaches.

In general, as the time interval increases, Bézier interpolation has a greater advantage in capturing the underlying dynamics of trajectories (**Supplementary Fig. 6**). However, for large enough time gaps, all interpolation methods suffer because data is sampled too sparsely to reveal any information about the underlying dynamics. For large enough Δt, there is no connection between the covariances at consecutively sampled points, and "trajectory information" is no longer contained in the data. This is also consistent with the negligible size of the autocorrelation for off-diagonal covariances at very large time gaps.

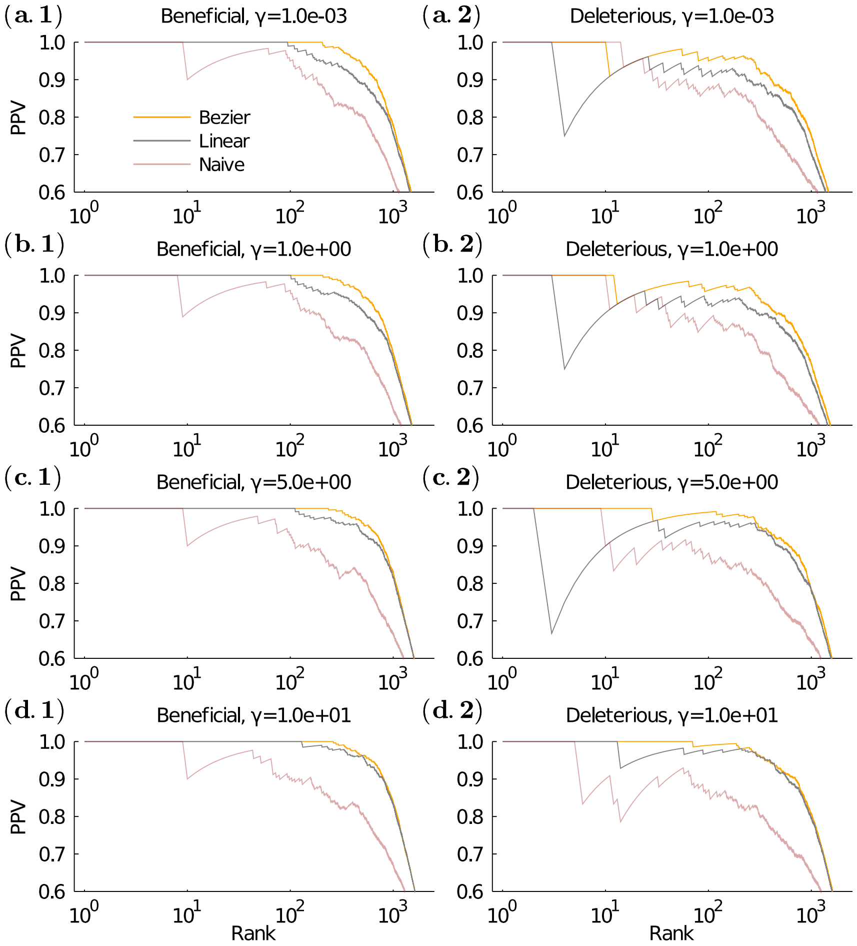

FIG. 3. **Across a wide range of regularization values, Bézier interpolation achieves more accurate inference than linear interpolation.** **(a.1)** PPV curves for beneficial selection coefficients using $\gamma = 10^{-3}$. Other conditions are the same as in the main text. **(a.2)** PPV curves for deleterious coefficients using $\gamma = 10^{-3}$. **(b)**, **(c)**, **(d)** the same type of figures but using $\gamma = 1.0, \gamma = 5.0$ and $\gamma = 10.0$. MPL with linear interpolation is sensitive to the regularization strength, and larger regularization is necessary to make more precise inferences. However, the most accurate PPV using MPL with linear interpolation ($\gamma = 30.0$) has almost the same performance as the least accurate inferences using MPL with Bézier interpolation (see **Supplementary Fig. 4**). Moreover, the larger regularization induces a strong estimation bias, as shown in **Supplementary Fig. 5**.

Positive semidefiniteness of the interpolated covariance

The eigenvalues of the covariance matrix are strictly non-negative. This positive semi-definiteness is an essential property of the covariance matrix and is practically important. We numerically confirmed the positive semidefiniteness of interpolated covariance matrices using the Bézier interpolation.

To evaluate the positive semidefiniteness, we generated a test data set by running the WF model 100 times. The dependent parameters of the WF model are the same as the main text. Then, we estimated integrated covariance matrices and their covariance matrix eigenvalues for different interpolation methods and different sampling intervals.

In either interpolation method, the eigenvalue distribution of the integrated covariance matrix showed little change, and only positive eigenvalues were observed in each case (**Supplementary Fig. 7**).

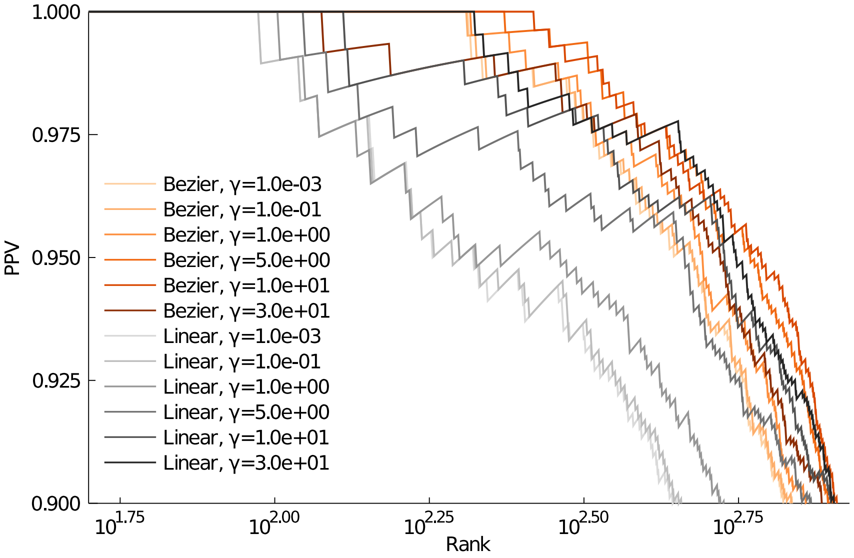

FIG. 4. **In a wide range of regularization, MPL with Bézier interpolation achieves higher PPVs than MPL with linear interpolation.** Here we show PPV curves between rank 60 and 900, where changes due to the different regularization values $\gamma = 10^{-3}, 0.1, 1.0, 5.0, 10.0$ and 30.0 are most noticeable. PPV curves of MPL with Bézier interpolation maintain high values stably. In contrast, PPV curves of MPL with linear interpolation are sensitive to the choice of the regularization strength and tend to be lower than those for MPL with Bézier interpolation. In the linear interpolation case, larger regularization yields higher the PPV curves, but also larger estimation bias.

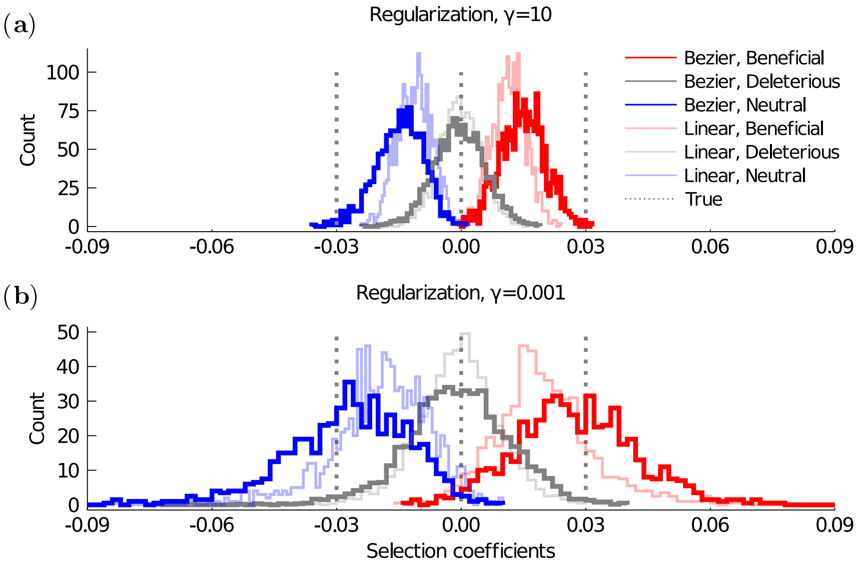

FIG. 5. **MPL with Bézier interpolation reduces estimation bias in a wide range of regularization, and small regularization is needed to avoid strong estimation bias. (a)** Distribution of inferred selection coefficients using a strong regularization $\gamma = 10$. Other conditions are the same as in the main text. **(b)** Estimated selection coefficients using a weak regularization $\gamma = 10^{-3}$. Smaller regularization $\gamma = 10^{-3}$ reduces estimation bias, especially for MPL with linear interpolation.

Selection coefficient inference without off-diagonals of integrated covariance elements

As shown in the main text, Bézier interpolation is better than linear interpolation in the sense of the more accurate reconstruction of the covariance matrix depending on perfectly observed trajectories (when the sampling interval $\Delta t = 1$) from the covariance matrix depending on "sparsely" observed trajectories, especially for the "off-diagonal" elements (corresponding to pair-

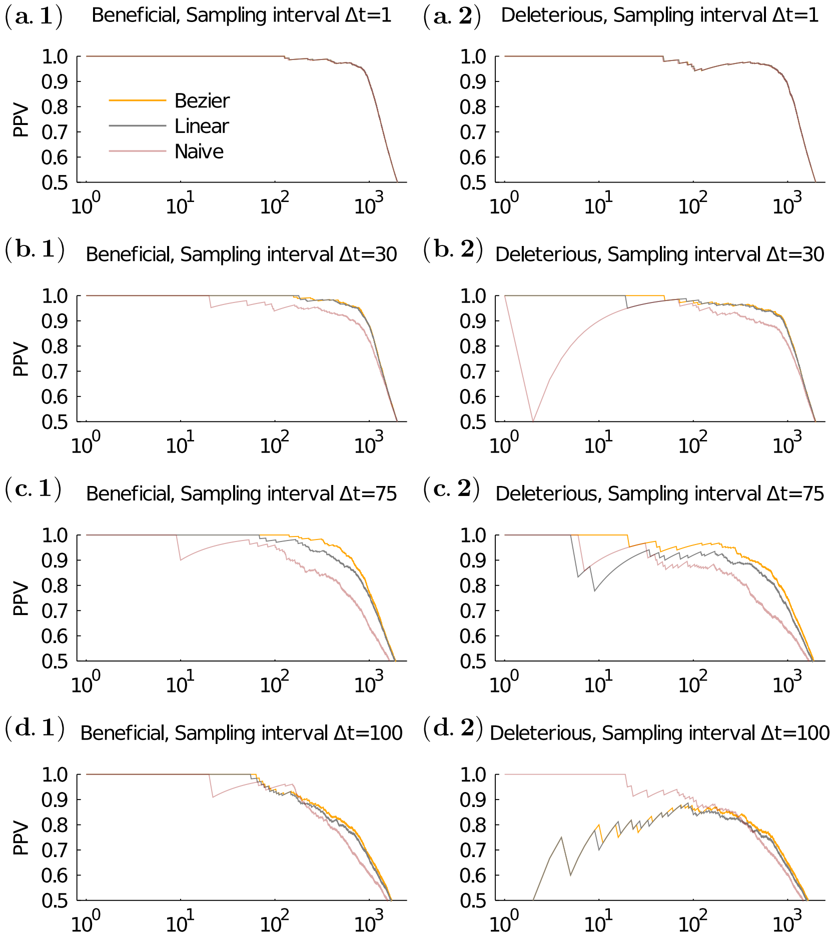

FIG. 6. **As the sampling interval increases, the advantage of the Bézier interpolation becomes more notable.** **(a.1)** PPV curves for beneficial selection coefficient inference using the sampling interval $\Delta t = 1$, using MPL with Bézier, linear, and piece-wise constant interpolations. Other conditions are the same as in the main text ($\gamma = 0.1, N = 10^3$, and $\mu = 10^{-3}$) **(a.2)** PPV curves for deleterious selection coefficient inference using $\Delta t = 1$. Subplots, **(b)**, **(c)** and **(d)** are the same type of figures but for $\Delta t = 30, 75$, and 100, respectively. As the sampling interval increases, the inference accuracy decreases in the PPV sense. However, inferences using Bézier interpolation degrade more slowly than other methods. For the longest sampling interval ($\Delta t = 100$), consecutive time points are poorly correlated. As a result, none of the interpolation methods can completely accurately infer selection coefficients, and hence the PPV curves roughly converge.

wise covariances $C_{ij} = x_{ij} - x_i x_j$, with $i \neq j$) of the integrated covariance matrix. On the other hand, the difference between linear and Bézier interpolation for the "diagonal" elements (variance $C_{ii} = x_i(1 - x_i)$) was relatively minor. To understand how exactly this observation is associated with the accuracy of the selection coefficients, we examine the effect of the off-diagonal entries of the integrated covariance matrix on the selection coefficients in this section.

Supplementary Fig. 8 shows the inference accuracy for both deleterious and beneficial mutations using MPL and the single locus (SL) method, a simplified inference method that ignores the off-diagonal of the integrated covariance matrix.

The PPV of MPL with Bézier interpolation achieves systematically higher values than the PPV of MPL with linear interpolation. However, the difference between linear and Bézier interpolation becomes unclear for inferences using SL. Thus, the main reason MPL with Bézier interpolation can infer better than MPL with linear interpolation is the accurate estimation of off-diagonal covariances (including pairwise frequencies).

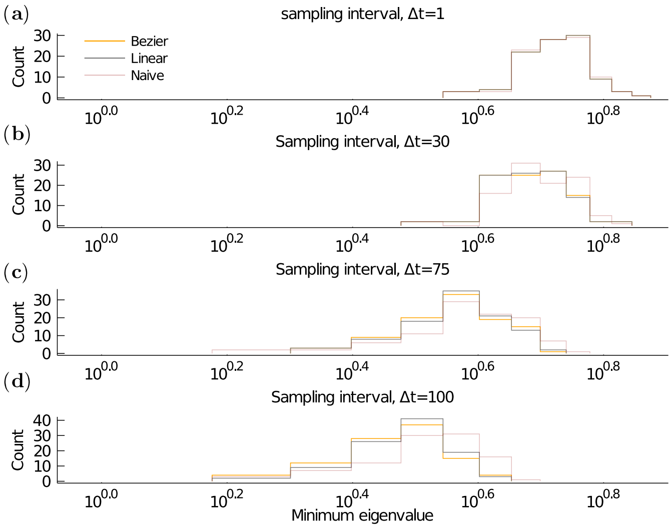

FIG. 7. **Covariance matrix with Bézier interpolation maintains positive semidefiniteness.** Comparison of the minimum eigenvalue distributions of the integrated covariance matrices: As the Δt increases, the minimum eigenvalues are smaller, but they remain nonnegative values. Thus, all the integrated covariances are positive definite.

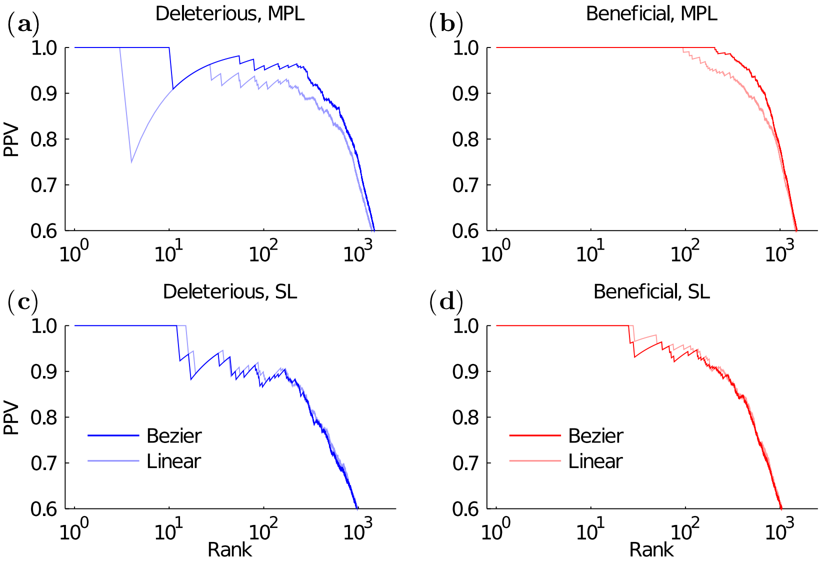

FIG. 8. **Improvement of the selection inference accuracy is due to the accurate restoration of pairwise frequencies.** PPV curves for **(a)** deleterious and **(b)** beneficial selection coefficients using MPL methods. The sampling interval is fixed as $\Delta t = 75$. PPV curves for **(c)** deleterious and **(d)** beneficial selection coefficients, but using the single locus (SL) method, a simplified version of MPL which sets off-diagonal elements of the covariance matrix to zero. Bézier interpolation improves the precision of MPL, but the choice between linear and Bézier interpolation does not significantly affect the accuracy of SL. This implies that the accurate estimation of pairwise frequencies (corresponding to off-diagonal covariances) improves selection inference accuracy.

Ornstein-Uhlenbeck process inference comparison

In this section, we report a more detailed analysis of the estimated coupling parameters of OU processes. The input data sets for the inference are the same as in the main section. To compare the inference accuracy between various inference methods, besides the path-likelihood-based methods, we included mean-field theory-based inference. In this approach, the effective solution is given by the inverse of the integrated covariance matrix, which effectively predicts interaction matrices for input data following an equilibrium distribution [16].

Supplementary Fig. 9 shows comparisons of a true interaction matrix and estimated interactions. The accuracy of the path-likelihood-based methods is significantly better than the the inverse of the covariance matrix in terms of Pearson's correlation and linear regression's slope. This is an anticipated result since the input data sets were generated from the relaxation processes, and the probability distributions that characterize these dynamics are in non-steady states. Therefore, MPL methods outperform inference methods assuming equilibrium states.

The path-likelihood-based inference method with Bézier interpolation achieves the best inference accuracy for both diagonal and off-diagonal interaction matrix elements in terms of Pearson's correlation coefficients and regression slope values.

Supplementary Fig. 10 shows sampling interval dependence for Pearson's correlation coefficients between true interaction matrices and inferred interaction matrices. The input data sets and conditions of the inferences are the same as the main text. As the sampling interval regime increases, the difference between Pearson's r of linear and Bézier interpolations becomes more pronounced, and the inferences using Bézier interpolation achieve higher Pearson's r values among all sampling intervals.

[1] Eid H Doha, Ali H Bhrawy, and MA Saker. Integrals of bernstein polynomials: an application for the solution of high even-order differential equations. *Applied Mathematics Letters*, 24(4):559–565, 2011.

[2] Ahmet Altürk. Application of the bernstein polynomials for solving volterra integral equations with convolution kernels. *Filomat*, 30(4):1045–1052, 2016.

[3] Daniel T Gillespie. Exact numerical simulation of the ornstein-uhlenbeck process and its integral. *Physical review E*, 54(2):2084, 1996.

[4] Hannes Risken. *The Fokker-Planck Equation: Methods of Solution and Applications.* Springer-Verlag, 2nd edition, 1989.

[5] Robert H Cameron and William T Martin. Transformations of weiner integrals under translations. *Annals of Mathematics*, pages 386–396, 1944.

[6] Igor Vladimirovich Girsanov. On transforming a certain class of stochastic processes by absolutely continuous substitution of measures. *Theory of Probability & Its Applications*, 5(3):285–301, 1960.

[7] Robert Shevilevich Liptser and Al'bert Nikolaevich Shiriaev. *Statistics of random processes: General theory*, volume 394. Springer, 1977.

[8] Stefano M Iacus. *Simulation and inference for stochastic differential equations: with R examples*, volume 486. Springer, 2008.

[9] Michael KP Liu, Natalie Hawkins, Adam J Ritchie, Vitaly V Ganusov, Victoria Whale, Simon Brackenridge, Hui Li, Jeffrey W Pavlicek, Fangping Cai, Melissa Rose-Abrahams, et al. Vertical t cell immunodominance and epitope entropy determine hiv-1 escape. *The Journal of clinical investigation*, 123(1), 2012.

[10] Muhammad Saqib Sohail, Raymond HY Louie, Matthew R McKay, and John P Barton. Mpl resolves genetic linkage in fitness inference from complex evolutionary histories. *Nature biotechnology*, 39(4):472–479, 2021.

[11] Fabio Zanini, Johanna Brodin, Lina Thebo, Christa Lanz, Göran Bratt, Jan Albert, and Richard A Neher. Population genomics of intrapatient HIV-1 evolution. *eLife*, 4:e11282, 2015.

[12] Ronald Christensen. *Advanced linear modeling.* Springer, 2019.

[13] David JC MacKay, David JC Mac Kay, et al. *Information theory, inference and learning algorithms.* Cambridge university press, 2003.

[14] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

[15] Qiushuang Lin and Chunxiang Li. Kriging based sequence interpolation and probability distribution correction for gaussian wind field data reconstruction. *Journal of Wind Engineering and Industrial Aerodynamics*, 205:104340, 2020.

[16] Faruck Morcos, Andrea Pagnani, Bryan Lunt, Arianna Bertolino, Debora S Marks, Chris Sander, Riccardo Zecchina, José N Onuchic, Terence Hwa, and Martin Weigt. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences*, 108(49):E1293–E1301, 2011.

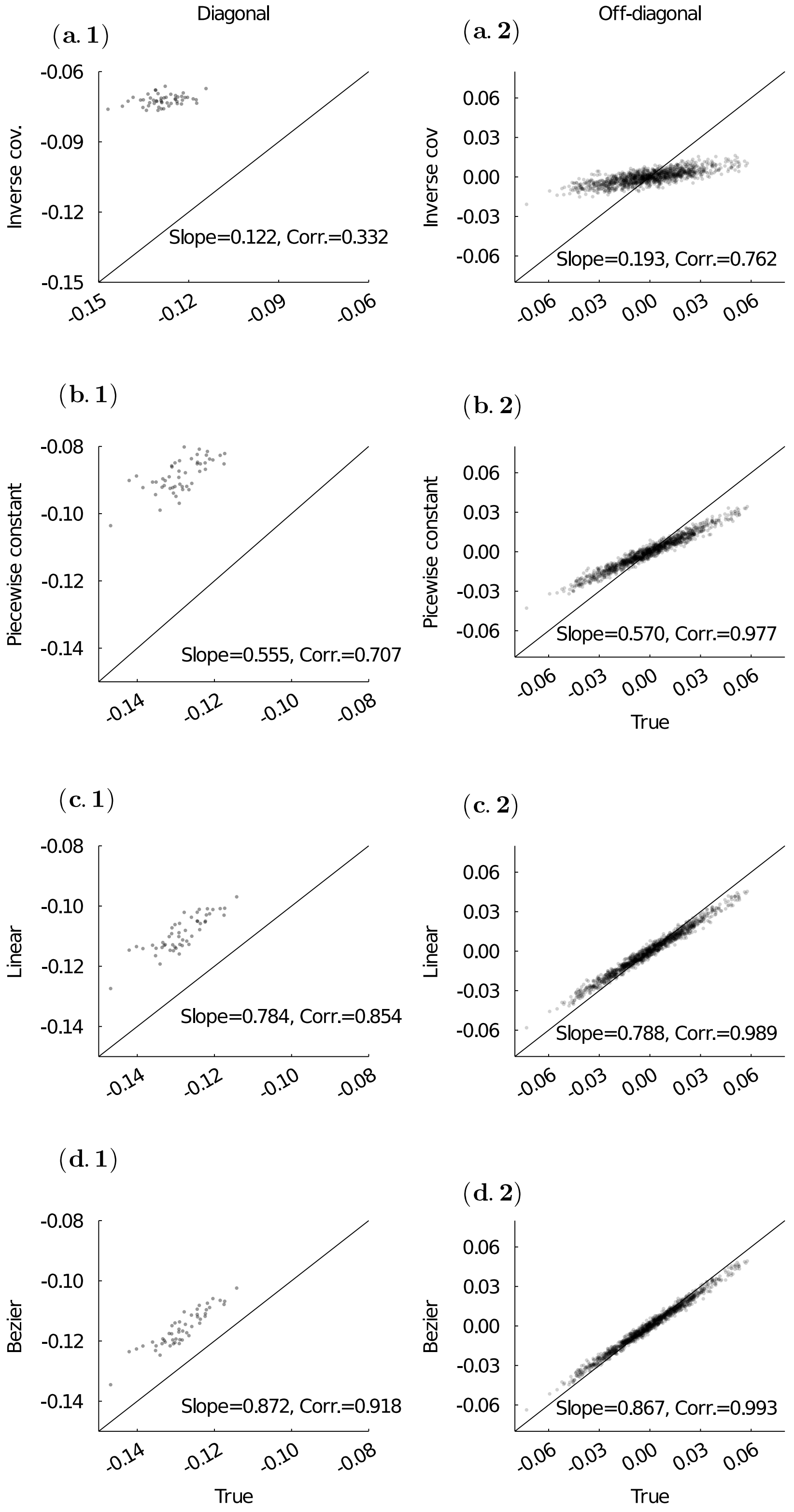

FIG. 9. **Path-likelihood-based inference method with Bézier interpolation achieves the best inference accuracy.** Comparison of true and inferred OU process interaction matrices. Mean-field based methods were used for **(a.1)** diagonal and **(a.2)** off-diagonal elements of the interaction matrix. Panels **(b)**, **(c)**, and **(d)** are the same type of plots, but using the path-likelihood-based inference with piecewise constant, linear, and Bézier interpolation, respectively. Among all the methods, inference with Bézier interpolation achieves the highest accuracy in terms of Pearson's correlation coefficient and regression slope value.

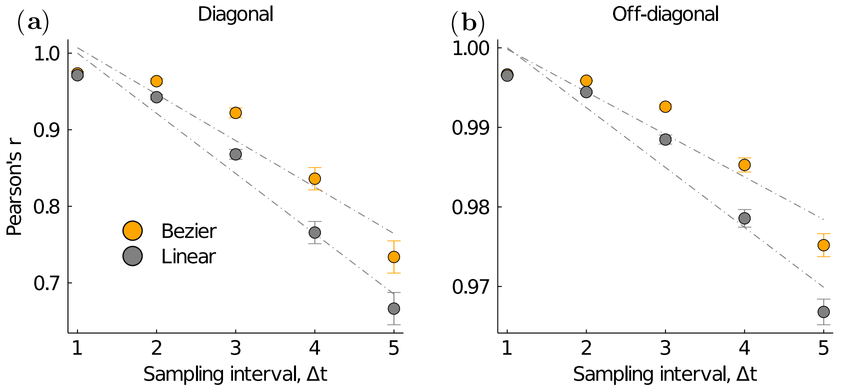

FIG. 10. **As the sampling interval increases, the advantage of Bézier interpolation becomes more pronounced.** Dependence of Pearson's correlation coefficients for **(a)** diagonal and **(b)** off-diagonal interaction matrices on the sampling interval. Pearson's correlation coefficients for Bézier interpolation are significantly higher than the ones for linear interpolation, especially when the sampling interval is large.