Title

Bézier interpolation improves the inference of dynamical models from data

Journal

Physical Review E, 107(2)

ISSN

2470-0045

Authors

Shimagaki, Kai
Barton, John P

Publication Date

2023-02-01

DOI

10.1103/physreve.107.024116

Peer reviewed

HHS Public Access

Author manuscript

Phys Rev E. Author manuscript; available in PMC 2023 March 20.

Published in final edited form as:

Phys Rev E. 2023 February ; 107(2-1): 024116. doi:10.1103/PhysRevE.107.024116.

Bézier interpolation improves the inference of dynamical models from data

Kai Shimagaki[1], **John P. Barton**[1,2,*]

[1]Department of Physics and Astronomy, University of California, Riverside, California 92521, USA

[2]Department of Computational and Systems Biology, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania 15213, USA

Abstract

Many dynamical systems, from quantum many-body systems to evolving populations to financial markets, are described by stochastic processes. Parameters characterizing such processes can often be inferred using information integrated over stochastic paths. However, estimating time-integrated quantities from real data with limited time resolution is challenging. Here, we propose a framework for accurately estimating time-integrated quantities using Bézier interpolation. We applied our approach to two dynamical inference problems: Determining fitness parameters for evolving populations and inferring forces driving Ornstein-Uhlenbeck processes. We found that Bézier interpolation reduces the estimation bias for both dynamical inference problems. This improvement was especially noticeable for data sets with limited time resolution. Our method could be broadly applied to improve accuracy for other dynamical inference problems using finitely sampled data.

I. INTRODUCTION

Stochastic processes are ubiquitous in nature. In biology, the evolution of genetic sequences can be formulated as a stochastic process. The Wright-Fisher (WF) model [1], a discrete-time stochastic process, has been used to study the evolution of organisms from viruses [2–4] to humans [5]. Models such as the Ornstein-Uhlenbeck (OU) process [6,7] have been applied to describe a wide range of phenomena, from the fluctuation of currency exchange rates [8] and cell migration [9] to driven quantum many-body systems [10].

Appropriate model parameters are needed to accurately describe the behavior or real systems. To infer such parameters from data, it is often necessary to compute statistics over a *path*, i.e., a complete realization of the stochastic processes. For example, the restoring force of the OU process can be estimated by taking the ratio of the deviation from the equilibrium position and the magnitude of the intrinsic fluctuations, both integrated over a stochastic path [11,12].

*Corresponding author: jpbarton@pitt.edu.

All authors contributed to methods development, data analysis, interpretation of results, and writing the paper. K.S. performed simulations and computational analyses. J.P.B. supervised the project.

However, real data often consists of incomplete, occasional measurements of a system, which may also be limited by experimental constraints. This makes it more difficult to accurately estimate model parameters since statistics over the path must be estimated from incomplete information.

Here, we propose a tractable nonlinear interpolation framework using Bézier curves [13–16]. In addition to incorporating nonlinearity, this approach has the added advantage of conserving sums of categorical variables. This property can be especially useful for conserved quantities such as probabilities.

We applied Bézier interpolation to two example problems: inferring natural selection in evolving populations through the WF model and inferring restoring forces for OU processes. Our method reduces estimation bias and improves the precision of model inferences. Furthermore, we show that the autocorrelation function of statistics over a path identifies time scales over which nonlinear interpolation is particularly effective, which is consistent with our observations in simulations. We show that Bézier interpolation can generically improve solutions of dynamical inference problems by accurately estimating statistics over stochastic paths. We expect that this nonlinear interpolation method can improve a wide range of dynamical inference problems beyond the specific examples we consider, such as parameter estimation for stochastic differential equations.

II. RESULTS

A. Bézier interpolation

Consider a function $x(t)$ sampled at discrete times t_k for $k \in \{0, 1, \ldots, K\}$. Then the interpolated value of the function $x_B^{(k)}(t)$ between two successive discrete time points t_k and t_k+1 is given by

$$x_B^{(k)}(t) = \sum_{n=0}^{P} \beta_n\left(\frac{t - t_k}{t_{k+1} - t_k}\right)\phi_n^{(k)}\left[(x(t_{k'}))_{k'=0}^{K}\right]. \tag{1}$$

Here, β_n is the nth Bernstein basis polynomial of degree P, with $\beta_n(\tau) = \binom{P}{n}\tau^n(1-\tau)^{P-n} \geqslant 0$. The control points $\phi_n^{(k)}\left[(x(t_{k'}))_{k'=0}^{K}\right]$ depend on the ensemble of data points $(x(t_k))_{k=0}^{K}$ and determine the outline of the interpolation curves.

For simplicity, we consider cubic ($P = 3$) interpolation, but our approach can be extended to polynomials of different degrees P. We impose the following conditions to ensure that the segment at each interval $[t_k, t_k+1]\forall k$ is seamlessly connected,

$$\phi_0^{(k)}\left[(x(t_{k'}))_{k'=0}^{K}\right] = x(t_k), \quad \phi_3^{(k)}\left[(x(t_{k'}))_{k'=0}^{K}\right] = x(t_{k+1}).$$

Other internal pointsare $\left\{(\phi_1^{(k)}, \phi_2^{(k)})\right\}_{k=0}^{K-1}$ are obtained by solving an optimization problem that reflects continuity and smoothness constraints imposed on the curves (see Supplemental Material [17]) (Fig. 1).

B. Wright-Fisher model of evolution

The WF model [1] is a classical model in evolutionary biology. In this model, a population of N individuals evolves over discrete generations under the influence of random mutations and natural selection. Each individual is represented by a genetic sequence of length L. For simplicity, we assume that each site in the genetic sequence is occupied by a mutant (1) or wild-type (0) nucleotide. There are thus $M = 2^L$ possible *genotypes* (i.e., genetic sequences) in the population.

The state of the population is described by a genotype frequency vector $z(t) = (z_a(t))_{a=1}^{M}$, where $z_a(t)$ represents the frequency of individuals with genotype a in the population at time t. In the WF model, the probability of obtaining a genotype frequency vector $\boldsymbol{z}'$ in the next generation is multinomial, with the succession probability of genotype a

$$p_a(\boldsymbol{z}(t)) \propto f_a z_a(t) + \sum_{b|b\neq a} \left[\mu_{ba} z_b(t) f_b - \mu_{ab} z_a(t) f_a\right]. \tag{2}$$

In (2), f_a denotes the *fitness* of genotype a. Individuals with higher fitness values reproduce more readily than those with lower fitness values. Here μ_{ab} is the probability to mutate from genotype a to genotype b.

Fitness values can be estimated from data by identifying the f_a that are most likely to generate the observed evolutionary history of a population, but this is challenging due to the enormous size of the genotype space. The problem can be simplified by assuming that fitness values are additive, $f_a = 1 + \sum_{i=1}^{L} \sigma_i^a s_i$ where $\sigma_i^a = 1$ if the nucleotide at site i in genotype a is a mutant and 0 otherwise. The s_i are referred to as *selection coefficients*, which are positive if the mutation at site i is beneficial for reproduction and negative if mutation at site i is deleterious. Similarly, the mutation rate μ_{ab} can be simplified to a constant μ if genotypes a and b differ from one another by only a single mutation and zero otherwise.

Sohail *et al.* solved this problem analytically in the limit that the population size $N \to \infty$, while the selection coefficients s_i and mutation rate μ scale as $1/N$ (Ref. [4]). In this case, the maximum *a posteriori* vector of selection coefficients $\hat{\boldsymbol{s}} = (\hat{s}_i)_{i=1}^{L}$ is

$$\hat{\boldsymbol{s}} = \left(\int_{t_0}^{t_K} dt\, \mathrm{C}(t) + \gamma \mathrm{I}\right)^{-1} \times \left[\boldsymbol{x}(t_K) - \boldsymbol{x}(t_0) - \mu \int_{t_0}^{t_K} dt [1 - 2\boldsymbol{x}(t)]\right], \tag{3}$$

where the time of observation runs from t_0 to t_K. In (3), $\boldsymbol{x}(t) = (x_i(t))_{i=1}^{L}$ is a vector of mutant frequencies (i.e., the number of individuals in the population with a mutation at site i at time t) and $\mathrm{C}(t)$ is the covariance matrix of mutant frequencies at time t. Here γ is the precision of a Gaussian prior distribution for the selection coefficients with mean zero and I is the identity matrix.

Extensive past work has also considered numerical solutions to this problem [3,5,18–22], though the analytical formula in (3) typically outperforms numerical approaches [4]. Sohail *et al.* referred to (3) as the marginal path likelihood (MPL) estimate for the selection coefficients, obtained by maximizing the posterior probability of an evolutionary history

with respect to the selection coefficients. The MPL approach has also been extended to consider more complex evolutionary models [23], missing covariance data [24], and epidemiological dynamics [25].

C. Bézier interpolation for WF model inference

In practice, Eq. (3) is not straightforward to evaluate because sequence data comes at discrete times $(t_k)_{k=0}^{K}$. However, Bézier interpolation allows us to analytically integrate both mutant frequency trajectories $\boldsymbol{x}(t)$ and covariances C(t), obtained by interpolating frequencies and computing $C_{ij}(t) = x_{ij}(t) - x_i(t)x_j(t)$. Here $x_{ij}(t)$ is the frequency of individuals in the population at time t that have mutations at both sites i and j.

To assess the performance of Bézier interpolation for inferring selection in the WF model, we generated a test data set by running 100 replicate simulations of WF evolution with identical parameters [Fig. 2(a)]. We then inferred selection coefficients from this data using MPL with linear and Bézier interpolation, applied to data sampled at discrete intervals Δt = 75 generations apart. While MPL with linear interpolation readily distinguishes between beneficial, neutral, and deleterious parameters, the inferred selection coefficients are shrunk towards zero. However, parameters inferred using Bézier interpolation are distributed around their true values [Fig. 2(b)]. Bézier interpolation reduces estimation bias due to long intervals between observation intervals by producing better estimates of underlying covariances (which we will quantify below). Here we used a regularization strength of γ = 0.1, but similar results are obtained with different choices for the regularization (see Supplemental Material [17]).

Next we studied how Bézier interpolation affects our ability to classify mutations as beneficial or deleterious, which we evaluated by ranking mutations according to their inferred selection coefficients. We quantified classification accuracy using positive predictive value (PPV), PPV = TP/(TP + FP), where TP and FP are the numbers of true positive and false positive predictions.

The PPV curves for beneficial or deleterious mutations estimated by MPL with Bézier interpolation are higher than those with linear interpolation, indicating more accurate classification [Figs. 3(a) and 3(b)]. This can be understood by observing reduced overlap between the distribution of inferred selection coefficients for beneficial, neutral, and deleterious mutations using Bézier interpolation [Fig. 3(c)].

D. Recovery of rapidly decaying correlations underlies improved accuracy

To understand why MPL with Bézier interpolation yields more accurate inferences, we studied errors between true and estimated parameters as a function of the time interval Δt between samples. For arbitrary matrices M we define an error function $\mathscr{E}(\Delta t) = \|\mathcal{I}_{\mathrm{M}}(\Delta t) - \mathcal{I}_{\mathrm{M}}(1)\|/\|\mathcal{I}_{\mathrm{M}}(1)\|$, normalizing by the matrix norm $\|\mathcal{I}_{\mathrm{M}}(1)\|$, which corresponds to perfect sampling for the WF model. Here, $\mathcal{I}_{\mathrm{M}}(\Delta t)$ is a time integral depending on the type of integration (piecewise constant, linear, and Bézier). For example, for piecewise constant integration, it will be $\mathcal{I}_{\mathrm{M}}(\Delta t) = \sum_{k=1}^{\lfloor T/\Delta t \rfloor} \mathrm{M}(k\Delta t)\Delta t$. In the discussion below we apply the L_2 norm, $\| \mathrm{M} \| = \sqrt{\left(\sum_{i,j} M_{ij}^2\right)}$, but other conventions could also be considered.

Using the metric defined above, we found that Bézier interpolation yields better estimates for both the diagonal and off-diagonal terms of the mutant frequency covariance matrix. However, the error for the off-diagonal covariances is larger and increases much more rapidly with increasing Δt than the error for the diagonal variances [Figs. 4(a) and 4(b)]. The reduction in error for Bézier interpolation is more substantial for off-diagonal terms compared to diagonal ones. Consistent with this observation, Bézier interpolation yields smaller improvements in performance for a simple version of MPL in which the off-diagonal terms of the integrated covariance matrix are ignored (see Supplemental Material [17]) [referred to as the single locus (SL) method in Ref. [4]].

To study the time scale τ on which nonlinear effects become important and Bézier interpolation is advantageous, we modeled the covariance elements using a simple Langevin equation, $\dot{z}(t) = -\lambda z(t) + \xi(t)$. Here $z(t)$ represents an element of the covariance matrix, $\lambda > 0$ is a damping coefficient, and $\xi(t)$ is a standard white noise with $\langle \xi(t) \rangle = 0$ and $\langle \xi(t)\xi(t+\tau) \rangle = 2\delta(\tau)$. Following this approach, a linear approximation should describe the evolution of $z(t)$ accurately if $\lambda \Delta t \ll 1$; otherwise, nonlinear nature of the $z(t)$ becomes significant and at this point the linear approximation cannot capture the actual evolution of $z(t)$.

The damping coefficient λ can be estimated by computing the autocorrelation function (ACF) of the covariance matrix elements, which can be matched to expectations from the Langevin equation, $\langle x(t)x(t+\tau) \rangle \propto \exp(-\lambda\tau)$. In our simulations, the exponents of the ACF for diagonal and off-diagonal terms are around $\lambda_d \sim 1/325$ and $\lambda_o \sim 1/50$, respectively [Fig. 4(c)]. When the time between sampling events is $\Delta t = 75$, where Bézier interpolation clearly has an advantage (Fig. 3), for diagonal and off-diagonal covariances we have $\lambda_d \Delta t \sim 0.23$ and $\lambda_o \Delta t = 1.5$, respectively. At this point, $\lambda_o \Delta t$ is $O(1)$, indicating the onset of nonlinearity for off-diagonal terms. Consistent with this observation, for this value of Δt, Bézier interpolation has notably lower error for off-diagonal covariances than linear interpolation, while errors for the diagonal terms are comparable.

While we focused specifically on the WF model in this example, the principle of autocorrelations and transitioning between linear and nonlinear behavior is general. This can allow us to anticipate the benefit of nonlinear interpolation for a wide range of problems.

E. Inference of forces in Ornstein-Uhlenbeck processes

We further applied Bézier interpolation to accurately infer the collective forces in Ornstein-Uhlenbeck (OU) processes, which plays important roles in various fields such as physics, biology, and mathematical finance [11,26–28]. Data has been used to infer the parameters of OU processes describing phenomena including cell migration [29], coevolution of species [30], and currency exchange rates [31], to name a few examples.

We consider the following multivariate OU process:

$$d\boldsymbol{X}_t = \mathrm{J}\boldsymbol{X}_t dt + \Sigma^{1/2} d\boldsymbol{W}_t . \tag{4}$$

Here t is the time variable, L is the number of OU variables, $\boldsymbol{X}_t \in \mathbb{R}^L$, $\mathbf{J} \in \mathbb{R}^{L \times L}$ is a negative semidefinite matrix, $\sum$ is a time-independent noise covariance, and $\boldsymbol{W}_t$ is a Wiener process.

We assume that the noise covariance matrix is constant over the evolution and given. Therefore, the unknown variable in the SDE in (4) is only the drift term, the interaction matrix J.

One of the most commonly used approaches for inferring stochastic force in OU processes is maximizing the likelihood ratio or Radon-Nikodym derivative, which is the ratio of two probability measures [12,32] and is commonly employed in fields such as mathematical finance [11]. In our problem, the likelihood ratio is defined as the probability density obeying the dynamics of (4) with interactions divided by the probability density of a "null" model with no interactions. Here, we inferred OU interactions by directly maximizing the path likelihood, as described for the WF model. Interestingly, this leads to exactly the same solution as the one for the standard likelihood or Radon-Nikdym derivative methods (see Supplemental Material [17]).

The interaction matrix $\hat{J}$ that best describes the data is given by

$$\hat{\mathbf{J}} = \left(\sum_{k=0}^{K-1} \Delta \boldsymbol{x}(t_k) \boldsymbol{x}(t_k)^\top \right) \left(\sum_{k=0}^{K-1} \Delta t_k \boldsymbol{x}(t_k) \boldsymbol{x}(t_k)^\top \right)^{-1} . \tag{5}$$

Here $(\boldsymbol{x}(t_k))_{k=0}^{K-1}$ is the observed trajectory following the OU process, $\Delta t_k = t_{k+1} - t_k$ is an observation interval, and $\Delta \boldsymbol{x}(t_k) = \boldsymbol{x}(t_{k+1}) - \boldsymbol{x}(t_k)$ is the amount of change during the kth interval.

To generate test data, we simulated the OU process using negative definite interaction matrices (see Supplemental Material [17]), which follows the construction of a Hopfield network [33]. Hopfield networks were first constructed to study associative memory [33] and have since been applied to inference problems in biology [34–37].

Interaction parameters estimated using Bézier interpolation matched better with the true, underlying parameters than those inferred using linear interpolation or a piecewise-constant assumption for the $\boldsymbol{x}(t)$ [Fig. 5(a)]. In particular, large parameters inferred with linear interpolation or the piecewise-constant assumption tended to be underestimated. In addition, we found that the slope relating the true and inferred parameters decreases as the sampling interval Δt increases. However, the slope decreases more slowly for Bézier interpolation compared to linear interpolation [Figs. 5(b) and 5(c)]. Overall, OU interaction parameters inferred using Bézier interpolation more closely match the true, underlying parameters than those inferred with simpler interpolation approaches, with gains in performance that increase as data becomes more limited.

III. DISCUSSION

Here we developed a nonlinear interpolation method using Bézier curves that improves the inference of dynamical models from finite data. We applied our approach to two problems: The inference of natural selection in evolving populations and interactions in multivariate Ornstein-Uhlenbeck processes. Bézier interpolation makes inference more precise and reduces bias, especially for data sets that are more sparsely sampled.

Because of its generality, Bézier interpolation could be broadly applied to give more reliable results for dynamic inference problems. For example, our approach could be combined with methods to learn forces from nonequilibrium dynamics [38,39] or ones used to learn parameters of stochastic differential equations from finitely sampled data [6,11,40].

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

The work of K.S. and J.P.B. reported in this publication was supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award No. RG35GM138233.

References

[1]. Ewens WJ, Mathematical Population Genetics: Theoretical Introduction (Springer, Berlin, 2004), Vol. 1.

[2]. Foll M, Poh Y-P, Renzette N, Ferrer-Admetlla A, Bank C, Shim H, Malaspinas A-S, Ewing G, Liu P, Wegmann D et al. , Influenza virus drug resistance: A time-sampled population genetics perspective, PLoS Genet. 10, e1004185 (2014). [PubMed: 24586206]

[3]. Ferrer-Admetlla A, Leuenberger C, Jensen JD, and Wegmann D, An approximate Markov model for the Wright–Fisher diffusion and its application to time series data, Genetics 203, 831 (2016). [PubMed: 27038112]

[4]. Sohail MS, Louie RHY, McKay MR, and Barton JP, Mpl resolves genetic linkage in fitness inference from complex evolutionary histories, Nat. Biotechnol 39, 472 (2021). [PubMed: 33257862]

[5]. Mathieson I and McVean G, Estimating selection coefficients in spatially structured populations from time series data of allele frequencies, Genetics 193, 973 (2013). [PubMed: 23307902]

[6]. Iacus SM, Simulation and Inference for Stochastic Differential Equations: With R Examples (Springer, Berlin, 2008), Vol. 486.

[7]. Gillespie DT, Exact numerical simulation of the Ornstein-Uhlenbeck process and its integral, Phys. Rev. E 54, 2084 (1996).

[8]. Roberts GO, Papaspiliopoulos O, and Dellaportas P, Bayesian inference for non-gaussian ornstein–uhlenbeck stochastic volatility processes, J. R. Stat. Soc.: Ser. B 66, 369 (2004).

[9]. Dieterich P, Klages R, Preuss R, and Schwab A, Anomalous dynamics of cell migration, Proc. Natl. Acad. Sci. USA 105, 459 (2008). [PubMed: 18182493]

[10]. Jung P, Periodically driven stochastic systems, Phys. Rep 234, 175 (1993).

[11]. Phillips PCB and Yu J, Maximum likelihood and gaussian estimation of continuous time models in finance, in Handbook of Financial Time Series (Springer, Berlin, 2009), pp. 497–530.

[12]. Liptser RS and Shiriaev AN, Statistics of Random Processes: General Theory (Springer, Berlin, 1977), Vol. 394.

[13]. Farouki RT, The Bernstein polynomial basis: A centennial retrospective, Computer Aided Geometric Design 29, 379 (2012).

[14]. Choi J-W, Curry R, and Elkaim G, Path planning based on bézier curve for autonomous ground vehicles, in Advances in Electrical and Electronics Engineering-IAENG Special Edition of the World Congress on Engineering and Computer Science 2008 (IEEE, New York, 2008), pp. 158–166.

[15]. Simba KR, Uchiyama N, and Sano S, Real-time obstacleavoidance motion planning for autonomous mobile robots, in Proceedings of the 2014 4th Australian Control Conference (AUCC) (IEEE, New York, 2014), pp. 267–272.

[16]. Forrest AR, Interactive interpolation and approximation by bézier polynomials, Comput. J 15, 71 (1972).

[17]. See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevE.107.024116 for details about (1) the data and code for the reported results, (2) the definition of Bézier interpolation, (3) analytical expressions of the integrated covariance matrix, (4) analysis of the normalizability and non-negativity of Bézier interpolation, (5) details of Ornstein-Uhlenbeck (OU) simulations, (6) maximum path-likelihood estimation for OU processes, (7) application of the Cameron-Martin-Girsanov theorem for OU inference, (8) application of Bézier interpolation for hiv-1 data, (9) the effects of regularization strength and sampling intervals, (10) numerical tests of the positive-semidefiniteness of interpolated covariance matrices, and (11) complementary simulations for inference of OU processes. It also includes Refs. [41–51].

[18]. Lacerda M and Seoighe C, Population genetics inference for longitudinally-sampled mutants under strong selection, Genetics 198, 1237 (2014). [PubMed: 25213172]

[19]. Tataru P, Simonsen M, Bataillon T, and Hobolth A, Statistical inference in the wright–fisher model using allele frequency data, Syst. Biol 66, e30 (2017). [PubMed: 28173553]

[20]. Schraiber JG, Evans SN, and Slatkin M, Bayesian inference of natural selection from allele frequency time series, Genetics 203, 493 (2016). [PubMed: 27010022]

[21]. Foll M, Shim H, and Jensen JD, WFABC: A Wright-Fisher ABC-based approach for inferring effective population sizes and selection coefficients from time-sampled data, Mol. Ecol. Resources 15, 87 (2015).

[22]. Iranmehr A, Akbari A, Schlötterer C, and Bafna V, Clear: Composition of likelihoods for evolve and resequence experiments, Genetics 206, 1011 (2017). [PubMed: 28396506]

[23]. Sohail MS, Louie RHY, Hong Z, Barton JP, and McKay MR, Inferring epistasis from genetic time-series data, Mol. Biol. Evol 39, msac199 (2022). [PubMed: 36130322]

[24]. Li Y and Barton JP, Estimating Linkage Disequilibrium and Selection from Allele Frequency Trajectories, Genetics (2023).

[25]. Lee B, Sohail MS, Finney E, Ahmed SF, Quadeer AA, McKay MR, and Barton JP, Inferring effects of mutations on SARS-Cov-2 transmission from genomic surveillance data, medRxiv 2021, 10.1101/2021.12.31.21268591.

[26]. Bouchaud J-P and Cont R, A Langevin approach to stock market fluctuations and crashes, Eur. Phys. J. B 6, 543 (1998).

[27]. Vasicek O, An equilibrium characterization of the term structure, J. Financ. Econ 5, 177 (1977).

[28]. Mamon RS, Three ways to solve for bond prices in the vasicek model, Adv. Decision Sci 8, 1 (2004).

[29]. Brückner DB, Fink A, Schreiber C, Röttgermann PJF, Rädler JO, and Broedersz CP, Stochastic nonlinear dynamics of confined cell migration in two-state systems, Nat. Phys 15, 595 (2019).

[30]. Ho LST and Ané C, Intrinsic inference difficulties for trait evolution with ornstein-uhlenbeck models, Methods Ecol. Evol 5, 1133 (2014).

[31]. Roberts GO and Rosenthal JS, Optimal scaling of discrete approximations to langevin diffusions, J. R. Stat. Soc.: Ser. B 60, 255 (1998).

[32]. Risken H, The Fokker-Planck Equation: Methods of Solution and Applications, 2nd ed. (Springer-Verlag, Berlin, 1989).

[33]. Hopfield JJ, Neural networks and physical systems with emergent collective computational abilities, Proc. Natl. Acad. Sci. USA 79, 2554 (1982). [PubMed: 6953413]

[34]. Cocco S, Monasson R, and Weigt M, From principal component to direct coupling analysis of coevolution in proteins: Low-eigenvalue modes are needed for structure prediction, PLoS Comput. Biol 9, e1003176 (2013). [PubMed: 23990764]

[35]. Tubiana J, Cocco S, and Monasson R, Learning protein constitutive motifs from sequence data, Elife 8, e39397 (2019). [PubMed: 30857591]

[36]. Shimagaki K and Weigt M, Selection of sequence motifs and generative Hopfield-Potts models for protein families, Phys. Rev. E 100, 032128 (2019). [PubMed: 31639992]

[37]. Shimagaki K and Weigt M, Selection of sequence motifs and generative Hopfield-Potts models for protein families, bioRxiv 652784 (2019), 10.1101/652784.

[38]. Frishman A and Ronceray P, Learning Force Fields from Stochastic Trajectories, Phys. Rev. X 10, 021009 (2020).

[39]. Brückner DB, Ronceray P, and Broedersz CP, Inferring the Dynamics of Underdamped Stochastic Systems, Phys. Rev. Lett 125, 058103 (2020). [PubMed: 32794851]

[40]. Ferretti F, Chardès V, Mora T, Walczak AM, and Giardina I, Building General Langevin Models from Discrete Datasets, Phys. Rev. X 10, 031018 (2020).

[41]. Doha EH, Bhrawy AH, and Saker MA, Integrals of bernstein polynomials: An application for the solution of high even-order differential equations, Appl. Math. Lett 24, 559 (2011).

[42]. Altürk A, Application of the bernstein polynomials for solving volterra integral equations with convolution kernels, Filomat 30, 1045 (2016).

[43]. Cameron RH and Martin WT, Transformations of weiner integrals under translations, Ann. Math 45, 386 (1944).

[44]. Girsanov IV, On transforming a certain class of stochastic processes by absolutely continuous substitution of measures, Theory Probab. Appl 5, 285 (1960).

[45]. Liu MKP, Hawkins N, Ritchie AJ, Ganusov VV, Whale V, Brackenridge S, Li H, Pavlicek JW, Cai F, Rose-Abrahams M et al. , Vertical T cell immunodominance and epitope entropy determine HIV-1 escape, J. Clin. Invest 123, 380 (2013). [PubMed: 23221345]

[46]. Zanini F, Brodin J, Thebo L, Lanz C, Bratt G, Albert J, and Neher RA, Population genomics of intrapatient HIV-1 evolution, eLife 4, e11282 (2015). [PubMed: 26652000]

[47]. Christensen R, Advanced Linear Modeling (Springer, New York, 2019).

[48]. MacKay DJC, Kay DJCM et al., Information Theory, Inference and Learning Algorithms (Cambridge University Press, Cambridge, UK, 2003).

[49]. Bishop CM and Nasrabadi NM, Pattern Recognition and Machine Learning (Springer, Berlin, 2006), Vol. 4.

[50]. Lin Q and Li C, Kriging based sequence interpolation and probability distribution correction for gaussian wind field data reconstruction, J. Wind Eng. Ind. Aerodyn 205, 104340 (2020).

[51]. Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, Zecchina R, Onuchic JN, Hwa T, and Weigt M, Direct-coupling analysis of residue coevolution captures native contacts across many protein families, Proc. Natl. Acad. Sci. USA 108, E1293 (2011). [PubMed: 22106262]

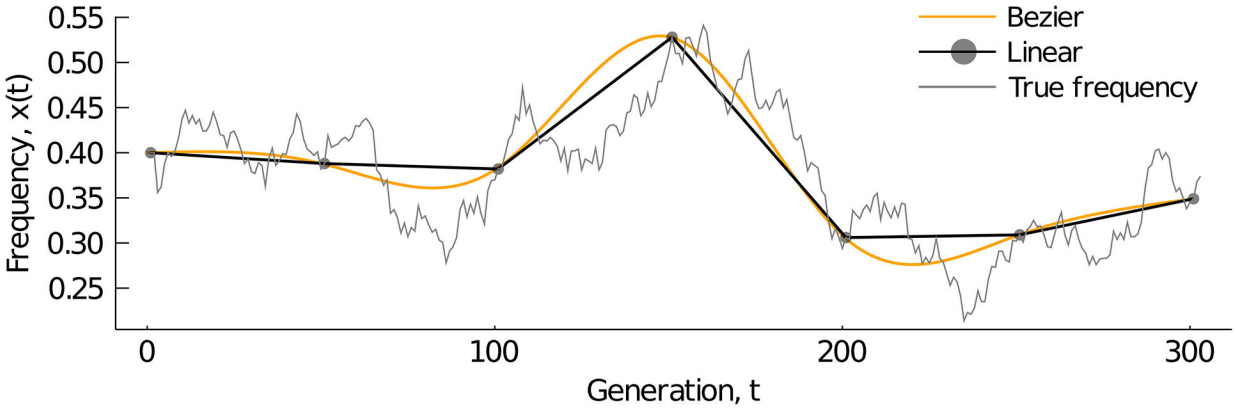

FIG. 1.

Bézier interpolation generates smooth curves. Cubic Bézier curves smoothly interpolate between discretely sampled frequency trajectories generated from a Wright-Fisher model. *Simulation parameters*. $L = 50$ sites, population size $N = 10^3$, mutation rate $\mu = 10^{-3}$, with simulations over $T = 300$ generations. Data points are sampled every 50 generations and interpolated using cubic Bézier and linear interpolation.

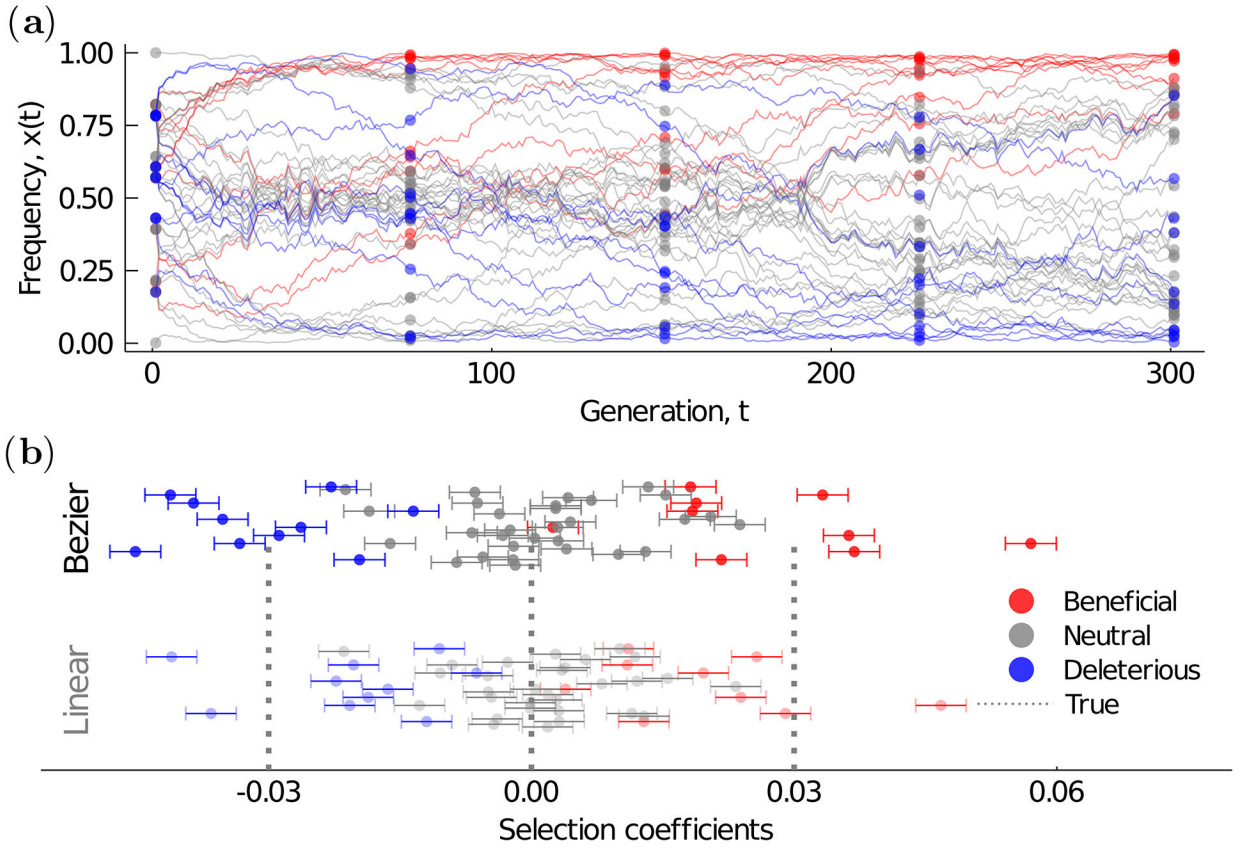

FIG. 2.

Bézier interpolation reduces bias in estimated selection coefficients. (a) Wright-Fisher simulation with selection and mutation. Each trajectory drawn as a solid line is true complete data, and filled circles are a subset of the complete data, which is observed every $\Delta t = 75$ generations and used for selection coefficient prediction. (b) Selection coefficients for the frequency trajectories in (a) were estimated by MPL with Bézier and linear interpolation. MPL with Bézier interpolation greatly reduces estimation bias for inferred selection coefficients when the time interval between sampled observations is large. *Simulation parameters*. $L = 50$ sites with 10 beneficial, 10 deleterious, and 30 neutral mutations with selection coefficients of $s = 0.03$, $s = -0.03$, and $s = 0$, respectively. Other parameters of the WF models are the same as in Fig. 1.

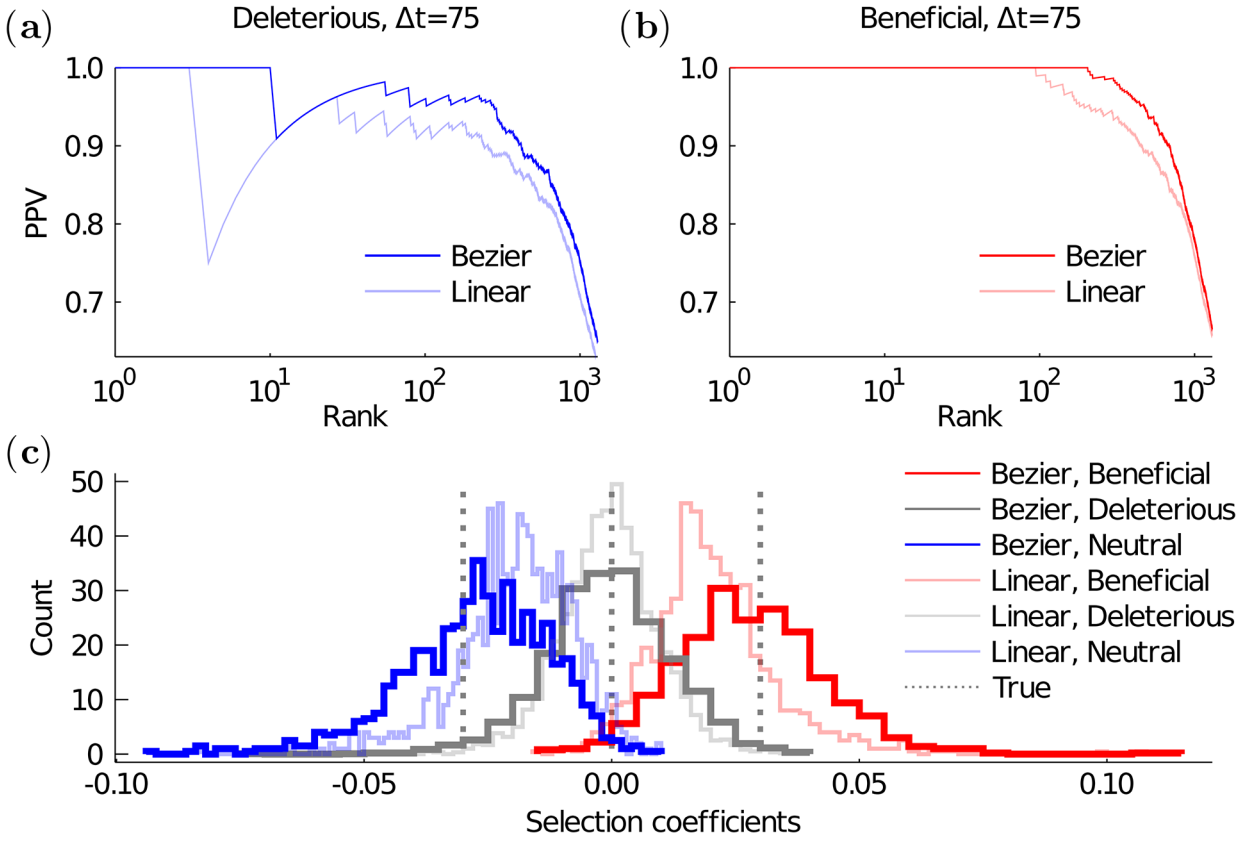

FIG. 3.

MPL with Bézier interpolation improves prediction precision and reduces estimation bias. (a),(b) When the observation time interval is longer ($\Delta t = 75$), the PPV curve for Bézier interpolation is universally higher than the curve for linear interpolation for both deleterious and beneficial cases. (c) The selection coefficient distributions estimated by MPL with linear interpolation visibly shrank toward zero and were biased, while distributions estimated by MPL with Bézeir interpolation did not considerably shrink and have the mean values near the true selection values.

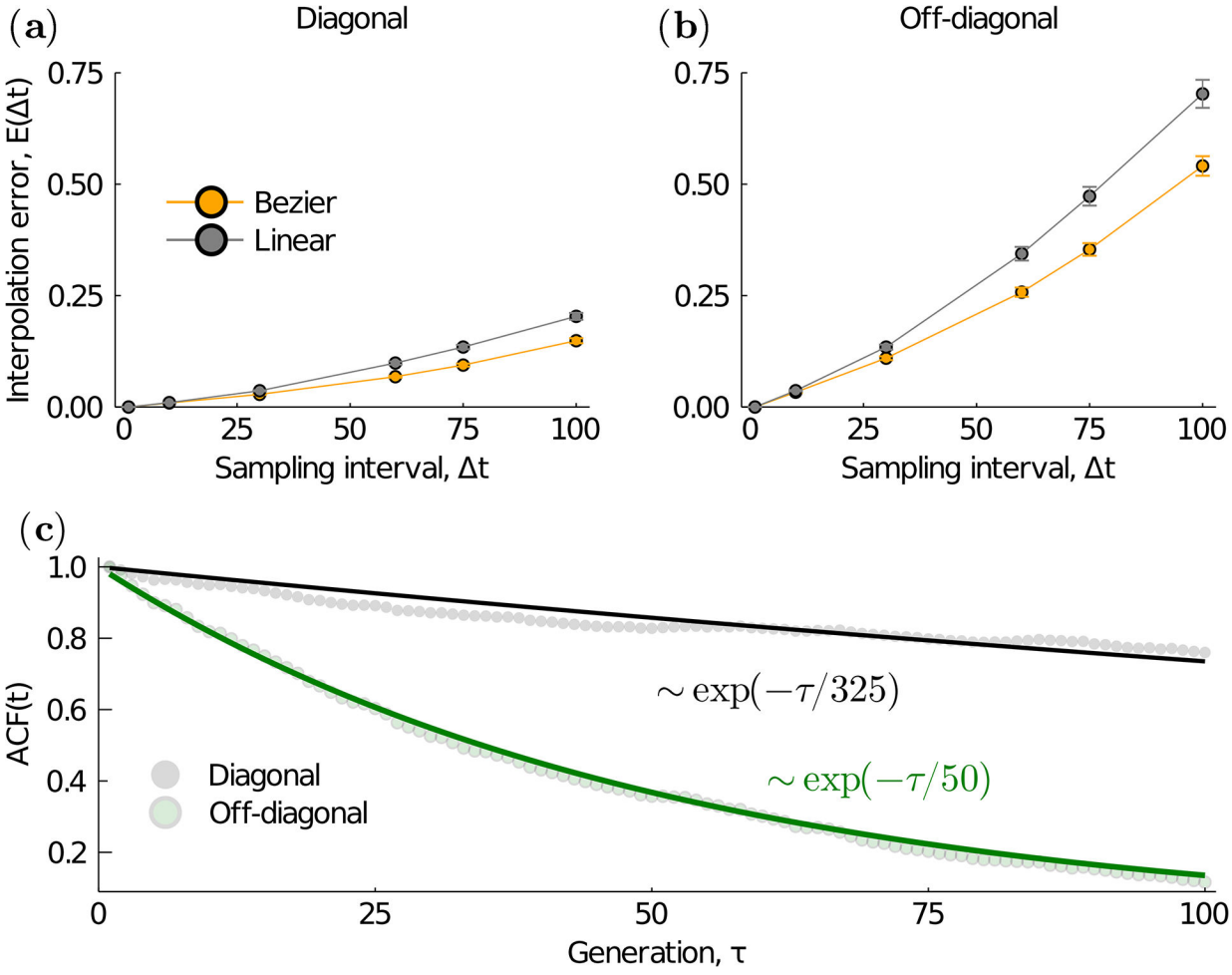

FIG. 4.

Bézier method suppresses interpolation error, especially for off-diagonal pairwise covariances. (a) Sampling time interval dependence for interpolation errors $\mathcal{E}(\Delta t)$ for diagonal covariances and (b) for off-diagonal pairwise covariances. We simulated WF dynamics using the model described in Fig. 2 and generated data sets that evolved to the 300th generation for each trial. For example, when $\Delta t = 100$, results only use data from $t = 0$, 100, 200, and 300. (c) The autocorrelation of off-diagonal covariance elements decays faster than diagonal ones. To simplify the analysis, we evaluated the autocorrelation function from generation $t = 50$. The diagonal autocorrelation shows nonmonotonic decay after long times due to mutant frequencies that approach the frequency boundaries (i.e., 0 and 1).

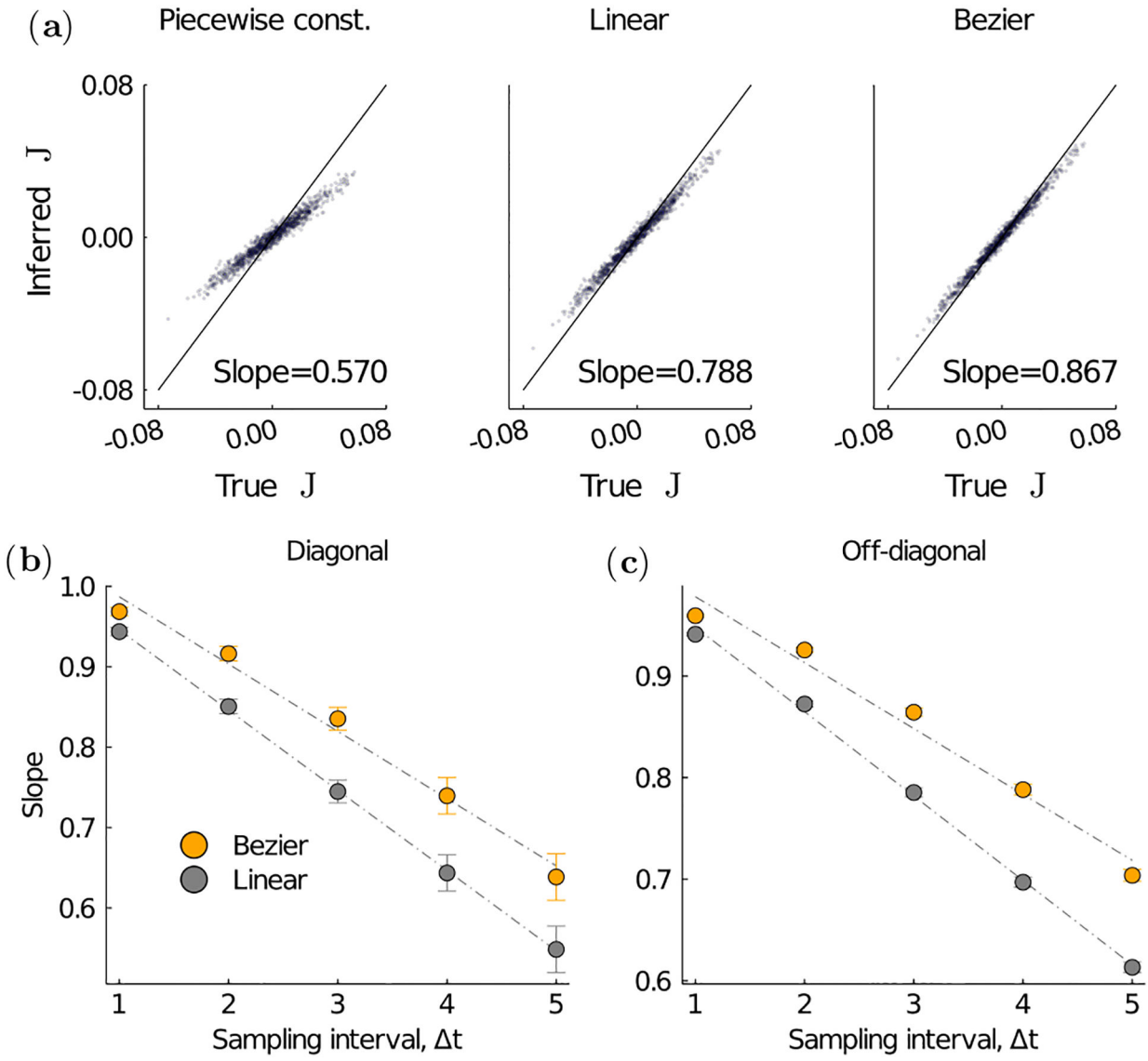

FIG. 5.

Bézier interpolation can improve the inference accuracy of parameters in the OU process. (a) Comparison between true and inferred OU parameters using piecewise constant, linear, and Bézier interpolation. Linear regression slope values are included in each panel. Inferred interaction parameters using Bézier interpolation correspond most closely with the true parameters. (b) Dependence of the slope between true and inferred parameters on the time sampling interval $\Delta t = 1$, shown separately for the (b) diagonal and (c) off-diagonal interaction parameters of the J matrix. In both diagonal and off diagonal, the slope values decrease more gradually with increasing Δt for Bézier interpolation than for linear interpolation.