A Report on "Inference in Molecular Population Genetics" by Stephens and Donnelly (2000)

Reviewer 2

April 07, 2026

isitcredible.com

Disclaimer

This report was generated by Reviewer 2, an automated system that uses large language models to assess academic texts. It did not have any input from a human editor and any claims made in it should be verified by a qualified expert.

> I am wiser than this person; for it is likely that neither of us knows anything fine and good, but he thinks he knows something when he does not know it, whereas I, just as I do not know, do not think I know, either. I seem, then, to be wiser than him in this small way, at least: that what I do not know, I do not think I know, either.
>
> Plato, *The Apology of Socrates*, 21d

> To err is human. All human knowledge is fallible and therefore uncertain. It follows that we must distinguish sharply between truth and certainty. That to err is human means not only that we must constantly struggle against error, but also that, even when we have taken the greatest care, we cannot be completely certain that we have not made a mistake.
>
> Karl Popper, 'Knowledge and the Shaping of Reality'

Overview

Citation: Stephens, M. and Donnelly, P. (2000). Inference in Molecular Population Genetics. *Journal of the Royal Statistical Society Series B* (*Statistical Methodology*). Vol. 62, No. 4, pp. 605–655.

URL: https://academic.oup.com/jrsssb/article/62/4/605/7083252

Abstract Summary: This paper introduces a new importance sampling (*IS*) algorithm for full likelihood-based inference in modern population genetics data, addressing the methodological and computational challenges. The authors compare their new method with existing *IS* and Markov chain Monte Carlo (*MCMC*) algorithms on various genetic examples, demonstrating substantial improvements in efficiency for certain problems.

Key Methodology: Importance Sampling (*IS*), Markov Chain Monte Carlo (*MCMC*), Coalescent theory, Genealogical processes, Simulation studies.

Research Question: How can full likelihood-based inference for modern population genetics data be performed more efficiently and accurately, particularly through improved importance sampling algorithms?

Editor's Note

Reviewer 2 raises valid theoretical concerns, particularly regarding the lack of a finite variance proof for the importance weights and the absence of strict theoretical bounds for the core approximation. However, the reviewer's demand for rigorous mathematical guarantees overlooks the practical reality of computationally intensive population genetics. In complex stochastic processes like the coalescent, exact analytical proofs are frequently intractable. The article's primary contribution is a highly effective practical heuristic. The authors could defend the empirical validation approach, noting that the inability to prove finite variance universally does not negate the algorithm's demonstrated utility, which is a commonly accepted trade-off in Monte Carlo methods. Furthermore, the reviewer's critique of the infinite sites model adaptation mischaracterizes a methodological strength as a confounding factor; the ability to bypass the driving value in this specific model is a feature of the framework's flexibility, not a flaw.

The reviewer also challenges the empirical comparisons, suggesting that competing Markov chain Monte Carlo methods were evaluated using suboptimal default parameters. The authors could counter this by emphasizing that default settings represent the standard use case for most practitioners, making it a highly relevant real-world benchmark. Nevertheless, to fortify the article against claims of unfairness, the authors might slightly soften the rhetoric regarding absolute superiority over competing approaches. Instead of broad assertions of dominance, the text could highlight that the new importance sampling method is a highly competitive alternative that excels particularly in constrained genetic structures. The authors could also explicitly acknowledge the shared vulnerability to the driving value problem, framing it as a universal challenge in the field rather than a unique failing of competing methods.

Several specific adjustments will immediately improve the article's rigor. The alge-

braic error in the derivation of the backward transition rates needs to be corrected by replacing the total sample size with the current number of lineages in the denominator (p. 615). Additionally, the authors could provide a brief justification for the state space truncation boundary chosen for the microsatellite data, perhaps noting why alleles beyond this length have negligible probability (p. 624). Regarding the tension between diagnostic claims and empirical evidence, the authors could revise the conclusion to present a more balanced view of convergence monitoring, acknowledging that while importance sampling diagnostics can be more straightforward than high-dimensional Markov chain monitoring, severe weight skewness remains a formidable challenge.

If conducting further standardized computational benchmarking or tuning the competitor algorithms proves too time-consuming, a pragmatic alternative is to systematically temper the article's most ambitious claims. The authors could remove assertions of orders-of-magnitude efficiency gains over all existing methods, instead focusing the narrative strictly on the improvements over the original Griffiths–Tavaré scheme. By prominently acknowledging the limitations regarding finite variance, the driving value dependency, and the reliance on empirical convergence benchmarks, the authors can preempt further theoretical objections and ensure the article serves as a valuable, albeit heuristic, computational advancement.

Summary

Is It Credible?

The article by Stephens and Donnelly introduces a novel importance sampling (IS) algorithm designed to tackle the computationally intensive problem of full likelihood-based inference in molecular population genetics. The authors claim that their new method, which is based on an approximation of the optimal IS proposal distribution, "substantially outperforms existing IS algorithms, with efficiency typically improved by several orders of magnitude" (p. 605). Furthermore, they assert that their approach "compares favourably with existing MCMC methods in some problems" and suggest that detecting convergence issues in IS is generally easier than monitoring Markov chain convergence in high-dimensional spaces (pp. 605, 630). While the article provides a valuable theoretical characterization of the optimal proposal distribution and introduces a clever practical heuristic, the sweeping claims regarding its comparative efficiency and diagnostic reliability are not entirely supported by the evidence presented.

The foundational mechanism of the new algorithm relies on approximating the true conditional sampling probability to construct an efficient proposal distribution. However, the theoretical justification for this core approximation is limited. The authors prove it is exact only in highly specific scenarios, such as parent-independent mutation models or when the sample size is one (p. 612). For more general and realistic genetic models, the quality of the approximation lacks theoretical bounds and relies heavily on numerical approximations, which the authors transparently acknowledge can be "rather rough" in some cases (p. 634). The utility of the method therefore rests almost entirely on empirical demonstrations rather than guaranteed analytical properties.

A critical vulnerability in the statistical validity of the method is the absence of a

proof that the importance weights possess a finite variance. The central limit theorem, which is necessary to justify the standard errors and confidence intervals of the likelihood estimates, requires finite variance. The authors openly admit that they "could not prove finiteness of the variance of our weights, except in the special case of the infinite sites model" (p. 628). Without this guarantee, the sample variance can severely underestimate the true uncertainty of the estimates. This theoretical gap manifests in the article's own empirical results: when testing the method on simulated data with a high mutation rate, the authors found that the standard error "severely under-estimates the standard deviation" (p. 623). This finding directly undermines the authors' broader claim that detecting pitfalls in IS is generally an easier problem than monitoring MCMC convergence.

The empirical comparisons used to establish the algorithm's superiority also warrant careful interpretation. The claim of "several orders of magnitude" improvement is primarily based on comparisons with the original Griffiths–Tavaré scheme, which serves as a relatively weak baseline. When comparing their method to competing MCMC approaches, the authors sometimes evaluate those competitors using default parameters that may be sub-optimal for the specific problem, or on datasets inherently ill-suited to the competitor's search strategy (p. 625). Additionally, the impressive efficiency gains reported for the infinite sites model benefit from a model-specific adaptation that bypasses the need to specify a "driving value" for the mutation rate (p. 627). Because the general framework still requires a fixed driving value—a structural limitation the authors explicitly critique in MCMC methods but acknowledge sharing—the exceptional performance on the infinite sites model cannot be generalized to the broader algorithm (p. 630).

Ultimately, the article makes a significant conceptual contribution by formalizing the optimal proposal distribution for coalescent histories and offering a creative, computationally tractable approximation. However, the headline claims of vast efficiency improvements and diagnostic advantages over MCMC methods are overstated. The

lack of a finite variance proof, the reliance on unverified benchmarks, and the confounding factors in the empirical comparisons suggest that while the method is a useful addition to the computational toolbox, its absolute superiority remains unestablished.

The Bottom Line

The article presents an innovative importance sampling framework that provides a useful heuristic for estimating likelihoods in population genetics, successfully characterizing the optimal proposal distribution in theory. However, its headline claims of orders-of-magnitude efficiency gains and diagnostic superiority over MCMC methods are not fully supported by the theoretical and empirical evidence. The lack of a proof for finite variance in the importance weights, combined with empirical comparisons against potentially sub-optimally tuned baselines, suggests that while the method is a significant conceptual step, its practical dominance is less certain than claimed.

Potential Issues

Absence of proof for finite variance of importance weights: The statistical validity of the standard errors and confidence intervals for the likelihood estimates produced by the importance sampling (IS) method relies on the central limit theorem, which requires the importance weights to have a finite variance. The authors acknowledge a significant theoretical gap regarding this requirement. On page 628, they state: "For example, we could not prove finiteness of the variance of our weights, except in the special case of the infinite sites model where the number of possible histories is finite." Without a guaranteed finite variance, the sample variance of the weights can severely underestimate the true variance, meaning the reported standard errors (such as those in Table 1) may not be theoretically justified in the general case. The authors are transparent about this limitation and discuss potential diagnostic workarounds, but the foundational proof remains absent.

Theoretical bounds of the core approximation: The efficiency of the proposed algorithm hinges entirely on how well the proposed distribution $\hat{\pi}(\cdot|A_n)$ approximates the true conditional sampling probability $\pi(\cdot|A_n)$. The article provides limited theoretical justification for this approximation, proving it is exact only in the special case of parent-independent mutation (PIM) and the trivial case of $n = 1$ (p. 612). For general cases, the quality of the approximation is not theoretically bounded. Furthermore, for sequence and multilocus models, the practical implementation relies on a further numerical approximation of $\hat{\pi}$ using Gaussian quadrature. The authors explicitly acknowledge this, noting in Appendix A (p. 634) that "in some cases the approximation to $\hat{\pi}(\cdot|\cdot)$ obtained through this procedure is rather rough," though they maintain it still yields a valid IS function. The method relies heavily on empirical demonstrations to validate the utility of these approximations.

Shared vulnerability to the driving value problem: The authors critique competing Markov chain Monte Carlo (MCMC) methods for fixing a "driving value" θ_0 to

estimate relative likelihood surfaces, arguing this leads to underestimation of likelihoods away from θ_0 and artificially peaked curves (p. 630). However, the authors' own IS method employs a similar strategy, generating samples from a single proposal distribution Q_{θ_0} based on a fixed driving value. The authors explicitly recognize this structural irony, stating on page 630: "In principle IS methods based on a driving value of θ will tend to share this undesirable property, as designing a single IS function Q_{θ_0} which is universally efficient for all θ may be extremely challenging." While they note this did not cause major problems in their specific examples, the conceptual flaw they identify in competing methods is inherently present in their own.

Tension between diagnostic claims and empirical evidence: In the conclusion, the authors assert that detecting the pitfalls of highly skewed importance weights "appears to be an easier problem (in most contexts) than monitoring the convergence of a Markov chain in such a high dimensional space" (p. 630). This optimistic claim is somewhat contradicted by the article's own empirical findings. When testing their method on simulated data with a high mutation rate ($\theta = 15.0$), they admit that standard diagnostics failed: "the change in the estimated standard error is less than a factor of 2, indicating that (at least for the short run) the standard error severely under-estimates the standard deviation in this case" (p. 623). The caption for Figure 12 (p. 651) further notes that the standard deviation is "even more substantially underestimated." While the authors dedicate Section 6.4 to discussing these genuine diagnostic difficulties, the assertion that IS diagnostics are generally "easier" than MCMC convergence monitoring lacks strong support within the presented data.

Empirical evaluation against sub-optimal competitor implementations: The article claims the new IS method is competitive with or superior to existing MCMC methods, but the empirical comparisons are sometimes made against MCMC programs running on default or potentially sub-optimal settings. For example, when comparing against the micsat program, the authors acknowledge: "...there are many ways

in which our use of the MCMC scheme could be improved (for example, the parameters of the MCMC scheme could be tuned to achieve better mixing over θ; we used the default values)" (p. 625). Similarly, the authors attribute the poor performance of the Fluctuate MCMC method in Figure 2 to its use of a fixed driving value (p. 630), but they do not test the alternative explanation offered by the program's developers in the discussion (p. 642) that the specific dataset used (short sequences) is inherently ill-suited to that MCMC's search strategy. Comparing a newly developed, tuned method against competitors running on default settings or ill-suited data can inflate the perceived relative efficiency of the new algorithm.

Confounding factors in infinite sites model evaluation: The article reports significant efficiency gains for the infinite sites model. However, the implementation for this specific model involves a major adaptation that is not part of the general theoretical framework. The authors note that due to technical challenges, they "adapt our earlier approach to this context by analogy" (p. 627). Crucially, this adapted procedure defines an IS function that is "independent of θ, removing the need to specify a driving value" (p. 627). While this is a clever model-specific simplification, it means the strong performance observed for the infinite sites model cannot be solely attributed to the general approximation framework, as it benefits from bypassing the driving value problem entirely.

Reliance on unverified benchmarks and indirect validation: To assess the accuracy of the proposed method on short runs, the authors compare the results to a benchmark generated from a very long run (10 million samples) of their own method, treating this as the "accurate" likelihood (p. 622). While this is a common practice when analytical solutions are unavailable, it relies on the untestable assumption that the 10-million-sample run has fully converged, which is not guaranteed given the severe skewness issues documented elsewhere in the article. Furthermore, the article evaluates the end-to-end performance of the algorithm but omits a direct, quantitative validation of the core approximation $\hat{\pi}(\cdot|A_n)$ against true conditional probabili-

ties on a tractable "toy problem." Such an isolated validation would have provided clearer insight into the error introduced specifically by the foundational heuristic.

Incorrect index in derivation of backward mutation rates: In the proof of Theorem 1 on page 615, there is an algebraic error in the derivation of the backward transition probabilities. The text defines the state of the ancestry at time t as consisting of k lineages: $A_k(t) = (\alpha_1, \dots, \alpha_{k-1}, \alpha)$. The derivation then calculates the probability of a mutation occurring on the k-th lineage backwards in time. However, in the subsequent equation representing the conditional probability ratio, the denominator incorrectly uses the total sample size n rather than the current number of lineages k. The published expression is $\frac{\pi(\alpha_1,\dots,\alpha_{n-1},\beta)\delta\theta P_{\beta\alpha}/2}{\pi(\alpha_1,\dots,\alpha_{n-1},\alpha)} + o(\delta)$. Because the state is defined over k lineages, this expression should be $\frac{\pi(\alpha_1,\dots,\alpha_{k-1},\beta)\delta\theta P_{\beta\alpha}/2}{\pi(\alpha_1,\dots,\alpha_{k-1},\alpha)} + o(\delta)$. While the surrounding text makes the intended logic clear, the formal mathematical expression is structurally incorrect.

Scope limitations and unverified state space truncation: The method is primarily developed and demonstrated on highly simplified biological models. The authors explicitly state they are considering "the simplest demographic and genetic scenario" (p. 607), which assumes a single randomly mating population of constant size with no recombination, and assumes the mutation transition matrix P is fixed and known (p. 610). While they discuss potential extensions (p. 631), these assumptions limit the immediate applicability of the method to more complex, realistic datasets. Additionally, for the microsatellite data analysis, the authors truncate the infinite state space of possible allele lengths to $\{0, 1, \dots, 19\}$, asserting this "will make little difference" (p. 624). The article omits sensitivity analyses to empirically demonstrate that this specific truncation boundary does not introduce bias.

Presentation of relative performance and computational costs: The narrative framing of the algorithm's efficiency is sometimes difficult to interpret cleanly. The headline claim of "several orders of magnitude" improvement (p. 605) is illustrated by comparing the new method to the original Griffiths–Tavaré (GT) results (p. 622).

However, the authors also transparently report that a modified GT scheme (the SEQUENCE program) produced a result statistically indistinguishable from their own in a similar timeframe (p. 622), which may make the comparison to the weakest baseline less representative of the method's general superiority. Furthermore, the reporting of computational costs is inconsistent across examples, mixing CPU times, iteration counts, and sample sizes (e.g., comparing 30 seconds of IS sampling to 1.5 hours of MCMC sampling on p. 626), without a standardized metric such as time-to-target-variance. The authors acknowledge the difficulty of making fair comparisons (p. 620), but the lack of standardized benchmarking complicates the interpretation of the efficiency claims.

Future Research

Theoretical bounds on importance weights: Future research should focus on establishing the mathematical conditions under which the importance weights in this class of IS algorithms possess finite variance. If analytical proofs remain intractable for complex demographic and genetic models, developing robust, extreme-value-theory-based diagnostics to reliably detect infinite variance in finite samples would be a critical step for validating the standard errors of these likelihood estimates.

Standardized computational benchmarking: To clarify the relative efficiency of IS versus MCMC methods, future studies should evaluate these algorithms using standardized metrics, such as time-to-target-variance, across a diverse suite of genetic models. Crucially, these comparisons must ensure that all competing algorithms are optimally tuned for the specific datasets rather than relying on default parameters, providing a fairer assessment of their fundamental computational limits.

Direct validation of the approximation heuristic: The core approximation of the conditional sampling probabilities should be validated in isolation from the full IS algorithm. By testing this heuristic against exactly calculable conditional probabilities in tractable toy models, researchers could quantify the specific error introduced by the approximation and identify the precise genetic scenarios where the heuristic breaks down.

Copyediting

The manuscript presents an innovative and highly useful importance sampling framework for population genetics. The theoretical characterization of the optimal proposal distribution and the practical heuristics are strong contributions to the field. However, to ensure the manuscript is as rigorous and unassailable as possible, some of the strongest comparative claims regarding efficiency and diagnostic superiority should be tempered. Additionally, a mathematical typo requires correction, and certain methodological choices, such as state space truncation and the reporting of computational costs, would benefit from further clarification.

- **p. 605** "Our approach substantially outperforms existing IS algorithms, with efficiency typically improved by several orders of magnitude." This claim is quite strong and primarily reflects improvements over the baseline Griffiths–Tavaré scheme rather than all existing methods. Softening this language will preempt critiques about the generalizability of these efficiency gains. Consider revising to: "Our approach offers substantial improvements over the original Griffiths–Tavaré IS algorithm, with efficiency typically improved by several orders of magnitude in our test cases."
- **p. 615** "$= \frac{\pi(\alpha_1,\ldots,\alpha_{n-1},\beta)\delta\theta P_{\beta\alpha}/2}{\pi(\alpha_1,\ldots,\alpha_{n-1},\alpha)} + o(\delta)$" There is an algebraic error in this derivation. Because the state of the ancestry at time t is defined over k lineages, the subscripts in the numerator and denominator of this conditional probability ratio should reflect $k-1$ rather than the total sample size $n-1$. Consider changing the subscripts to correct the mathematical expression: "$= \frac{\pi(\alpha_1,\ldots,\alpha_{k-1},\beta)\delta\theta P_{\beta\alpha}/2}{\pi(\alpha_1,\ldots,\alpha_{k-1},\alpha)} + o(\delta)$".
- **pp. 622-626** Throughout Section 5, computational costs are reported using a mix of metrics, such as "1 million samples took 72 h" (p. 622), "took 23 min" (p. 622), and "each run took about 30 s" (p. 626). Mixing central processor

unit times, iteration counts, and sample sizes makes it difficult for readers to cleanly interpret the relative efficiency of the algorithms. Consider standardizing the framing of computational costs across these examples, clearly distinguishing between time, iterations, and sample sizes to prevent misleading efficiency equivalencies.

- **p. 624** "...truncating the type space E to $\{0, 1, \ldots, 19\}$ by insisting that all mutations to alleles of length 0 or 19 involve the gain or loss respectively of a single repeat. This truncation will make little difference to the likelihood of samples whose allele lengths are not too close to these boundaries." While intuitively reasonable, the lack of a brief justification for why this specific boundary does not introduce meaningful bias leaves the methodological choice open to criticism. Consider adding a brief explanatory sentence, such as: "Because the probability of observing alleles beyond this length under our tested mutation rates is negligible, this specific boundary does not introduce meaningful bias into the likelihood estimates."

- **p. 630** "MCMC methods which fix θ at a 'driving value' θ_0, and use IS to estimate the relative likelihood, appear to make things unnecessarily difficult for themselves..." The critique of Markov chain Monte Carlo methods using a driving value could be perceived as overly harsh given that the proposed importance sampling method shares this exact vulnerability, as acknowledged later in the same paragraph. Consider softening the critique here and explicitly framing the driving value dependency as a universal challenge in the field rather than a unique failing of competing methods.

- **p. 630** "...detecting this pitfall once we are aware of it appears to be an easier problem (in most contexts) than monitoring the convergence of a Markov chain in such a high dimensional space." This optimistic claim is somewhat in tension with the empirical difficulties documented earlier in the text (e.g., on

p. 623, where the standard error severely underestimates the standard deviation). Consider revising to present a more balanced view of convergence monitoring. For example: "...detecting this pitfall once we are aware of it can sometimes be more straightforward than monitoring the convergence of a Markov chain in such a high dimensional space, though severe weight skewness remains a formidable diagnostic challenge for IS methods."

Proofreading

No issues found.

isitcredible.com