Study Design Reviewer Feedback

The manuscript presents a novel proposal distribution for Importance Sampling (IS) in population genetics, demonstrating significant efficiency gains over existing IS and MCMC methods in specific contexts. To strengthen the validity of the comparative claims and ensure the robustness of the proposed method, I recommend more rigorous benchmarking against MCMC controls and a deeper validation of the state-space truncation used in the microsatellite analysis.

Comments

1. **Confounding in MCMC Comparisons due to Parameter Tuning:** In Section 5 (Applications), particularly the comparison with `micsat` (Wilson and Balding) and `Fluctuate` (Kuhner et al.), the manuscript notes that default parameters were used for the MCMC algorithms (e.g., "we used the default values"). This introduces a confounding variable: the level of optimization applied to the algorithm. The observed superiority of the proposed IS method could be attributed to a lack of tuning in the MCMC competitor rather than inherent algorithmic superiority. I recommend performing a "best-effort" tuning of the MCMC parameters (e.g., heating schemes or mixing parameters) to establish a valid baseline for the efficiency comparison.

2. **Validation of State Space Truncation:** In Section 5.4 (Microsatellites), the analysis centers the sample distribution and truncates the type space E to $\{0, \ldots, 19\}$. The authors assert this "will make little difference to the likelihood," but no data is presented to support this assumption. Boundary effects in stepwise mutation models can significantly bias likelihood estimates if the probability mass at the boundaries is non-negligible. I recommend including a control experiment where the boundaries are expanded (e.g., to $\{0, \ldots, 40\}$) to quantitatively verify that the truncation does not alter the likelihood estimates for the datasets analyzed.

3. **Metrics for Algorithmic Efficiency:** throughout the results (e.g., Figure 3 caption, Table 1), efficiency comparisons often rely on the number of samples/iterations (e.g., comparing 10,000 IS samples to 50,000 MCMC iterations). This comparison is difficult to interpret without explicit data on the computational cost (CPU time) per independent sample for the new method versus the MCMC steps. To support the claim of "orders of magnitude" improvement, the authors should standardize the comparison by plotting the variance of the estimator against wall-clock time for both methods.

4. **Diagnostics for Importance Weight Distributions:** The manuscript relies heavily on standard errors (SE) to demonstrate the accuracy of the likelihood estimates (Table 1). However, as noted in the discussion of the Griffiths-Tavaré method, IS estimators can have infinite variance or extremely skewed weight distributions where the Central Limit Theorem approximation for SE fails. I recommend reporting the Effective Sample Size (ESS) or the ratio of the maximum weight to the sum of weights for the proposed method. This is necessary to prove that the improved performance is due to a better proposal distribution and not simply a failure to encounter high-weight rare events during the simulation.

5. **Baseline for "Constrained" vs. "Unconstrained" Problems:** The authors suggest that their IS method performs favorably in "constrained" problems but perhaps less so in "less constrained" problems compared to MCMC. This distinction is qualitative. I suggest a simulation experiment that systematically varies the mutation parameter θ (as a proxy for constraint/tree depth) to identify the specific crossover point where MCMC becomes more efficient than the proposed IS method. This would provide a clearer guideline for when researchers should choose one method over the other.
6. **Representative Qualitative Data:** In Figure 2, the likelihood surfaces are shown. For Figure 2(c), the `Fluctuate` program results (dashed lines) show significant variance and failure to converge to the peak. While this supports the author's claim, it is important to clarify if these represent single runs or averages. If they are single runs, they may be outliers. I recommend plotting the mean likelihood surface with confidence bands calculated from multiple independent runs of the MCMC software to ensure Figure 2(c) is a representative characterization of the MCMC performance.

Reproducibility Reviewer Feedback

The manuscript presents a novel Importance Sampling (IS) approach for likelihood-based inference in population genetics, offering a theoretically grounded approximation for the optimal proposal distribution. While the mathematical derivation is rigorous and the potential for variance reduction is clear, the empirical benchmarking against existing methods (both IS and MCMC) requires stricter statistical controls and more transparent reporting of computational costs to fully support the claims of superior efficiency.

Comments

1. **Normalization of Computational Efficiency Metrics** The manuscript frequently claims substantial improvements in efficiency compared to the Griffiths-Tavaré (GT) scheme and MCMC methods. However, the comparisons in Section 5 (e.g., Table 1 and Figure 2) rely primarily on the number of iterations or samples. This is a potentially misleading metric because the computational cost per iteration likely differs between methods. Specifically, the proposed method involves calculating $\hat{\pi}(\cdot|A_n)$, which appears more computationally intensive than the proposal steps in the GT scheme. To support the claim of practical efficiency, the authors must report the wall-clock time or CPU operations required to achieve a specific variance or Effective Sample Size (ESS) for both the proposed method and the benchmarks.

2. **Rigorous Benchmarking against MCMC** In Sections 5.3 and 5.4, the authors compare their IS method against MCMC implementations (e.g., `Fluctuate` and `micsat`). The text notes instances where MCMC produces inaccurate estimates or fails to converge (e.g., Figure 2c). However, MCMC performance is highly sensitive to tuning parameters (chain length, burn-in, heating, proposal step sizes). To ensure a fair comparison and avoid a "straw man" argument, the authors must provide evidence that the MCMC algorithms were optimally tuned. Please report standard convergence diagnostics for the MCMC runs (e.g., Effective Sample Size, Gelman-Rubin statistics, or autocorrelation times) to demonstrate that the poor performance of MCMC was inherent to the method and not a result of insufficient sampling or poor tuning.

3. **Validation of "True" Likelihood Values** In the simulation studies (e.g., Table 1), the authors compare estimates from different methods. However, for the cases where θ is large or the model is complex, the "true" likelihood value is not analytically available. The manuscript implies that the long-run estimate from the proposed Q^{SD} method is treated as the ground truth. This introduces a risk of circularity if the proposed method has any unrecognized bias. The authors should explicitly state how the "true" values were established. If possible, validation against a completely independent method (e.g., a massive grid search on a very small dataset or rejection sampling) would strengthen the validation.

4. **Reliability of Standard Error Estimates** The authors correctly identify that the GT method underestimates the standard error (SE) when important weights are missed (Section 5.2). However, the manuscript should address whether the proposed Q^{SD} method is immune to this issue in higher-dimensional parameter spaces or with larger datasets. The authors should provide a diagnostic assessment (such as the distribution of importance weights or the effective sample size of the weights) to demonstrate that the variance of the importance weights in their method is finite and well-behaved for the reported examples.

5. **Reproducibility of the Proposal Approximation** The core of the method relies on the approximation $\hat{\pi}(\cdot|A_n)$. While the mathematical properties are described in the Appendix, the algorithmic implementation details necessary for reproducibility are sparse. Specifically, for the "stepwise mutation model" and "infinite sites" variations, the handling of edge cases and the specific numerical methods used to solve the associated linear systems (or approximations thereof) should be detailed. I recommend providing a pseudocode algorithm or making the source code available to ensure that the specific implementation of the proposal distribution can be replicated.

6. **Clarification of Sample Sizes in Figures** In Figure 2, the caption mentions "10,000 samples" for the IS function and "long chains of 50,000 iterations" for MCMC. Later, the text mentions "short chains of 10,000 iterations." The reporting of sample sizes across the text and figure captions is occasionally inconsistent or difficult to parse. Please standardize the reporting of n (samples/iterations) across all figures and ensure that the distinction between burn-in, thinning, and recorded samples is explicitly stated for all MCMC comparisons.

Limitations & Context Reviewer Feedback

The manuscript presents a significant advancement in Importance Sampling (IS) for population genetics by approximating the optimal proposal distribution, demonstrating clear efficiency gains over existing methods. However, to fully contextualize the utility of this approach, the authors must more rigorously address the scalability of the method to larger sample sizes and its robustness under demographic scenarios that deviate from the stationary assumptions used to derive the proposal distribution.

Comments

1. **Scalability and Computational Complexity:** The empirical evaluations provided in the manuscript focus on relatively small sample sizes (e.g., $n = 20$ to 50). While the reduction in variance is impressive, the calculation of the proposal distribution $\hat{\pi}(\cdot|A_n)$ appears to involve computationally intensive steps at every addition to the genealogy. The authors should explicitly discuss the computational complexity of their algorithm as a function of n. Specifically, is there a crossover point where the computational cost of constructing the high-efficiency proposal outweighs the gain in variance reduction compared to simpler MCMC updates?

2. **Limitations in Non-Stationary Demographics:** The derivation of the approximate conditional distribution $\hat{\pi}$ relies heavily on properties of the coalescent at stationarity (constant population size). However, a primary goal of molecular population genetics is often to infer demographic changes (e.g., bottlenecks, expansions). The authors should discuss the limitations of applying a proposal distribution derived from stationary assumptions to data generated under strong demographic changes. Does the efficiency of the IS scheme degrade significantly if the true demographic history deviates sharply from the constant-size assumption implicit in the proposal?

3. **Robustness of the "Driving Value" (θ_0):** While the manuscript demonstrates that the proposed method is more robust to the choice of the driving value θ_0 than the Griffiths-Tavaré scheme, it remains an Importance Sampling method subject to weight degeneracy. The authors should discuss the limitations of this robustness. Specifically, are there regimes (e.g., highly multimodal likelihood surfaces or extremely high mutation rates) where a poor choice of θ_0 still leads to effective sample sizes close to zero, despite the improved proposal distribution?

4. **Definition of "Constrained" vs. "Unconstrained" Spaces:** The discussion distinguishes between the performance of IS and MCMC based on whether the genetic structure is "constrained" or "unconstrained." This distinction is currently qualitative. To improve generalizability, the authors should attempt to operationalize these terms. Can the authors provide specific metrics (e.g., ratio of segregating sites to sample size, or specific parameter ranges of θ) that would allow a user to predict *a priori* whether IS or MCMC would be the superior inference tool for a given dataset?

5. **Handling of Recombination:** The current derivation and examples assume a tree-like genealogy (no recombination). Given that recombination is a fundamental feature of nuclear DNA in many organisms, the omission of this factor limits the generalizability of the method. The authors should discuss the theoretical hurdles involved in extending the proposal distribution $\hat{\pi}$ to Ancestral Recombination Graphs (ARGs). Would the "look-ahead" property of the proposal distribution remain tractable if lineages can split (recombination) as well as coalesce?

6. **Diagnostics for Convergence:** The manuscript uses standard errors and comparison between runs to assess accuracy. However, IS weights can have infinite variance, making standard error estimates unreliable. The authors should discuss more rigorous diagnostics for the reliability of the IS estimates. For instance, how should users detect if the proposal distribution has failed to cover a significant mode of the posterior, particularly in high-dimensional genealogy spaces where standard diagnostics might be misleading?

7. **Approximation Validity for Complex Mutation Models:** The approximation $\hat{\pi}$ is shown to be exact for Parent-Independent Mutation (PIM) models but is an approximation for others (e.g., stepwise mutation). The authors should discuss the limitations of this approximation for more complex, biologically realistic mutation models (e.g., those with bias or heterogeneity in mutation rates across sites). At what point does the divergence between the true conditional distribution and the approximation $\hat{\pi}$ render the importance weights too variable for practical inference?