Robust Automated Equilibration Detection for Molecular Simulation

Finlay Clark[1]; Graeme Robb[2]; Daniel Cole[3]; Julien Michel[1]

[1]*EaStCHEM School of Chemistry, University of Edinburgh, David Brewster Road, Edinburgh EH9 3FJ, United Kingdom*

[2]*Oncology R&D, AstraZeneca, Cambridge CB4 0WG, United Kingdom*

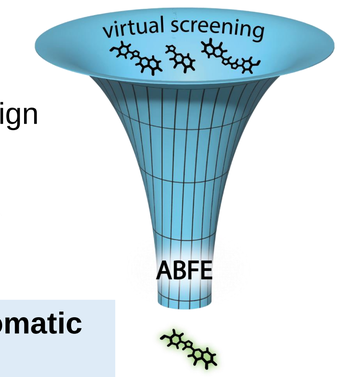
[3]*School of Natural and Environmental Sciences, Newcastle University, Newcastle upon Tyne NE1 7RU, United Kingdom*

Corresponding author: finlay.clark@ed.ac.uk

Molecular simulations are often subject to a large initial bias

Molecular simulations are:

- Widely used to inform structure-based drug design (e.g. alchemical free energy calculations)
- Often subject to initial bias – must truncate data from the start of the simulation to reduce bias.



Given simulation data, what is the best automatic method for selecting the truncation point?

Popular automated bias removal methods minimise the marginal standard error

General Algorithm

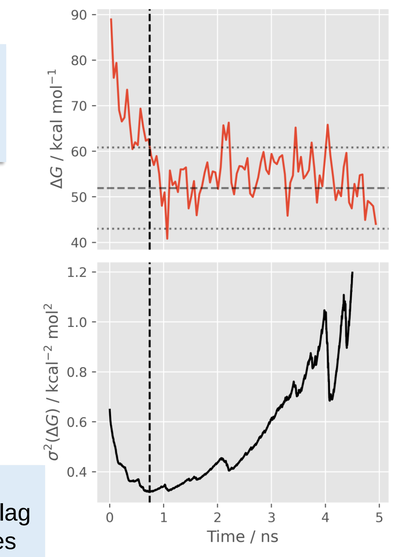
- Discard increasing amounts of data from start
- Calculate SEM of remaining data each time
- Minimum SEM truncation point is selected

Examples

- White's Marginal Standard Error Rule (MSER)[1] is popular in operational research
- Chodera's method[2] is closely related and popular in molecular simulation

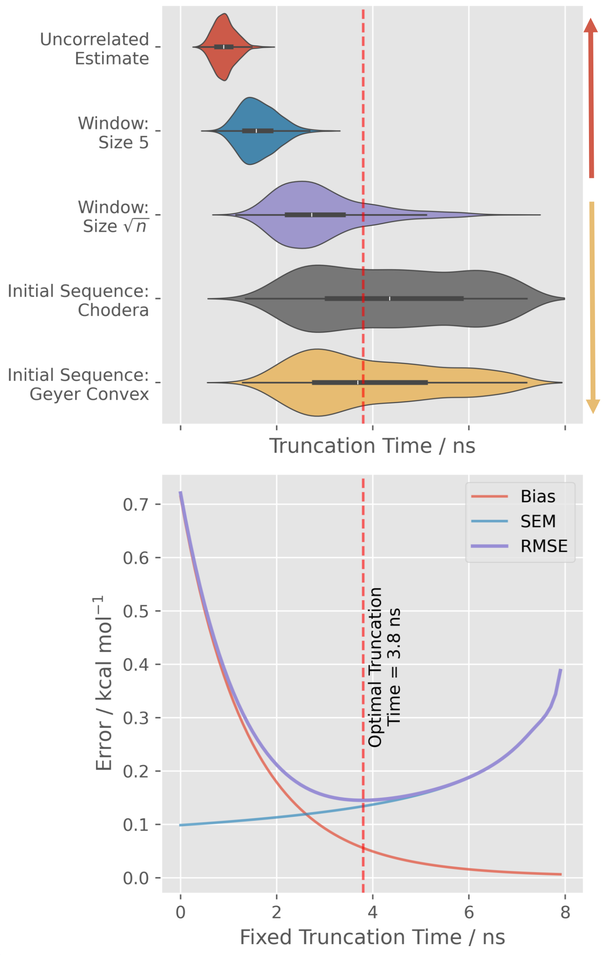
Methods differ in the way the variance is calculated.

Estimated variance $\hat{\sigma}_n^2 = \sum_{t=-\infty}^{\infty} \hat{\gamma}_{n,t}$ Estimated autocovariance with lag time *t* and *n* samples



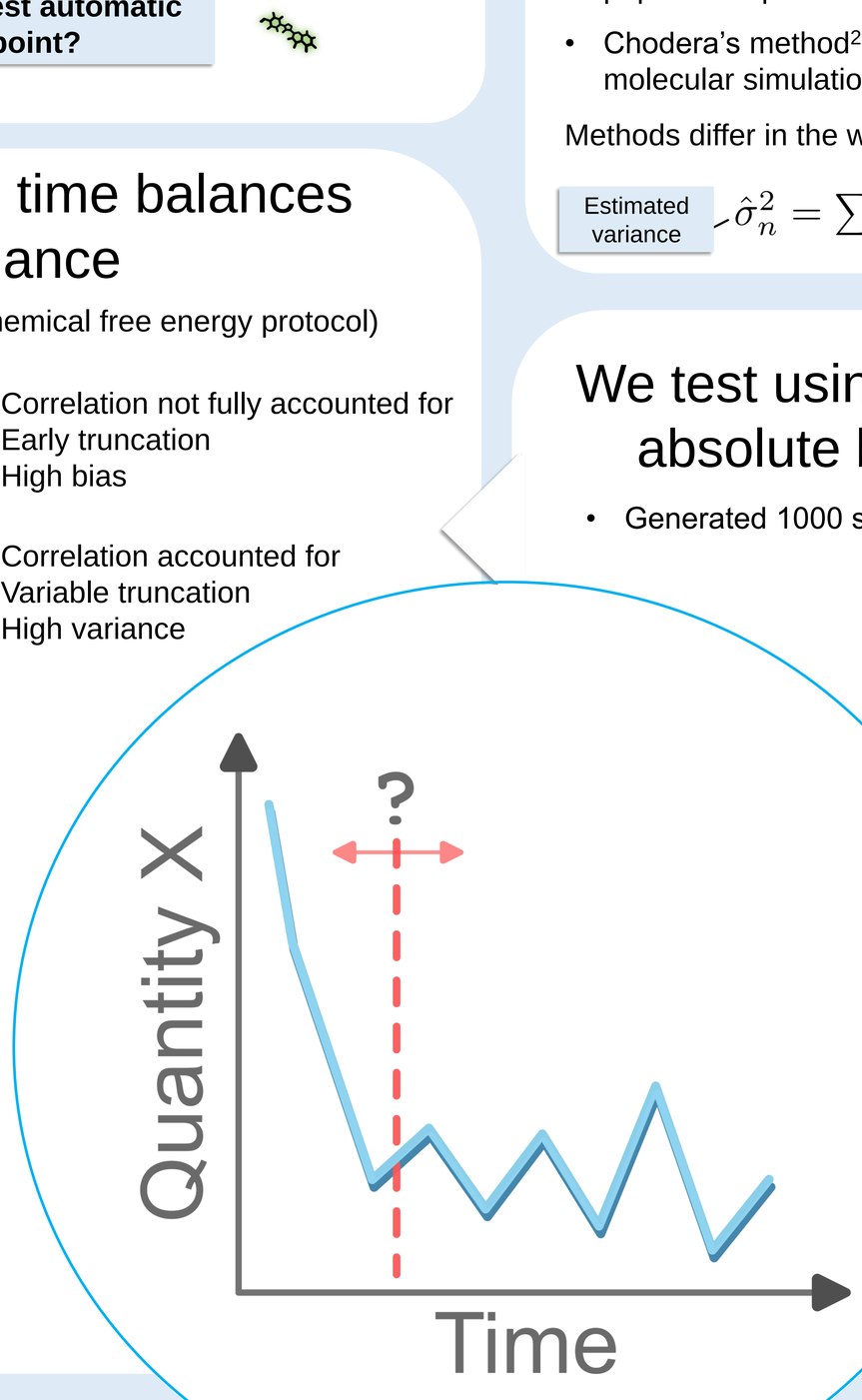
The optimal truncation time balances bias and variance

- Tested on initial 8 ns of data (typical alchemical free energy protocol)



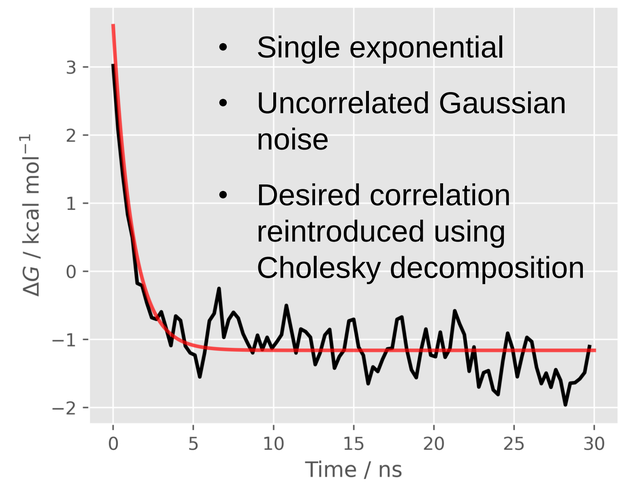
- Correlation not fully accounted for
- Early truncation
- High bias

- Correlation accounted for
- Variable truncation
- High variance



We test using synthetic data modelled on long absolute binding free energy calculations

- Generated 1000 synthetic "runs" with different noise modelled on 5 real runs



- Single exponential
- Uncorrelated Gaussian noise
- Desired correlation reintroduced using Cholesky decomposition

We test different methods for calculating the variance

- **Uncorrelated variance estimate** (similar to MSER)
- **Windowed estimates of variance** using Bartlett window with window size 5 ($w_n(t) = 0 \text{ if } t > 5$, similar to MSER-5) and size $\sqrt{n_{samp.}}$:

$$\hat{\sigma}_n^2 = \sum_{t=-\infty}^{\infty} w_n(t)\hat{\gamma}_{n,t}$$

- **Initial sequence estimates of variance** using Chodera's method and Geyer's Initial Convex Sequence[3] method

The window estimator with size √n appears robust

- Window variance estimator with size √n produces minimum RMSE by balancing bias and variance
- This method is also substantially faster than the initial sequence methods (~40 times faster for this example)

Similar results are obtained when the data are modified

Variation	Summary
Discard 99 of every 100 data points	Window size √N remains lowest RMSE automated method
Only use the first 0.2 ns of data	
Increase variance by a factor of 5	
Block average data with block size 100	Uncorrelated estimate gives lowest RMSE

Conclusions

- **Failure to account for autocorrelation produces large bias** (uncorrelated estimate and widow size 5)
- **Initial sequence methods** which more thoroughly account for autocorrelation **reduce bias but increase variance**
- **Window method with size √n** strikes a **good balance between bias and variance** for this data
- Further testing with data modelled on other molecular dynamics simulations required

All methods implemented in python package: **https://github.com/fjclark/red**

References

[1] K. P. White, *Simulation*, 1997, **69**, 323–334.
[2] J. D. Chodera, *J. Chem. Theory Comput.*, 2016, **12**, 1799–1805.
[3] C. J. Geyer, *Stat. Sci.*, 1992, **7**, 473–483.