Markus Wallerberger*, **Aleksander Gaenko**, **Emanuel Gull**

UNIVERSITY OF MICHIGAN

Hypothesis testing of quantum Monte Carlo simulations

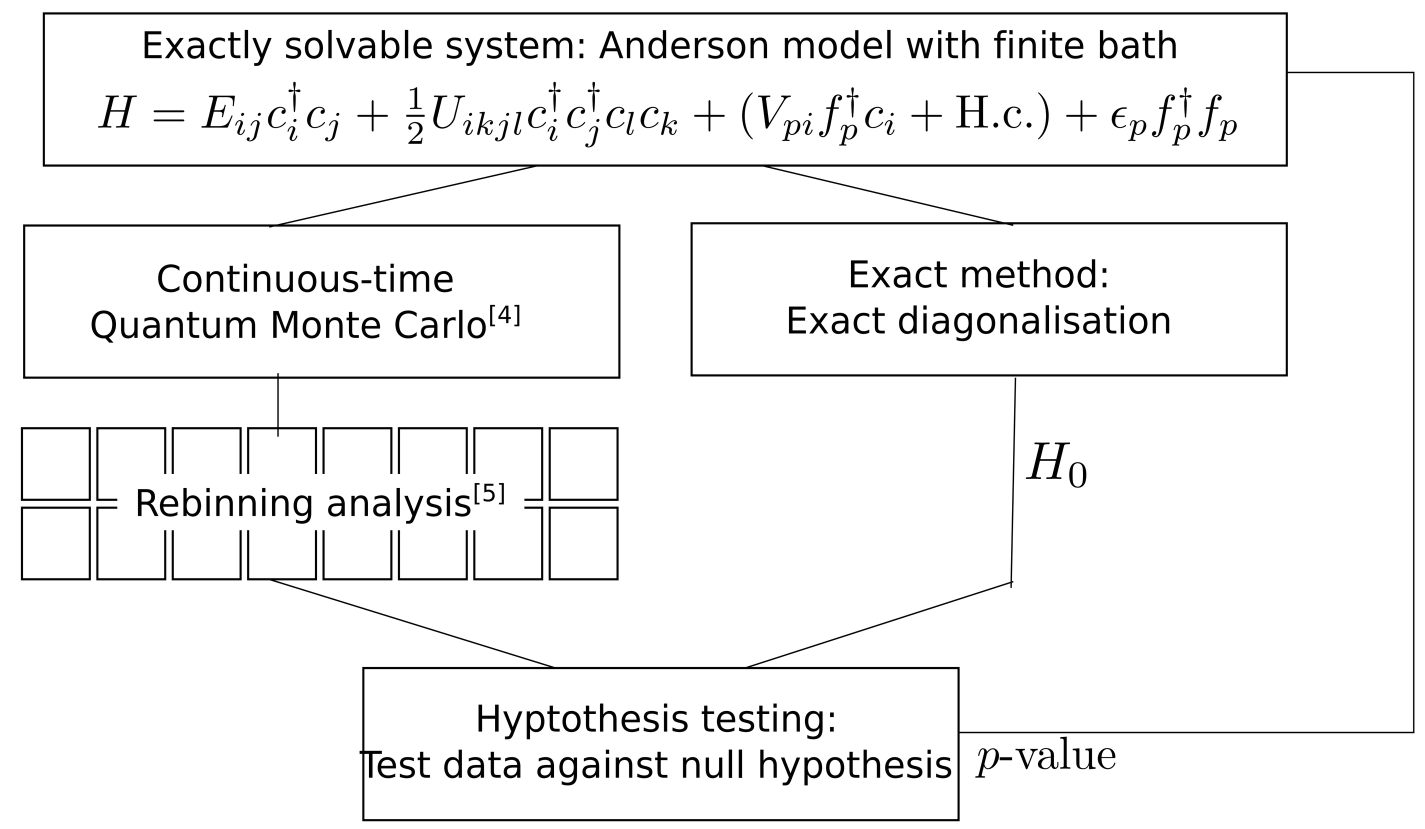
Introduction

- Many-body algorithms become more and more involved.
  E.g., Typical implementation size of Anderson model solvers:
  - Exact diagonalisation: ~ 1,000 LOC
  - Hirsch-Fye quantum Monte Carlo: ~ 10,000 LOC
  - Continuous-time quantum Monte Carlo:[4] ~ 100,000 LOC
- Testing and benchmarking becomes important!
- **Benchmarks**: check agaianst "known" physical results.
  → Important, but "high-level"
- **Contract-oriented programming**: testing invariants (asserts)
  → "low-level" step verification but no test of, e.g. Markov chain
- **Unit tests**: small, fast, self-contained, user-facing tests
  pointwise correctness tests for the interface
  verified at every development step
- **Fuzzing**: behavior tests for large number of random inputs
  reliability and stability tests for the code
  trivially ("embarassingly") parallelizable
  verified once for releases or continuously
- unit tests and fuzzing are **deterministic**
  stochastic algorithm results are not verifyable
  fixed seed tests break at valid changes to algorithm

References

[1] K. Subr and J. Arvo: *Proc. 15th Pacific Conf. on Comput. Graph. Appl.*, p. 106 (2007)
[2] H. Ševčíková et al.: *Proc. 2006 Int. Symp. on Softw. Testing and Analysis*, p. 215 (2006)
[3] A. Gaenko et al.: arXiv 1609.03930; accepted in *Comput. Phys. Commun.* (2017)
[4] E. Gull et al.: *Rev. Mod. Phys.* 83, 349 (2011)
[5] e.g., T. Hesterberg et al.: Bootstrap methods and permutation tests, in: Moore and McCabe, eds., *Introduction to the practice of statistics* (2005)
[6] M. Marozzi: *Stat. Meth. Med. Res.* 25(6) 2593–2610 (2016)

Hypothesis testing[1,2]

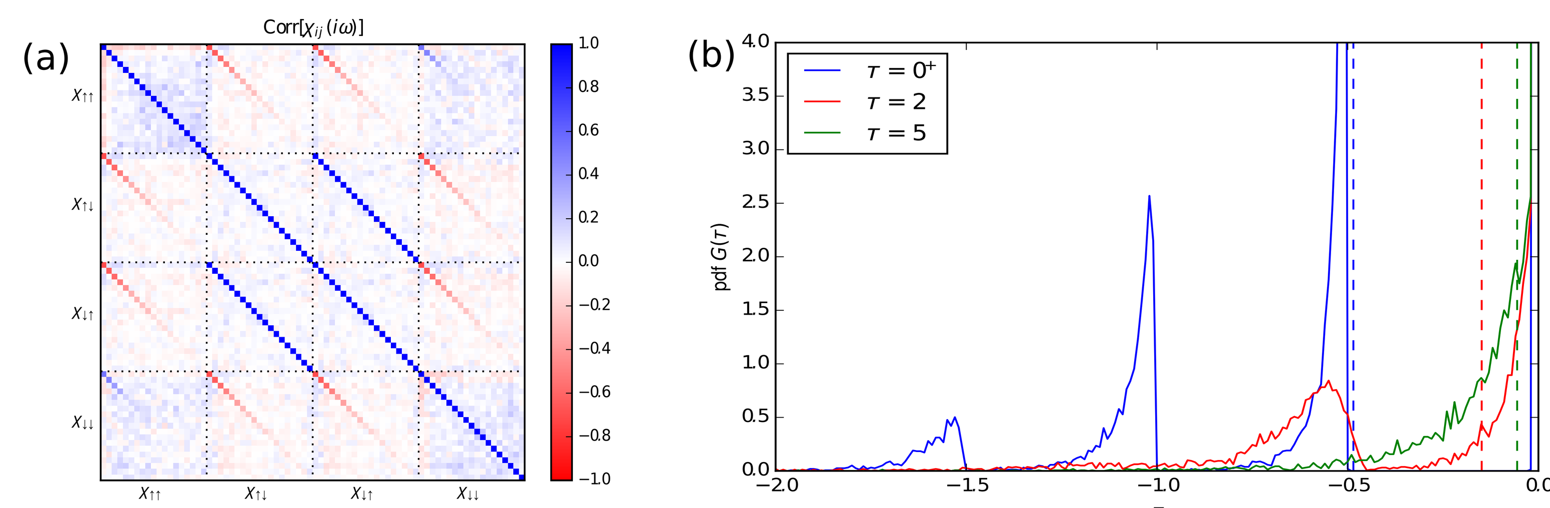


- **Student's *t*-test:** one (few) data points (densities etc.)
- **Hotellings's T^2-test:** series of data points (Green's function etc.)

$$T^2 = N(\langle x\rangle - x_0)^\dagger \Sigma_x^{-1} (\langle x\rangle - x_0) \qquad T^2 \sim \frac{n(N-1)}{N-n} F_{n,N-n}$$

  - $\sigma > \sigma_0; P(F \geq t^2) < p$: systematic error or error bars too small
  - $\sigma < \sigma_0; P(F \leq t^2) < p$: error bars too **large**

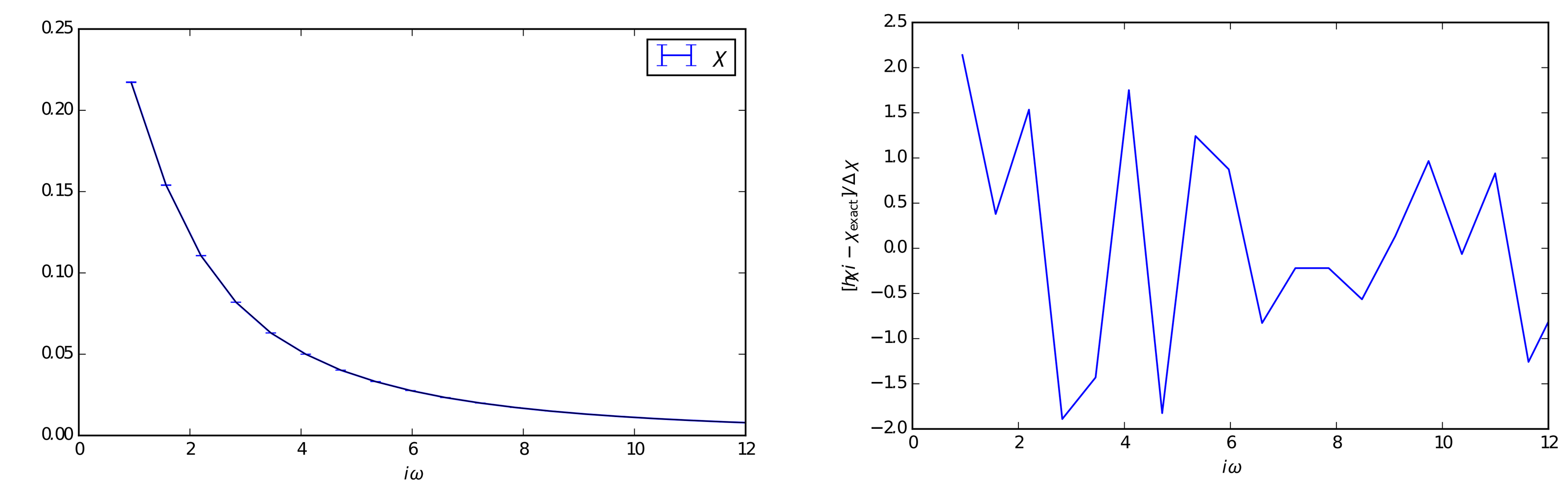
- **Complications:** (a) correlation/clones; (b) non-normality; and (c) cases when more datapoints than bins[6]

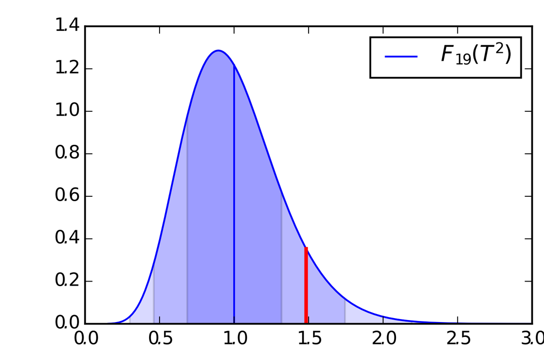


- **Stochastic fuzzing:** *p*-value-guided sampling of parameter space
  propose $H_\mathrm{new} = H + \delta H$, if $p[H_\mathrm{new}] < p[H]$ then $H \leftarrow H_\mathrm{new}$
  –> improve on sampling of discontinuous indicator function
- **Outlook:** part of testing framework of AlpsCore[3]

Example I: $\chi_{ij}(i\omega)$

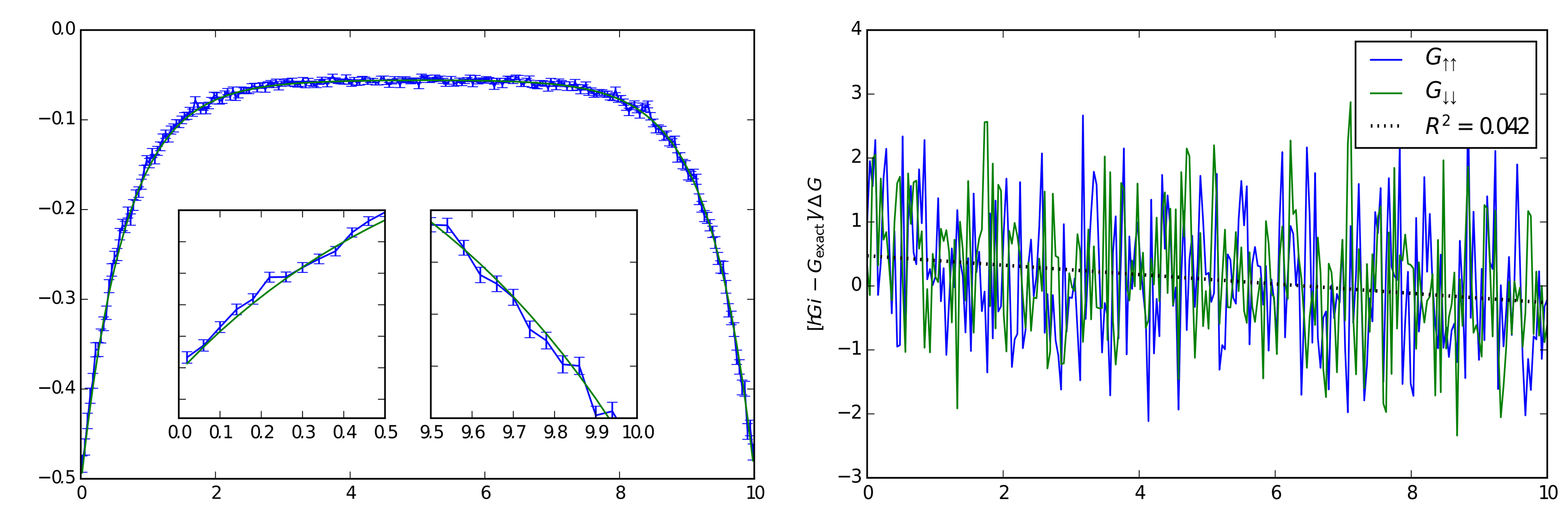
- Single-orbital AIM, two bath states (+0.5, -0.5), V=1, U=1, μ=0.42

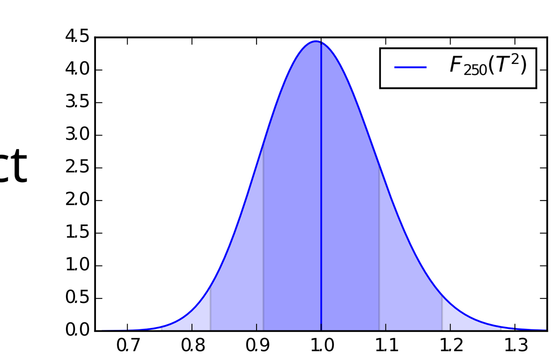




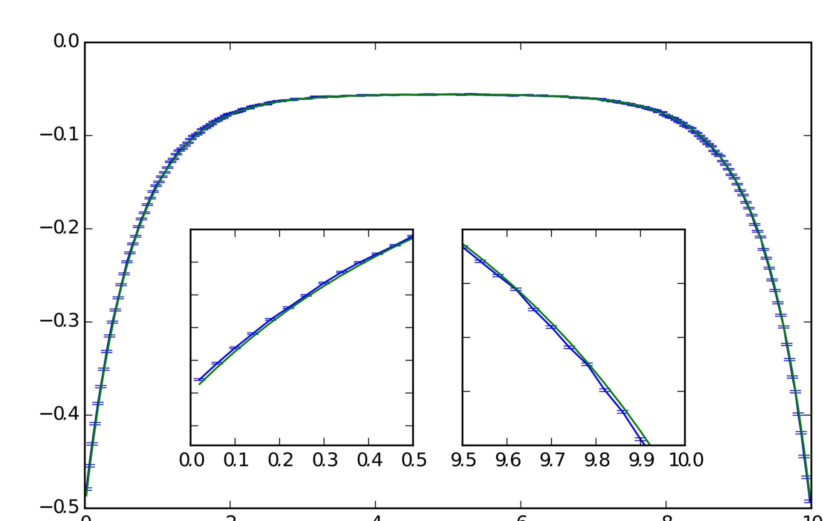
- **visual inspection**: looks O.K.
  maybe small bias to lower side
- **hypothesis testing**: one-sided F test
  $T^2 \approx 1.48$ ($p \approx 0.08$, OK at the 2σ level)

Example II: $G_{ij}(\tau)$





- **visual inspection**: looks O.K.
  slight bias visible in data, but no clear reject
- **hypothesis testing**: one-sided F test
  $T^2 = 2.3$ (rejection at the 15σ level)



- reason: hybridisation function discretization + linear interpol.
- solution: exponential models

$$G_\mathrm{model}(\tau) = A_+ \mathrm{e}^{-\tau B_+} + A_- \mathrm{e}^{(\beta-\tau)B_-}$$