QMC and thinning for empirical datasets

Mike Giles, Fei Xie

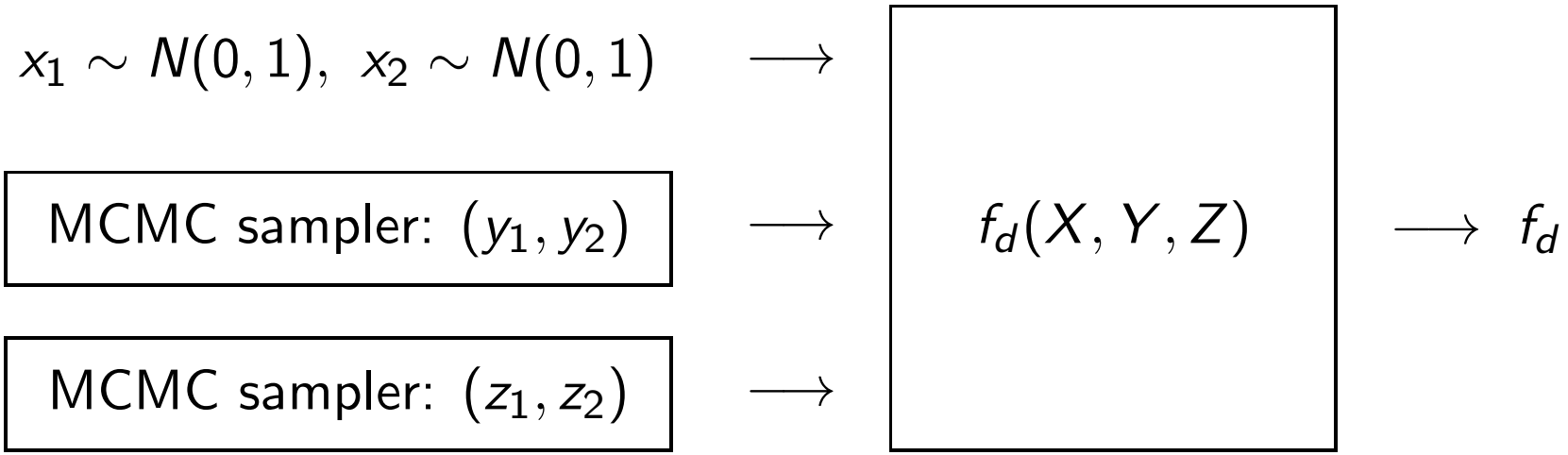
Mathematical Institute, University of Oxford

Tractability of High Dimensional Problems and Discrepancy

Sept 26, 2017

Motivation

In a nested expectation application (EVPPI), some of the random inputs come from a Bayesian posterior distribution, with samples generated by MCMC methods.



If the MCMC posteriors are approximated by multi-variate Gaussians, then QMC can be used, and is often very effective.

But how can we avoid that approximation error and apply QMC directly to the empirical dataset?

Motivation

An alternative idea is to first pre-generate and store a large collection of MCMC samples:

$$\{Y_1, Y_2, Y_3, \ldots, Y_N\}$$

When applying Monte Carlo methods (or MLMC) random samples for Y can be obtained from this collection by uniformly distributed selection.

This avoids the problem of strong correlation between successive MCMC samples (and also works fine for the MLMC calculation).

This approach also leads to ideas on QMC for empirical datasets, and dataset thinning, reducing the number of samples which need to held in memory during calculations.

Two questions

QMC methods generate quasi-uniformly distributed points in $[0,1]^d$, then map them to points in some target domain X.

Requires a mapping $f : [0,1]^d \to X$ giving the desired distribution on X.

Question 1: what do we do if we want a uniform distribution on a large empirical dataset $X = \{X_0, X_1, \ldots, X_{N-1}\}$?

Our aim is to end up with the following mappings when $N = 2^{dk}$ for some integer k:

$$[0,1]^d \longrightarrow \{0,1,\ldots 2^k-1\}^d \longrightarrow \{0,1,\ldots 2^{dk}-1\} \longrightarrow X$$

- first step is simply $U \to \lfloor 2^k U \rfloor$ applied to each dimension
- other two steps are 1–1 giving uniform distribution on X

Two questions

Question 2: if N is very large, can we thin it down to a subset of size $M \ll N$ without introducing much error?

e.g. for 3D dataset, maybe generate $2^{3\times 8} \approx 16$M samples, then thin it down to perhaps $2^{3\times 6} \approx 260$k samples

The benefit of this is that 260k samples can be held in the L3 cache of a CPU, so it will improve the execution efficiency of the code.

There are a number of other approaches to thinning datasets, e.g.

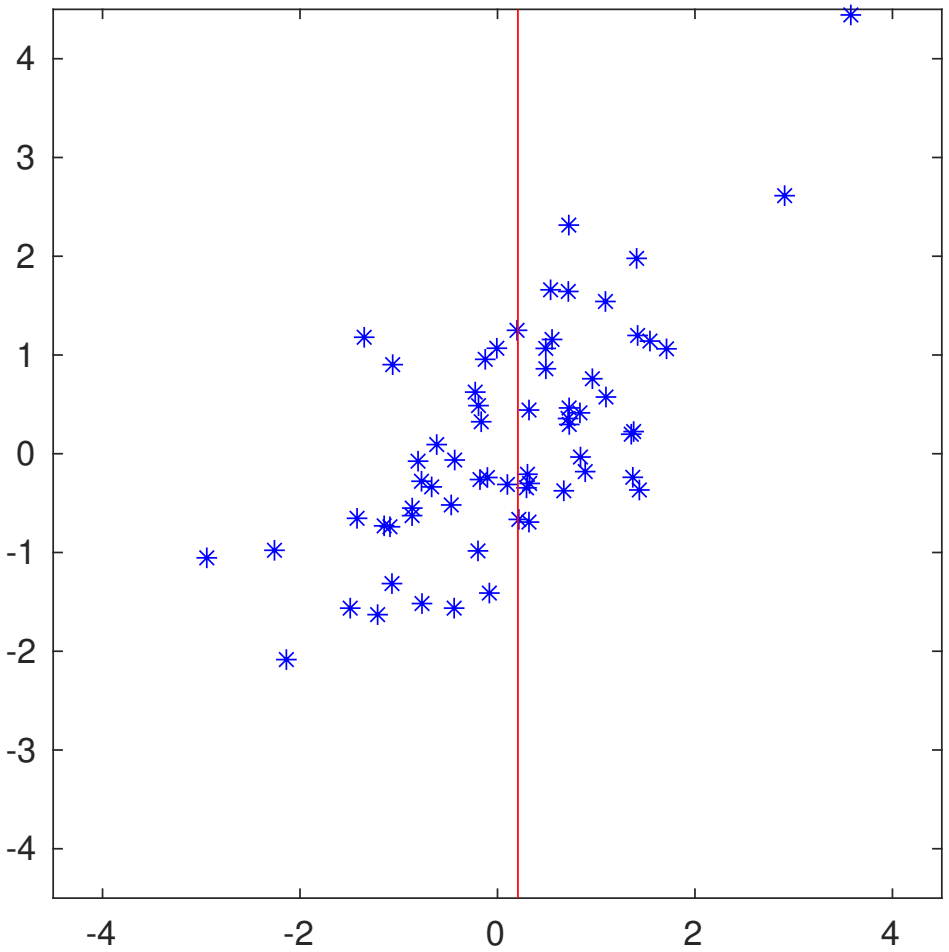
- Principal points (Flury, 1990)
- Support points (Mak & Joseph, 2017)

The relative merits of these different approaches have not yet been investigated – Question 1 is probably the more important one, and the more interesting one perhaps for this workshop.

Recursive bisection

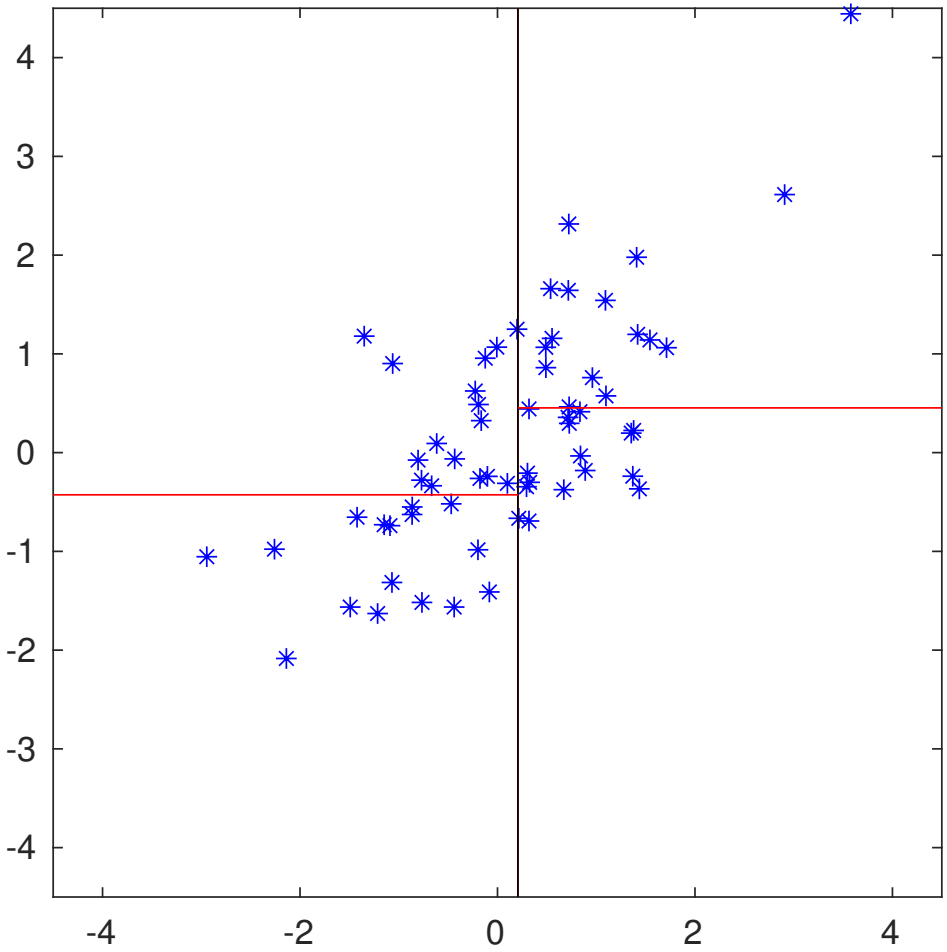
A simple idea: apply recursive bisection in alternating directions.

Step 1: sort points based on first coordinate to bisect into two halves



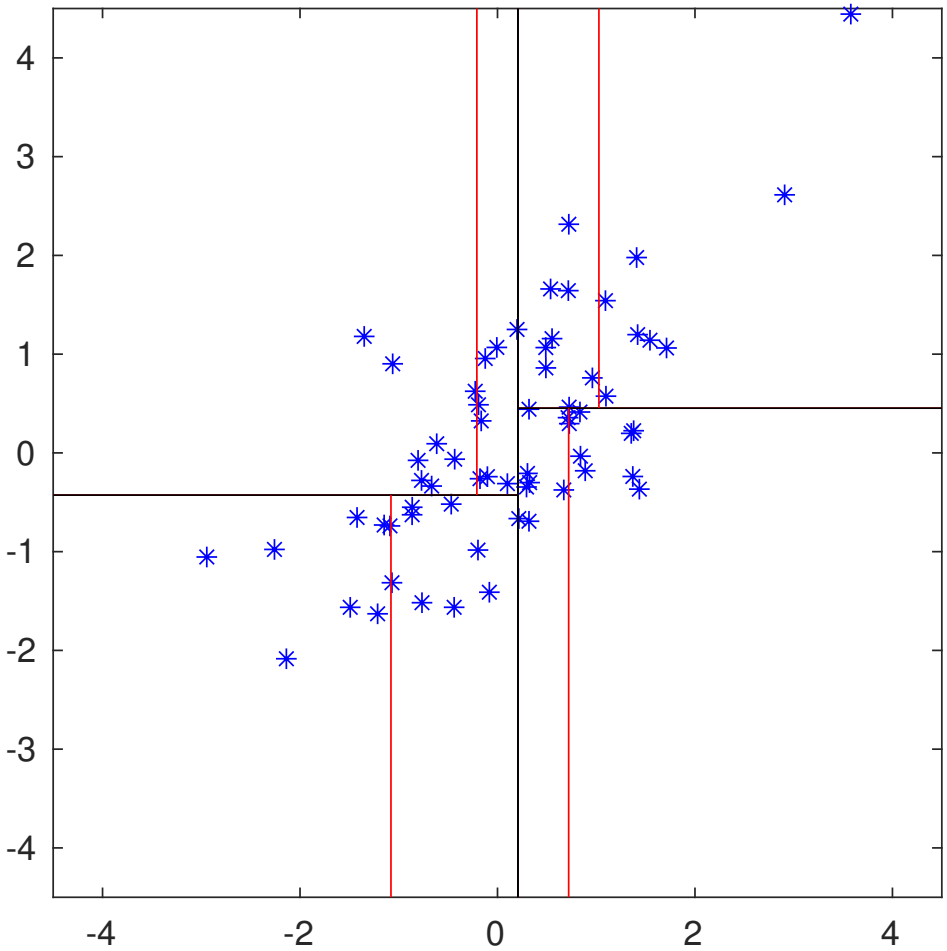
Recursive bisection

Step 2: within each half, sort points based on second coordinate to again bisect



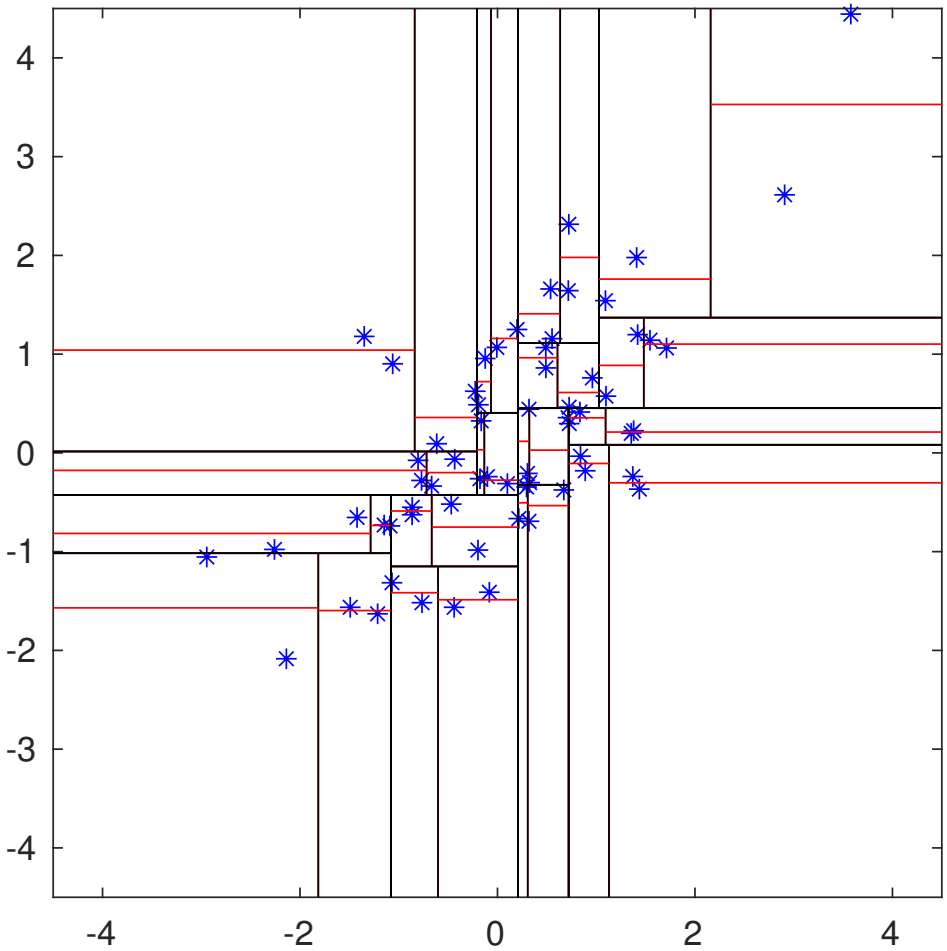
Recursive bisection

Step 3: in 2D, within each quarter, sort points based on first coordinate



Recursive bisection

Repeat until there is just one point in each subset



Recursive bisection

In this 2D example, we end up with $2^k \times 2^k$ subsets (with one point each) with the single entry for subset (i_1, i_2) stored at index I formed by interleaving the bits:

$$\begin{aligned} i_1 &= 1\ 0\ 1 \\ i_2 &= \ \ 0\ 1\ 0 \\ \Longrightarrow \quad I &= 1\,0\,0\,1\,1\,0 \end{aligned}$$

This gives us our mapping: $\{0, 1, \ldots 2^k - 1\}^2 \longrightarrow \{0, 1, \ldots 2^{2d} - 1\}$ for the re-ordered data, but it is convenient to do a further re-ordering so that we can access the same data through the simpler mapping $(i_1, i_2) \longrightarrow i_1 + 2^k i_2$

This all generalises naturally to higher dimensions.

QMC tests

We hope that this new approach applied to a dataset generated from a known distribution is comparable to applying standard QMC to the same distribution.

To test this, we consider the following 4 distributions:

- 2D independent Gaussians: $X_1 \sim N(0,1), X_2 \sim N(0,1)$,
- 2D multivariate Gaussian: $(X_1, X_2) \sim N\left(0, \begin{pmatrix} 1 & 1/\sqrt{2} \\ 1/\sqrt{2} & 1 \end{pmatrix}\right)$
- 2D distorted Gaussian: $X_1 = Z_1,\ X_2 = Z_2 + Z_1^2$ where $Z_1 \sim N(0,1), Z_2 \sim N(0,1)$,
- double well distribution: sum of two Gaussian distributions

and in each case use the test function $f(x_1, x_2) = \sin(x_1 + x_2)$.

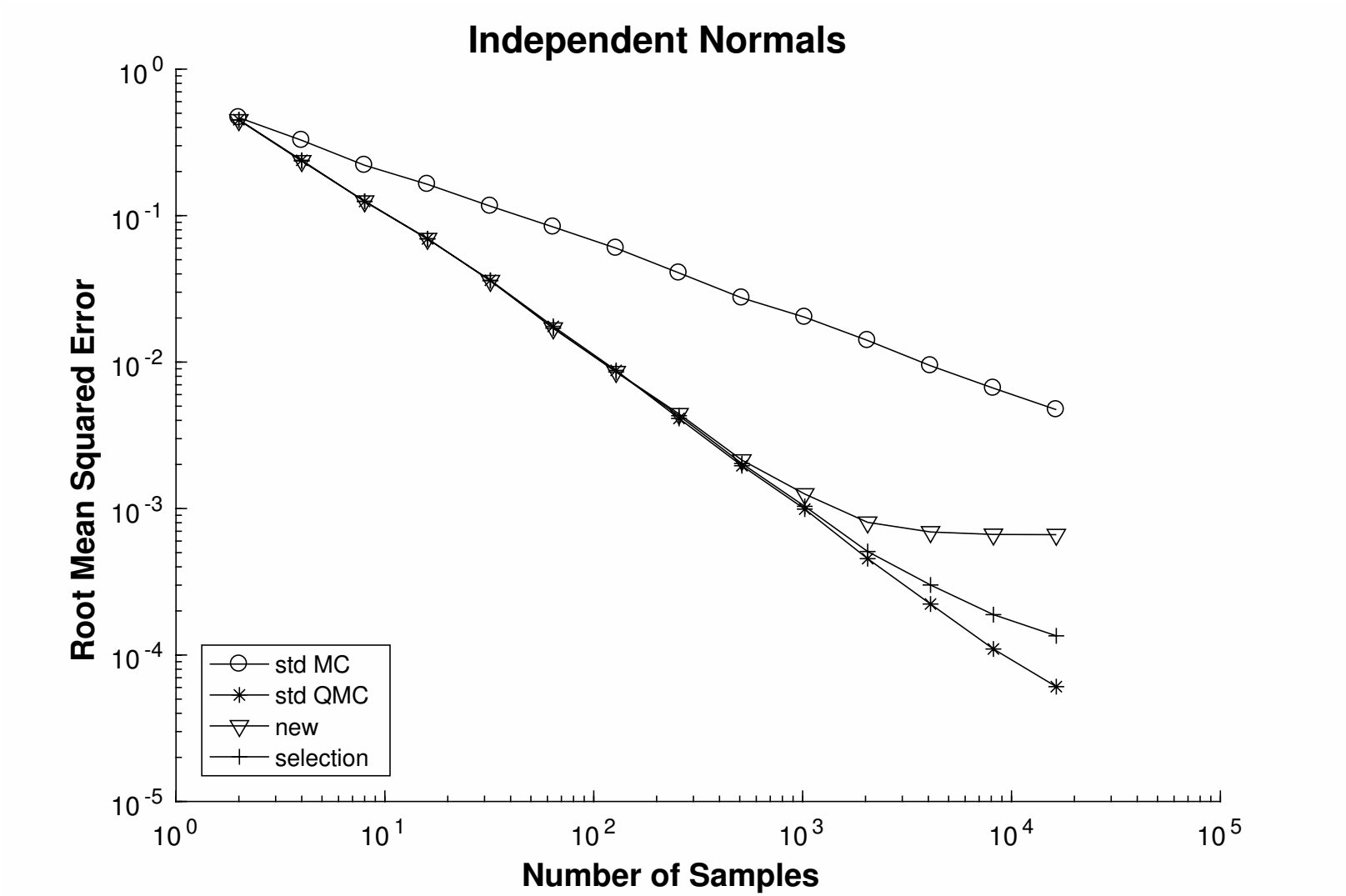
QMC tests

In each case, we

- generate multiple random datasets of 2^{20} points
- for each, generate randomised QMC estimates using varying numbers of Sobol points with digital scrambling
- compare overall RMS error from 256 randomisations (of datasets and scramblings) to that given by standard QMC and MC applied to same distributions

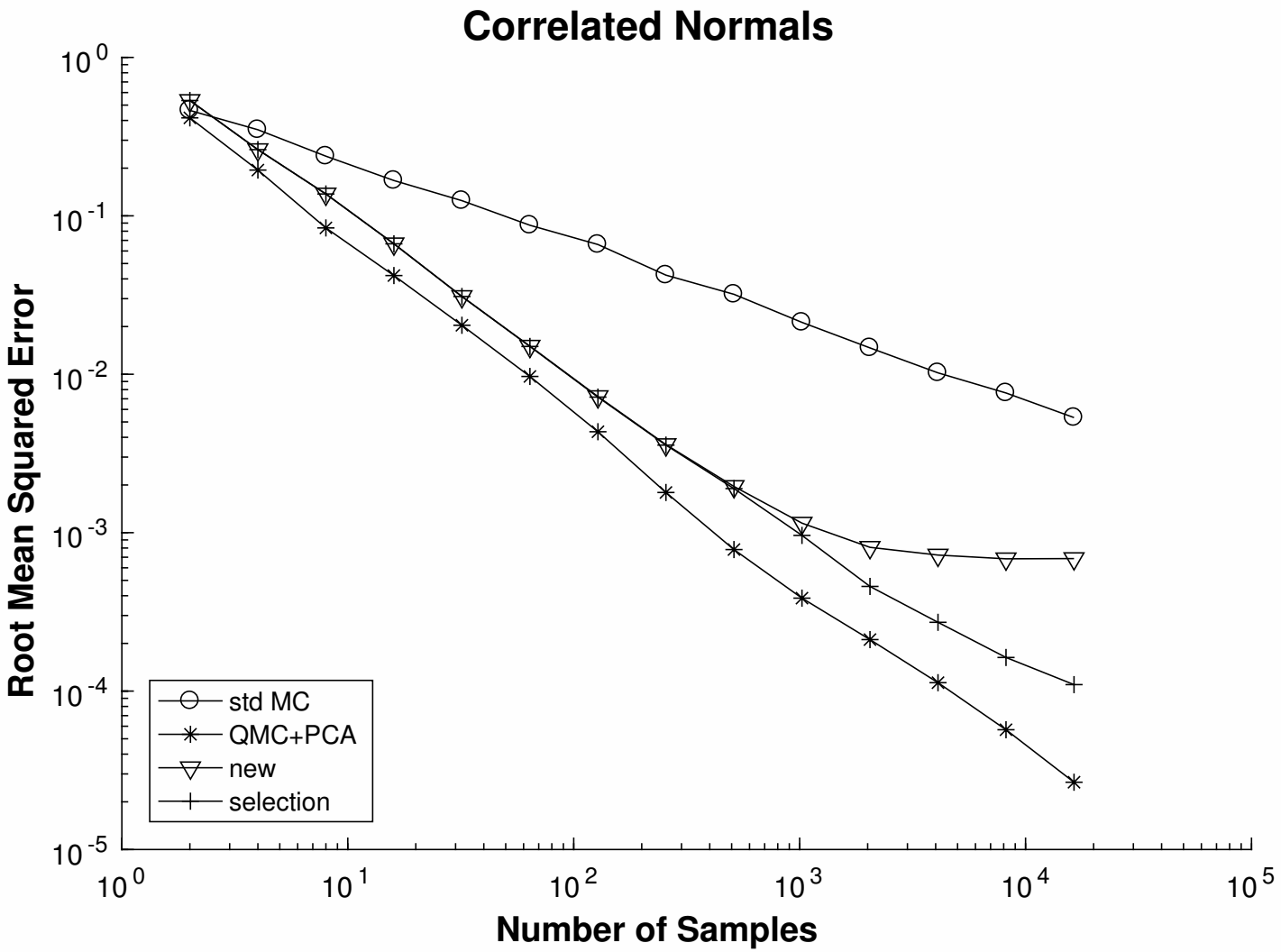
QMC tests

Results for independent Gaussians



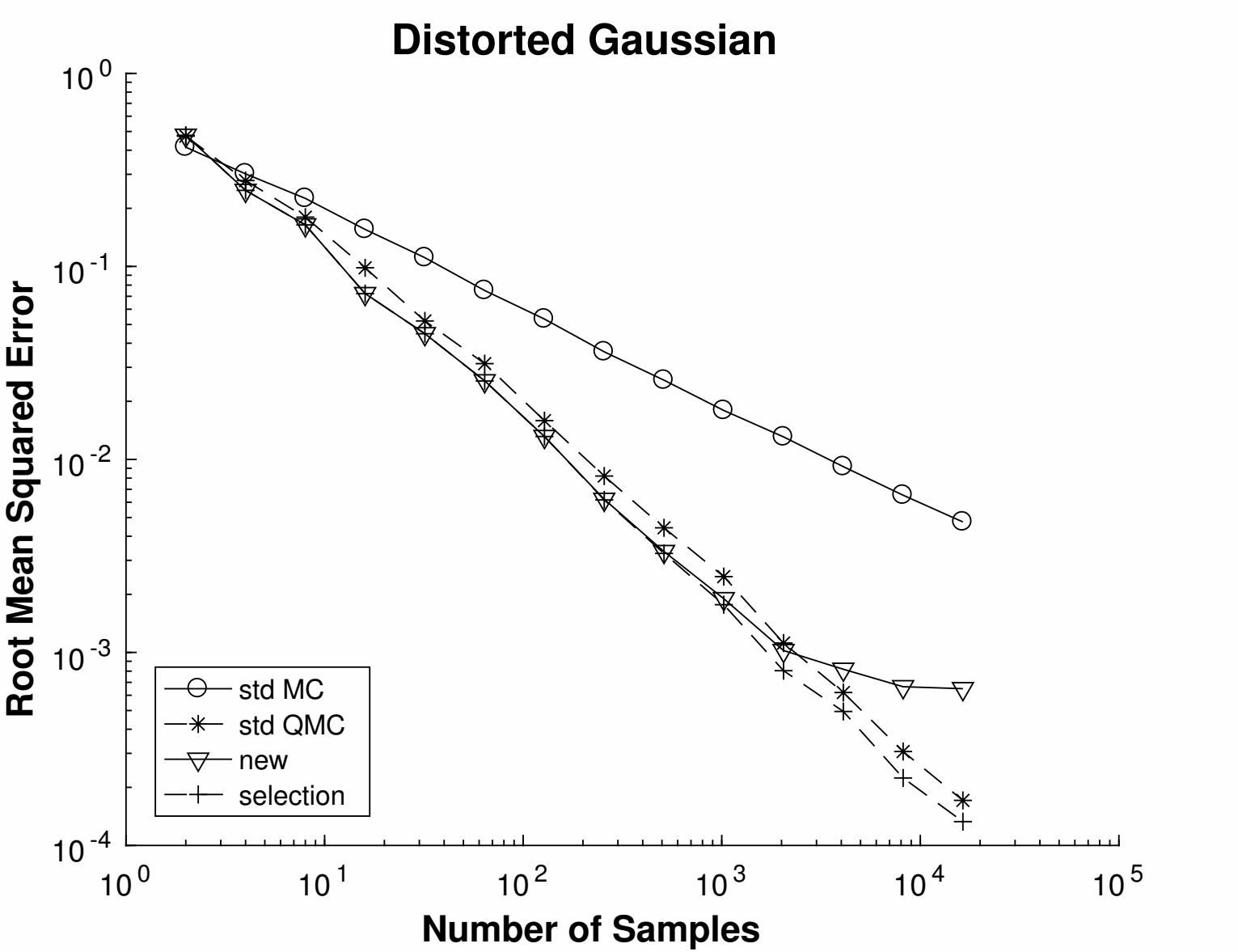
QMC tests

Results for multivariate Gaussian



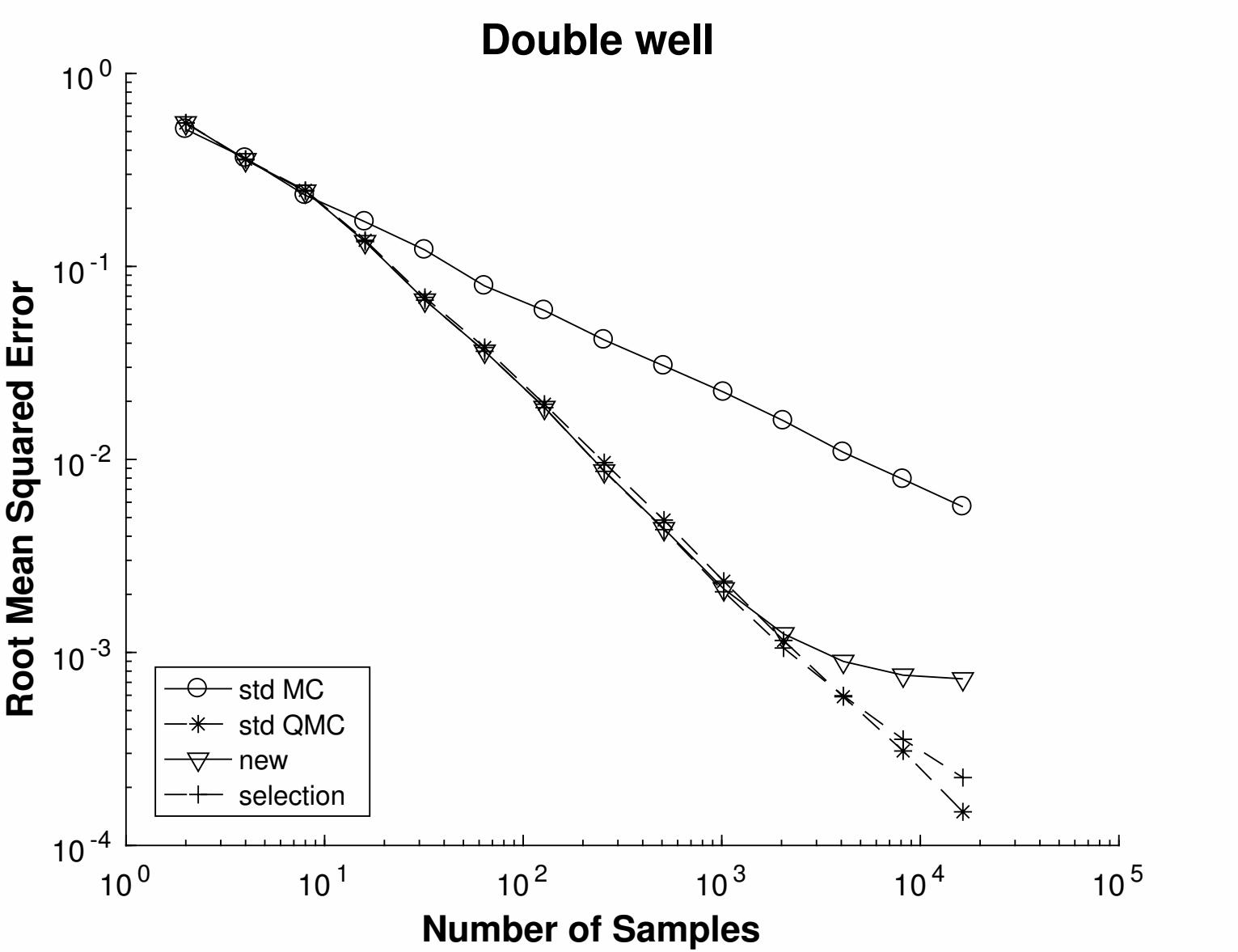
QMC tests

Results for distorted Gaussian



QMC tests

Results for double well:



Thinning

The thinning idea builds on the same recursive bisection:

- start with $N = 2^{dk_1}$ samples
- recursively bisect into $M = 2^{dk_2}$ subsets, with $k_2 < k_1$
- from each subset S_k, randomly select one point $Y^{(j)}$

Note that the expectation over all random selections is unbiased, i.e.

$$\mathbb{E}\left[\frac{1}{M}\sum_j f(Y^{(j)})\right] = \frac{1}{M}\sum_j\left(\frac{M}{N}\sum_{X^{(j)}\in S_k} f(X^{(j)})\right)$$
$$= \frac{1}{N}\sum_j f(X^{(j)})$$

Surprisingly (?), this is better in general (particularly in higher dimensions) than using the mean of each subset, which gives a biased estimator.

Thinning

We are able to do some error analysis if we make the following assumption:

Assumption

The recursive bisection results in subsets $S_k, k = 1, \ldots, M$ *in which*

$$\max_k \max_{X^{(i)}, X^{(j)} \in S_k} \|X^{(i)} - X^{(j)}\| = O(M^{-1/d}).$$

This seems reasonable, at least for distributions on bounded domains with strictly positive densities, since there are $O(M^{1/d})$ blocks in each direction.

Thinning

Theorem

If the Assumption is satisfied, and the function $f(x)$ is Lipschitz, with $|f(x)-f(y)| \le \|x-y\|$, then the RMS thinning error is $O(M^{-1/2-1/d})$.

Proof: the error in subset S_k,

$$e_k \equiv f(Y^{(k)}) - \frac{M}{N} \sum_{X^{(j)} \in S_k} f(X^{(j)})$$

is zero mean, and because of the Assumption and the Lipschitz property we have $|e_k| \le c\ M^{-1/d}$.

Hence, due to independence, the variance of the overall error is

$$\mathbb{V}\left[M^{-1}\sum_k e_k\right] = M^{-2}\sum_k \mathbb{V}[e_k] \le M^{-1} c^2 M^{-2/d}.$$

Thinning

A planar cut through a regular d-dimensional grid of blocks with $N^{1/d}$ in each direction cuts at most $d\,N^{(d-1)/d}$ of the blocks. This inspires the following theorem which considers functions with discontinuities.

Theorem

If the Assumption is satisfied, and the function $f(x)$ is such that

- $|f(x)-f(y)| \le 1$ *for pairs* (x,y) *in* $O(N^{(d-1)/d})$ *of the subsets*
- $|f(x)-f(y)| \le \|x-y\|$ *for pairs* (x,y) *in the other subsets*

then the RMS thinning error is $O(M^{-1/2-1/2d})$.

Proof: very similar to the previous one, with the dominant error contribution from the first category of subsets.

Thinning

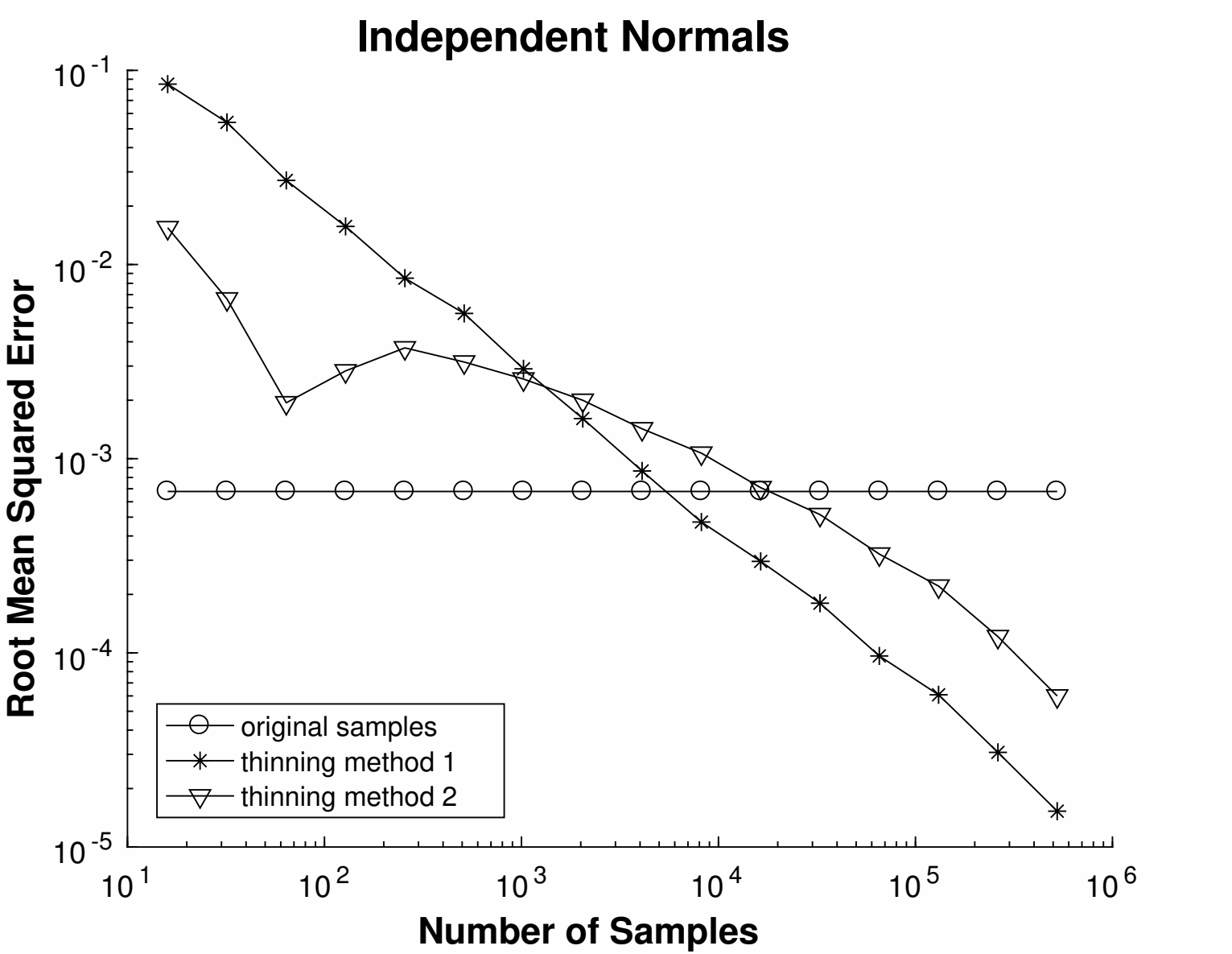
In both cases, the thinning error should be compared to the $O(N^{-1/2})$ sampling error arising from the original sample set being a collection of iid samples from the true distribution.

Equating the two errors gives

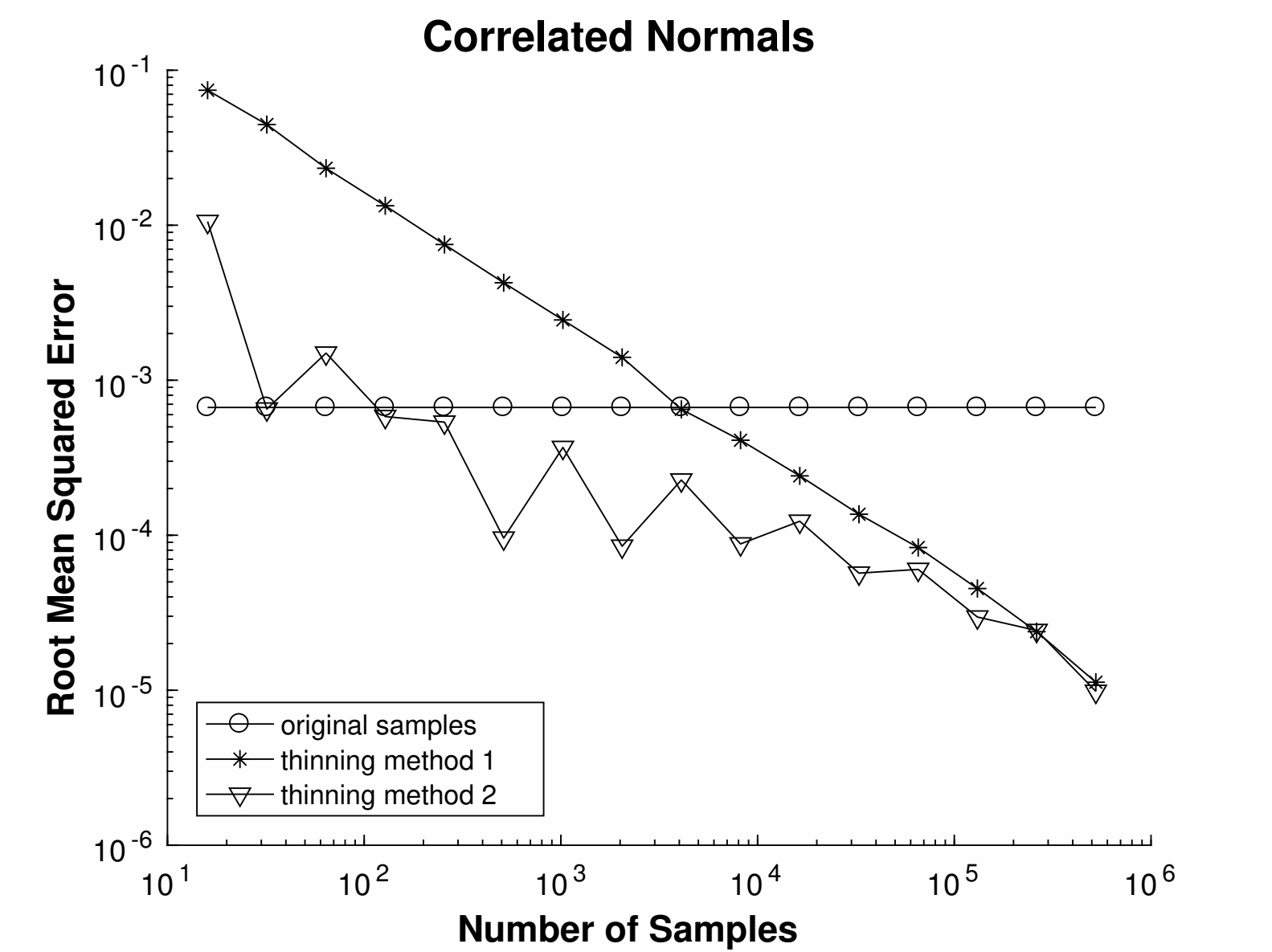
$$M = \begin{cases} O(N^{d/(d+2)}), & \text{Lipschitz case} \\ O(N^{d/(d+1)}), & \text{discontinuous case} \end{cases}$$

This indicates the potential for significant savings in the size of M when the dimension is low, but not so much in higher dimensions.

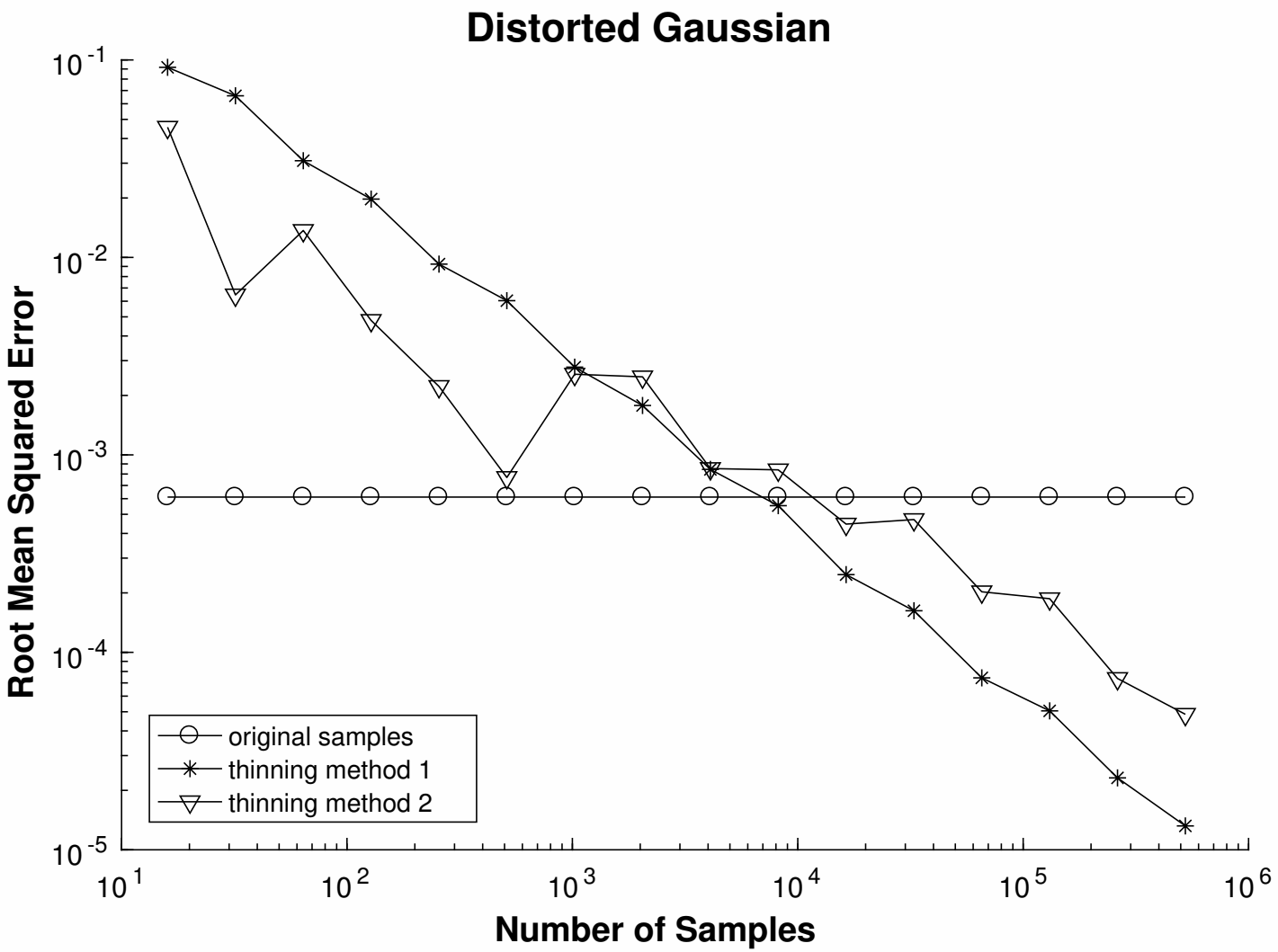
Thinning tests



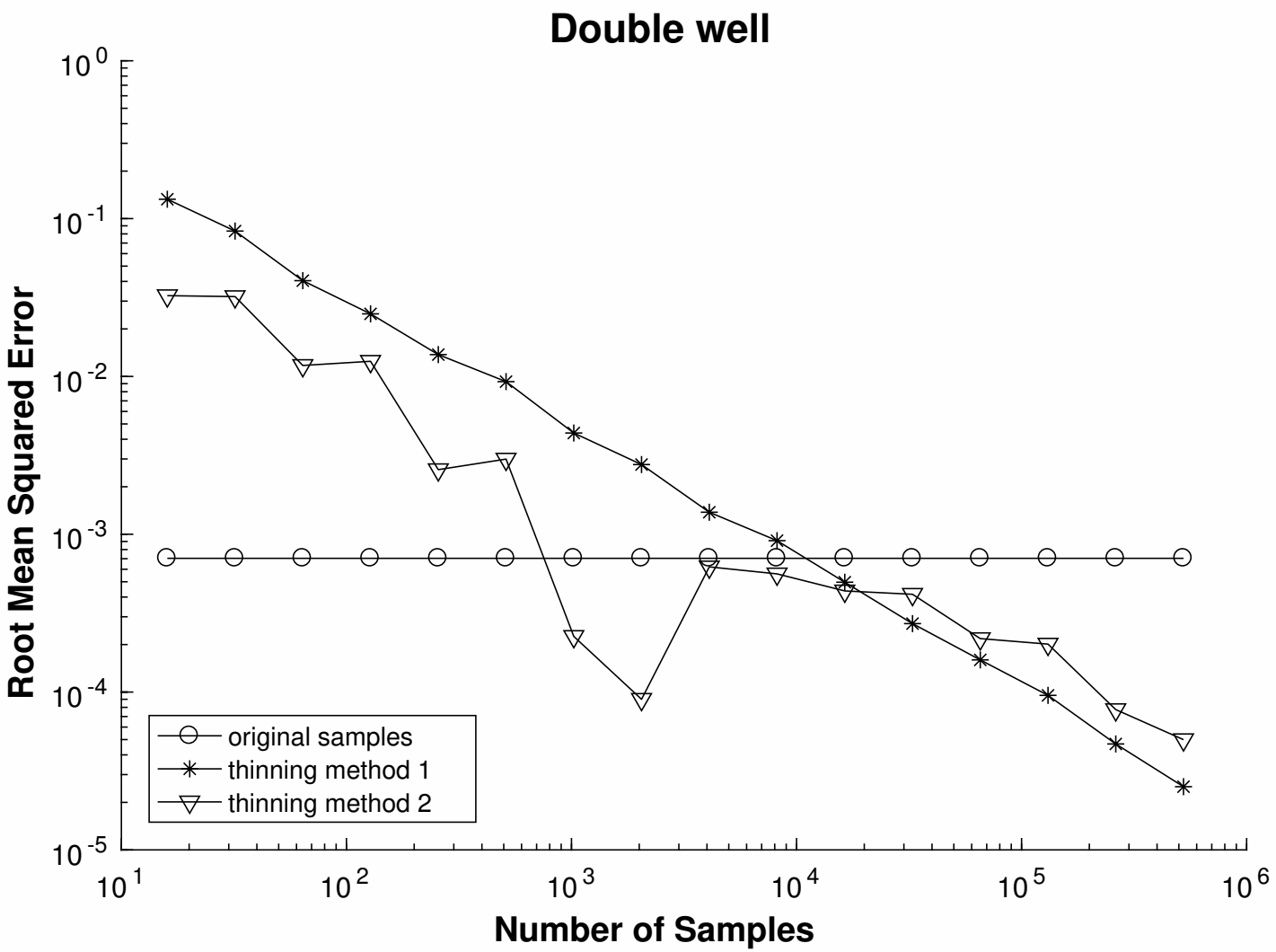
Thinning tests



Thinning tests



Thinning tests



Conclusions

- it seems the QMC approach is effective, at least in low dimensions, so maybe a good way to apply QMC to samples generated by MCMC – would be good to have some supporting theory
- the thinning also seems effective, particularly in low dimensions, with its effectiveness depending on the smoothness of the integrand, but it needs to be compared against other thinning technqiues

Webpage: `http://people.maths.ox.ac.uk/gilesm/slides.html`

Email: `mike.giles@maths.ox.ac.uk`