Improved data covariance estimation techniques in lattice QCD

J.N. Simone
Fermilab

20 June 2017

Lattice2017
35TH INTERNATIONAL SYMPOSIUM ON LATTICE FIELD THEORY

covariant fits

We minimize the covariant sum of squares

$$\chi^2\left(\vec{a}\right) = \sum_{j,k=1}^{p} \left(\frac{\bar{y}_j - f_j(\vec{x},\vec{a})}{\sigma_{\bar{y}_j}}\right) \mathrm{C}^{-1}_{j,k} \left(\frac{\bar{y}_k - f_k(\vec{x},\vec{a})}{\sigma_{\bar{y}_k}}\right)$$

over p points to estimate the fit parameters $\vec{a}$. The $\bar{y}$ are averaged over n samples, expected to be independent (no auto-correlations).

y-correlation matrix, C, is often estimated to be the sample correlation matrix.

As $p/n \to 1$, $\mathrm{C}_{\mathrm{sample}}$ has small singular values which can destabilize the fit.

There are one or more zero singular values when $p/n \geq 1$.

typical mitigations

-  **KEEP CALM and FIT ON!**
- Generate more samples increasing *n*.
- Thin the data, reducing the number of points *p*.
- Employ SVD to eliminate / modify modes having small singular values.
- How and where to set the SVD floor / cut?
  - A fixed floor: $\epsilon\langle\texttt{float}\rangle$, $\sqrt{\epsilon\langle\texttt{float}\rangle}$, or something else.
  - Keep principal modes accounting for some given percentage of total variance.
  - Look for some feature (a knee, a break) in the singular value mode spectrum.
  - Look for and modify modes (sv's) suspected of being "noisy".

experiment I: “vine” synthetic data

Start from a known population correlation matrix, Σ.

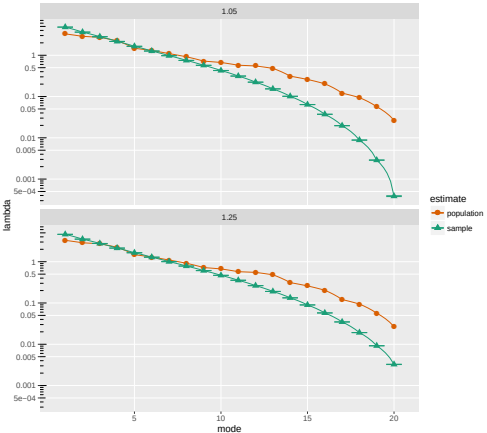
We construct Σ via the ‘vine” algorithm[1].

From Σ generate many (4000) trial data sets for each given sample size n.

Study averaged properties of the sample correlation matrix as a function of n.

[1] DOI:10.1016/j.jmva.2009.04.008

“vine” synthetic data: spectrum



Correlation mat. sing. value spectrum shown for $n/p = 21/20$, and $25/20$.

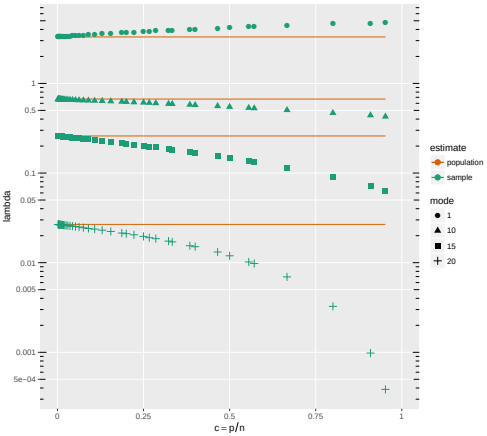
Averaged over 4000 random trials.

Sample errors not visible.

Sampling under estimates smaller s.v. and over estimates the larger s.v.

Systematic effect.

“vine” synthetic data: sing. values vs $c = p/n$



Sing. values of selected modes plotted *vs* $c = p/n$.

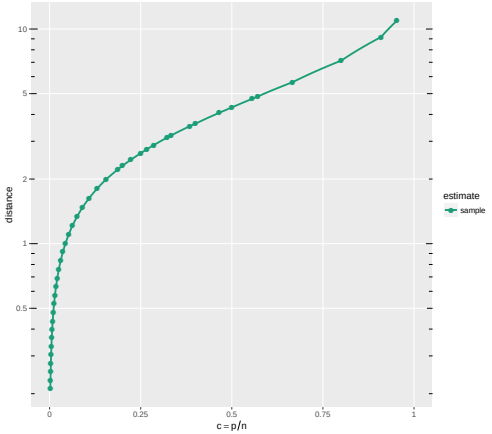
Limiting population values as redish lines.

Sample values (errors invisible) as greenish points.

Over/under shoot of large/small sample values.

Sample values converge slowly to population values.

“vine” synthetic data: distance



Shows average $\mathrm{dist}_{\mathrm{R}}(\Sigma, \mathrm{C}_{\mathrm{sample}})$ *vs* $c = p/n$.

Riemann distance[1] between two equal rank covariance matrices

$$\mathrm{dist}_{\mathrm{R}}(S_1, S_2)^2 \equiv \sum_{j=1}^{p} \left(\log \lambda_j\right)^2 ,$$

where

$$S_1 \phi_j = \lambda_j S_2 \phi_j ,$$

the generalized eigenvalue problem.

[1] DOI:10.1111/j.1558-5646.2008.00587.x

experiment II: HISQ pion two-point data

HISQ physical mass pion two-point data at $a = 0.15$ fm.

Have 3630 analyzed configurations available.

Two pion correlators each with a local sink, one with a smeared source.

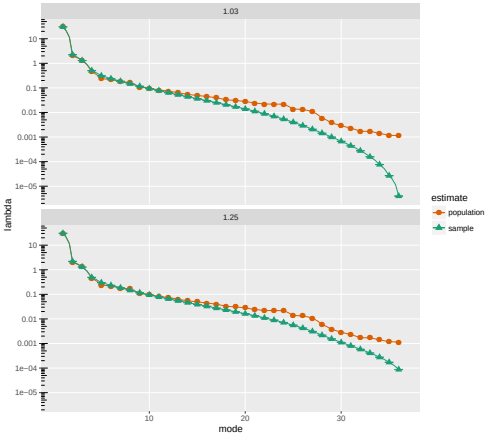
Test $p = 36$ data set using both correlators and timslices $\in [3, 20]$.

Approximate population Σ from full data set.

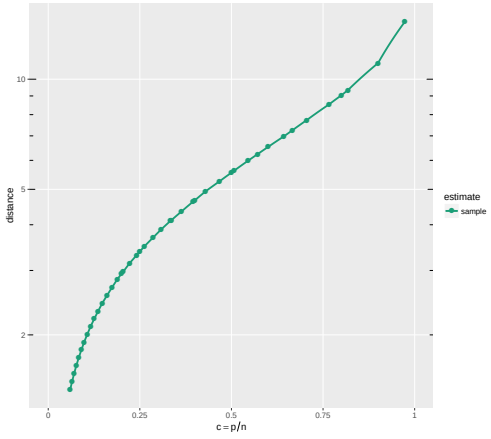
Sample n configurations (with no replacement) for $36 < n \leq 600$.

HISQ pion two-point: spectrum and distance

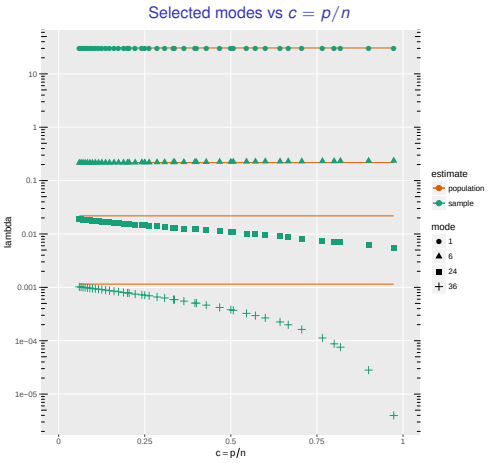
Spectrum $n/p = 37/36$, and $45/36$



$\mathrm{dist}(\mathrm{C}_{\mathrm{full}}, \mathrm{C}_n)$ vs p/n



HISQ pion two-point: finite sampling effects



Find a similar pattern of $C_{\mathrm{sample},n}$ singular value spectrum deviations from the inferred population spectrum.

Want to do further such tests on even larger lattice data sets.

What is known from the literature about such effects?

Random Matrix Theory

Distributions of singular values in the limit $c = p/n$ fixed, and $n \to \infty$.

Marčenko and Pastur[1]: population e.d.f., H_n, has a limiting distribution $H_n(\tau) \to H(\tau)$.

Sample e.d.f., $F_n(\lambda)$, has Stieljes transform

$$\forall z \in \mathbb{C}^+, \quad m_{F_n}(z) = \frac{1}{p} \sum_{j=1}^{p} \frac{1}{\lambda_j - z} \, .$$

Silverstein[2] writes Marčenko-Pastur relation between F and H,

$$\forall z \in \mathbb{C}^+, \quad m_F(z) = \int_{-\infty}^{\infty} dH(\tau) \, \frac{1}{\tau \left[1 - c - c\, z\, m_F(z)\right] - z}$$

[1] DOI:10.1070%2FSM1967v001n04ABEH001994

[2] J.Multivar.Anal. **55**(2)(1995)

Numerical solution of Marčenko-Pastur relation

Ledoit & Wolf[1] introduce the QuEST (Quantized Eigenvalues Sampling Transform) function which discretizes the relation between the F and H distributions.

They present their numerical implementation of the QuEST function in a second publication[2].

Map population eigenvalues to an estimate of eigenvalues for sample size n,

$$\{\tau_j : j = 1 \ldots p\} \xrightarrow{\text{QuEST}} \{\lambda_j^{(n)} : j = 1 \ldots p\}$$

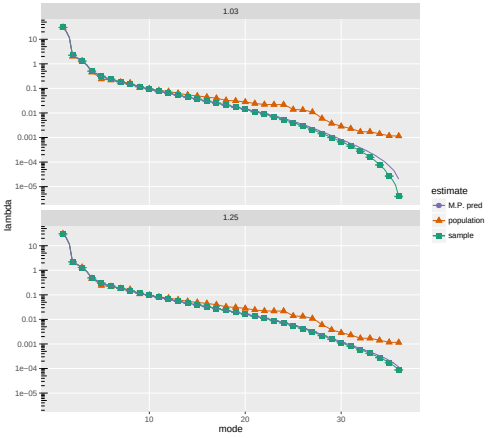
Test predicted sample spectrum for our two eamples...

[1] arXiv:1406.6085

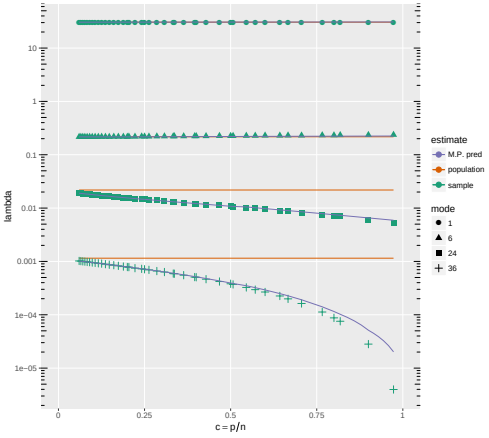
[2] arXiv:1601.05870

HISQ pion two-point: Marčenko-Pastur prediction

Spectrum $n/p = 37/36$, and $45/36$

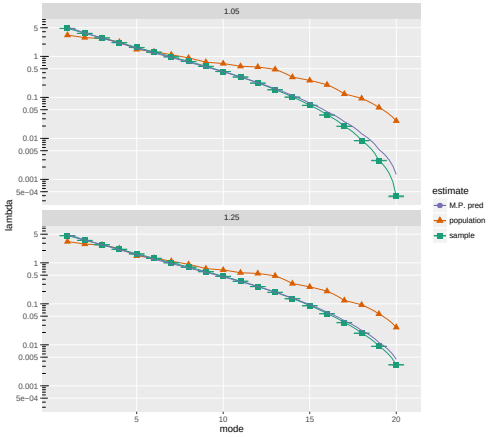


Selected modes vs $c = p/n$

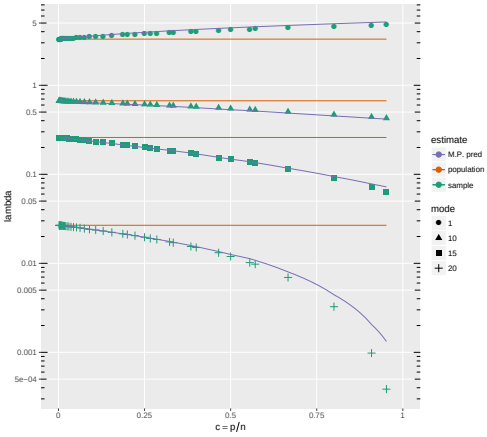


“vine” synthetic data: Marčenko-Pastur prediction

Spectrum $n/p = 21/20$, and $25/20$



Selected modes vs $c = p/n$



non-linear shrinkage

The goal is to improve the estimation of correlation.

Ledoit and Wolf consider estimators

$$\mathsf{C}_{\mathrm{shrink}} = \mathsf{U}\ \mathrm{diag}(\{d_j\})\ \mathsf{U}^T$$

where the above is the sample estimator when $d = \lambda$, the sample eigenvalues.

The optimal choice is found by minimizing the Frobenius norm

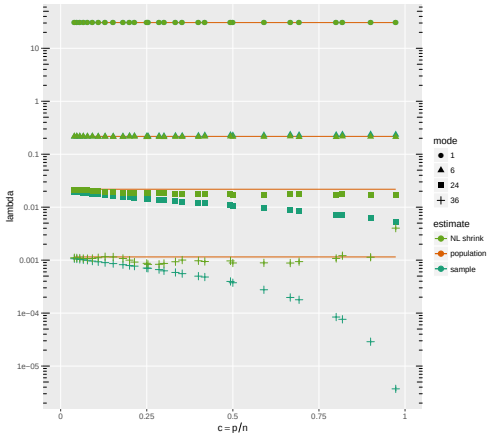
$$\min_{\{d_j\}} \left\| \mathsf{U}\ \mathrm{diag}(\{d_j\})\ \mathsf{U}^T - \Sigma_n \right\|_F$$

where the $\{d_j\}$ are computed using the QuEST function.

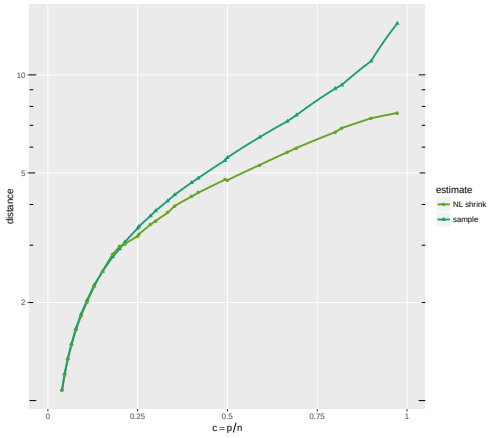
Try for our two examples...

HISQ pion two-point: non-linear shrink

Selected modes vs $c = p/n$

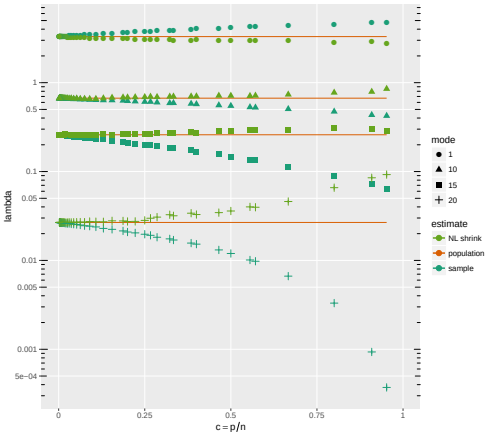


sample and shrink distances vs p/n

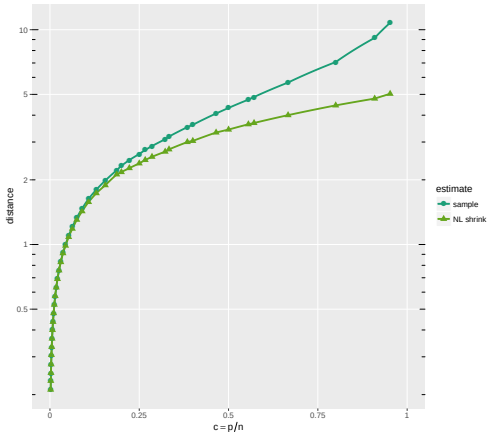


“vine” synthetic data: non-linear shrink

Selected modes vs $c = p/n$



sample and shrink distances vs p/n



summary and outlook

- C_{sample} singular values look problematic, especially when $p/n > 0.33$.
- In these tests, non-linear shrinkage estimate for C converged faster, on average, to the population correlation matrix.
- More tests with actual and simulated data; examine effect on fits.
- These results produced using  package `nlshrink`.
- I'm working on a (second) QuEST function implementation in 

.