

Bias reduction through regularization

Miron Ivanov

August 3, 2017

Here we conducted a simple simulation experiment that shows that oracle regularized estimator for Min-Var portfolio is more robust(in terms of bias) to decrease in concentration ratio than sample estimator and nonlinear oracle estimator with k-fold cross-validation and has similar to Ledoit-Wolf(LW) model accuracy in case of sufficient number of simulations. For details of the k-fold and LW estimators see [alex write up here].

Non-negative Least Squares (NNLSQ) with Regularization estimator

Following the notations in [alex write up here] for oracle optimal shrinkage estimator of Min-Var portfolio, we would like to find optimal shrunk eigenvalues $d_i = 1/z_i$ by solving the system:

$$CAz = \alpha$$

for z where $C = U^\top \Sigma U$, $A = \text{Diag}(\alpha_1, \dots, \alpha_N)$ and $\alpha = U^\top \mathbb{1}$.

Unfortunately, solution to this system is not guaranteed to be positive and in fact appears to be very noisy. Reasons for the former are not quite intuitive and are related to the fact that we are targeting the Min-Var portfolio which has the property of error maximization during variance minimization procedure. To keep eigenvalues positive and decrease the amount of noise, we restrict the values of z to be non-negative and add Tikhonov L_2 Regularization:

$$z = \arg \min \|CAz - \alpha\|_2^2 + \lambda^2 \|z\|_2^2 : z \geq 0$$

or

$$z = \arg \min \|Gz - b\|_2^2 : z \geq 0$$

where

$$G = \begin{bmatrix} CA \\ \lambda I \end{bmatrix}, b = \begin{bmatrix} \alpha \\ 0 \end{bmatrix}$$

To avoid cases when $z = 0$ we used linear interpolation between neighborhood values.

Simulation setup

In order to test the performance of the introduced estimator we ran simulations of $\Sigma_N = UDU^\top$ with the following parameters:

$$N = 100; T = \{200, 500, 1000\}; \lambda = 0.000005; \#simulations = \{1, 50, 100, 200\}$$

We considered a simple case in which Σ_N is a diagonal matrix with uniformly distributed variances mapped to an interval corresponding to about 5-30% volatility per year and assumed return matrix X to be iid with zero mean.

We employed sample, asymptotic LW, nonlinear oracle with k-fold cross-validation and NNLSQ with regularization estimators to extract optimal shrinkage values from Σ_N . We were particularly interested in the reduction of bias while we were decreasing the concentration ratio and varying the number of simulations. More specifically, we define bias as the difference between population eigenvalues and optimal shrinkage eigenvalues :

$$bias = \sum_{i=1}^N |\tau_i - d_i^*|$$

Note that direct comparison with LW estimator cannot be strictly justified since the former one uses a different loss function and is optimized for Max-Sharpe portfolio. It should apriori exhibit higher accuracy as compared to other estimators because of the property of Min-Var optimization routine to maximize errors and return noisier estimations of eigenvalues. Thus we conduct this comparison for information purposes - to see that optimal eigenvalues of LW and other estimators are actually different.

Simulation results

First we test how estimators behave when we change the number of data samples. As we can see, three of them return somewhat better results: the bias decreases proportionally (linearly) to increase in T . However, NNLSQ estimator shows almost no improvement, which is because of its inherently high volatility.

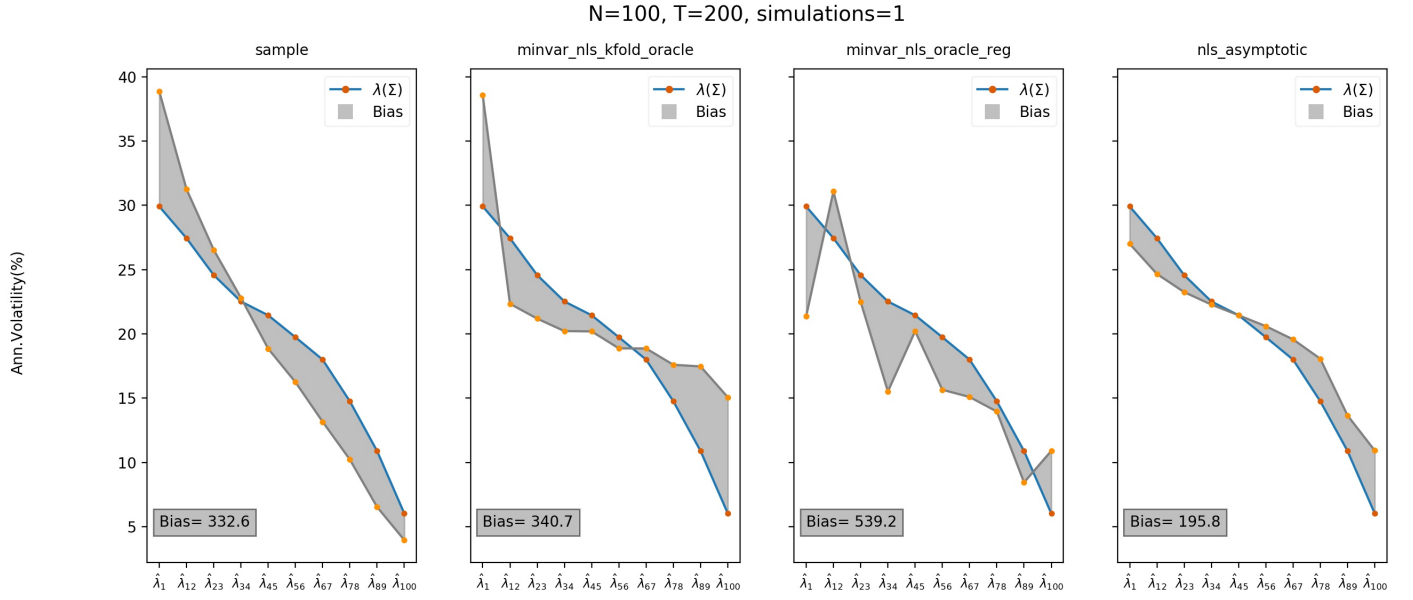


Figure 1a

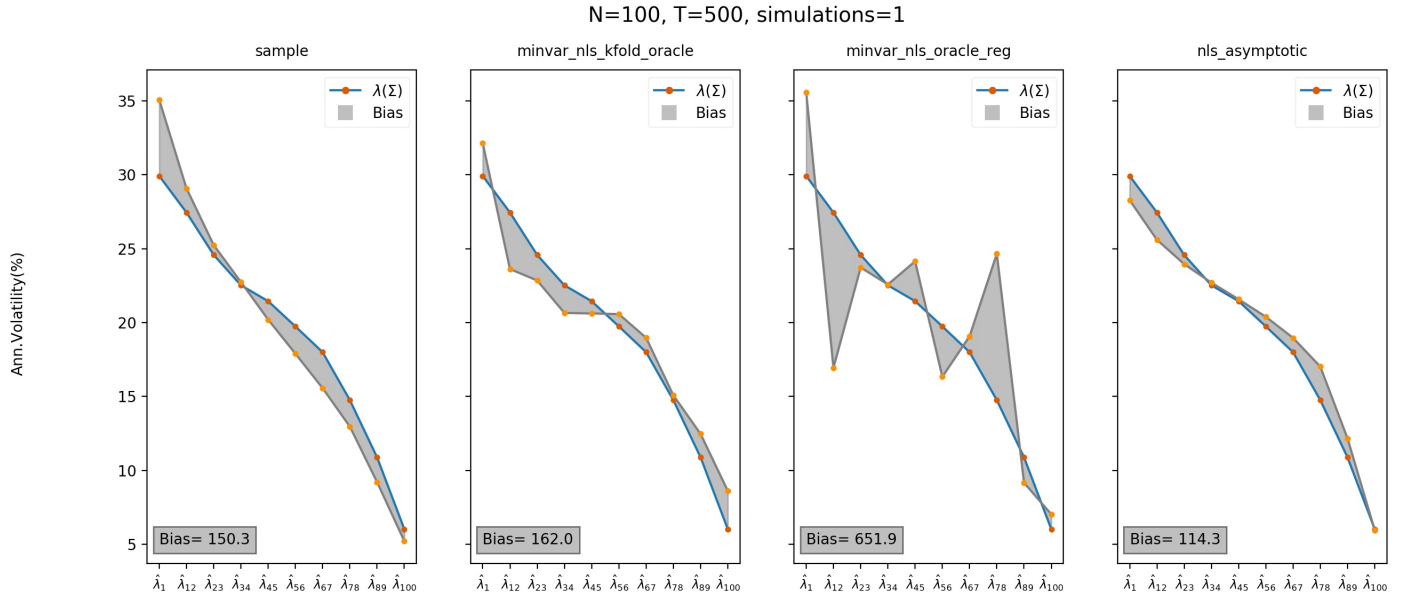


Figure 1b

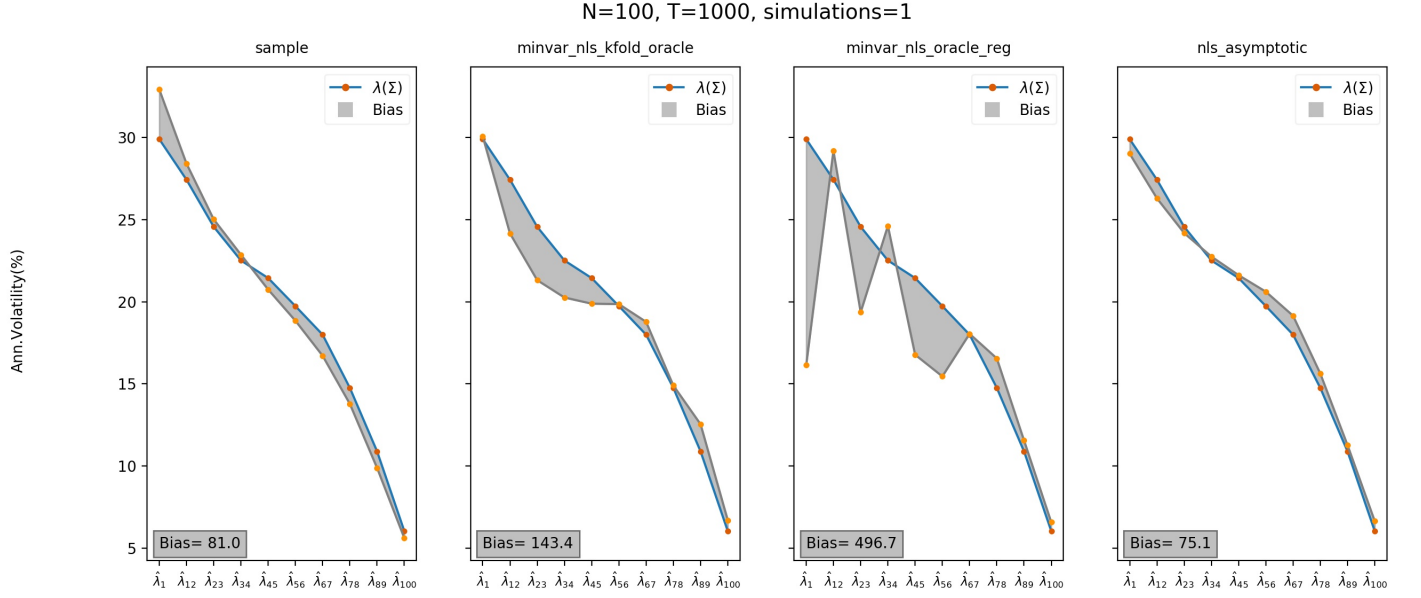


Figure 1c

Second, we fix the number of data samples and features such that concentration ratio $c = T/N$ equals to 2 and vary the number of simulations. Such value of concentration ratio should be sufficient enough to establish a pronounced level of bias and attribute any improvement to the model rather than randomness.

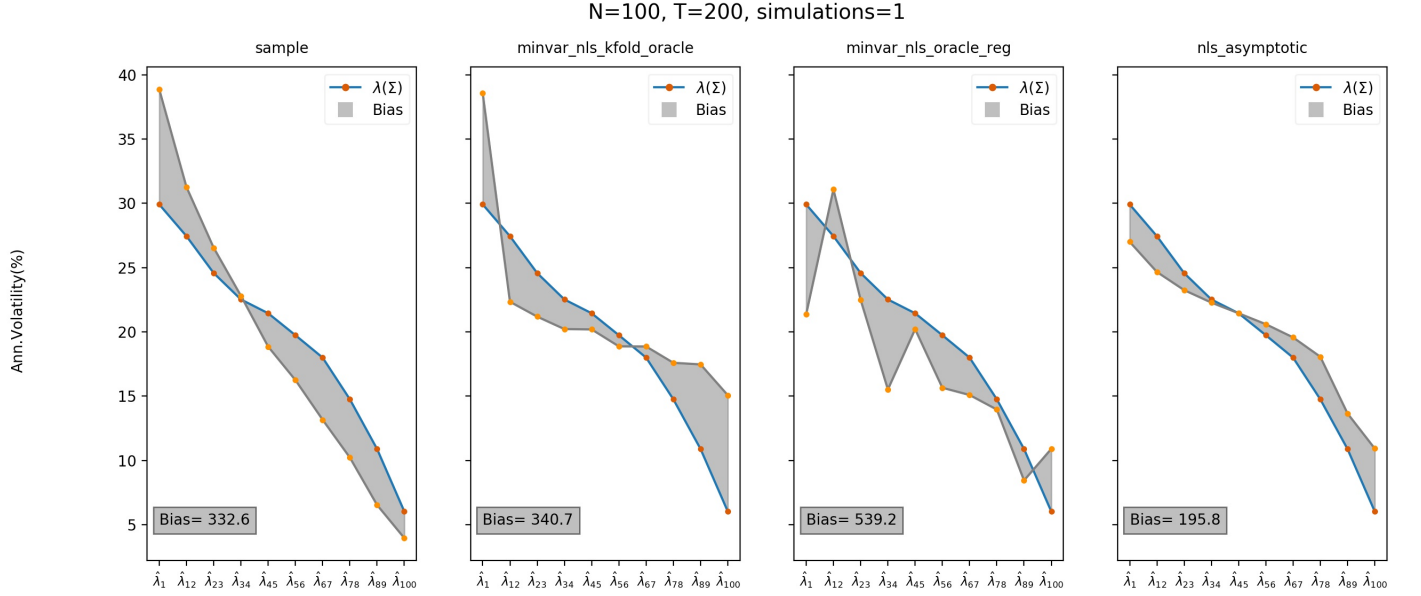


Figure 2a

N=100, T=200, simulations=50

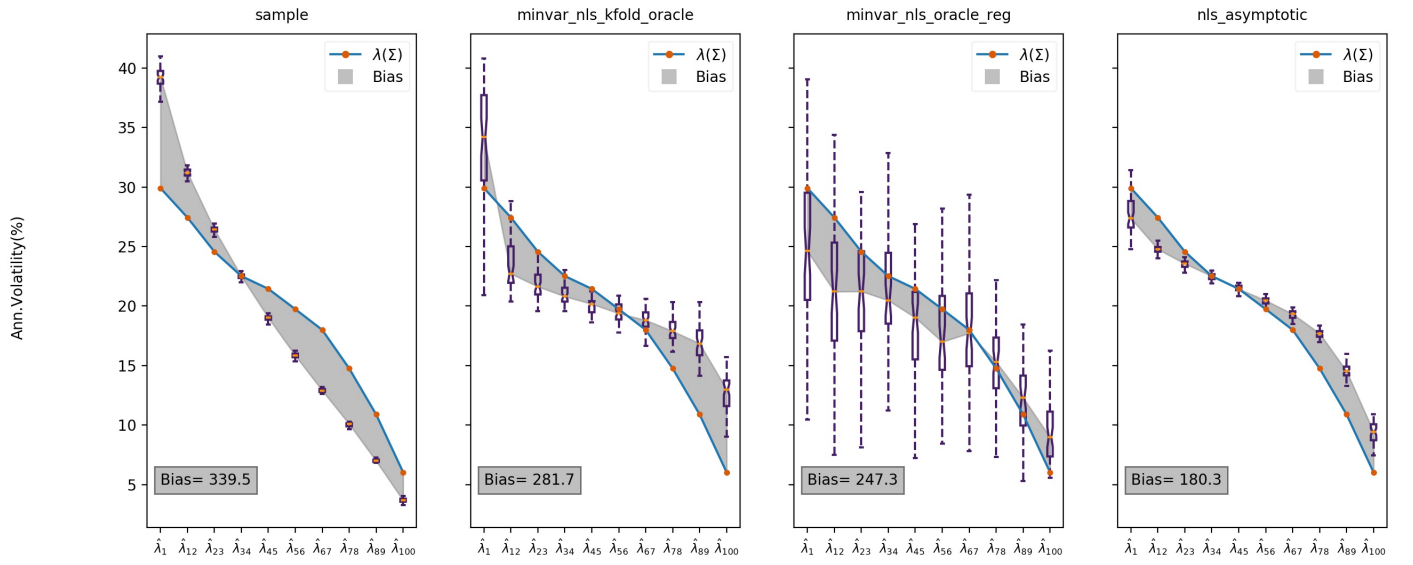


Figure 2b

N=100, T=200, simulations=100

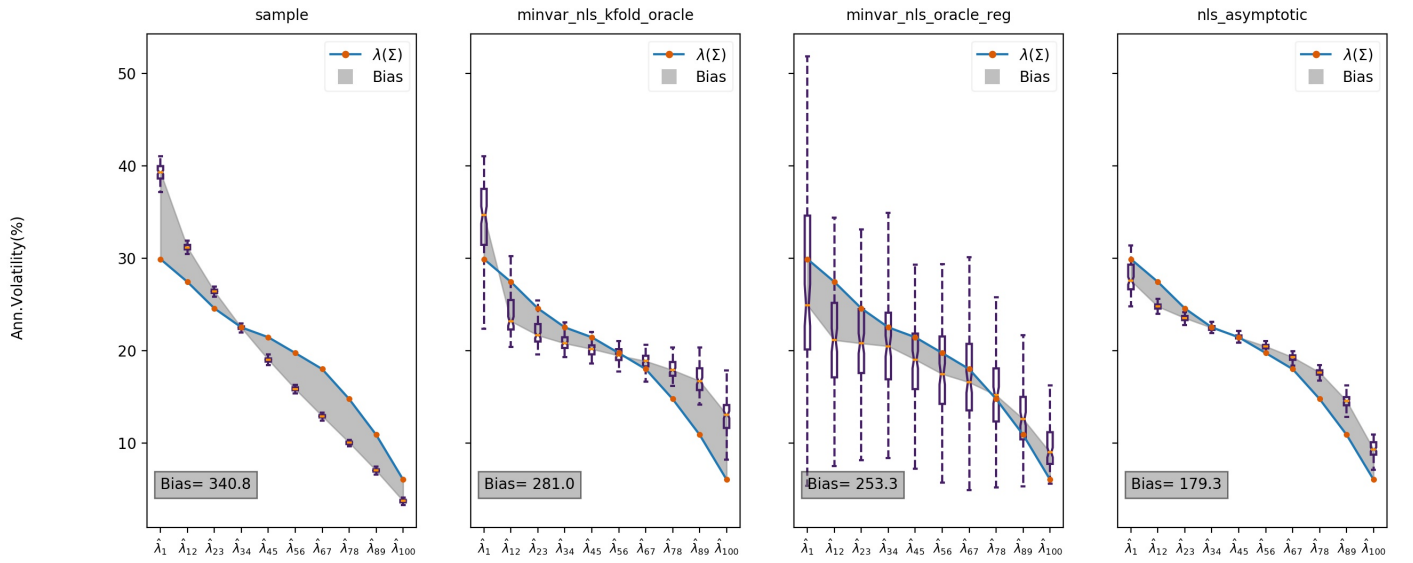


Figure 2c

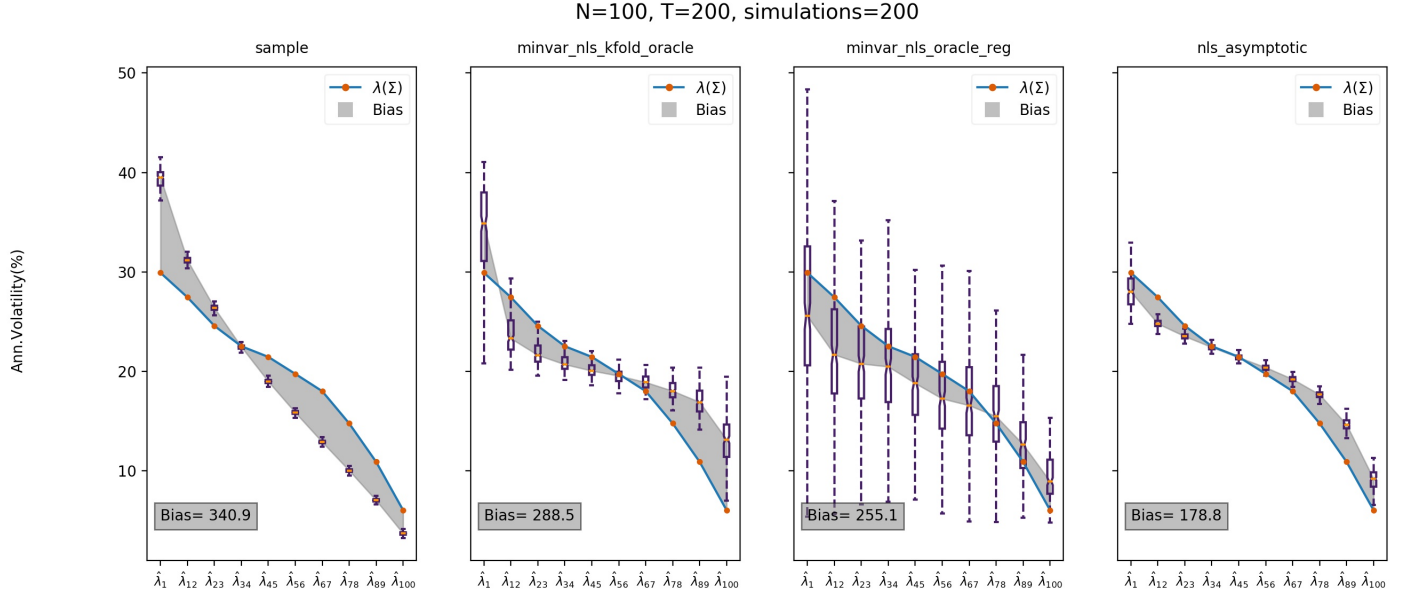


Figure 2d

As a result we can see that on average LW estimator is indisputably most accurate estimator which is not a surprise because of its different loss function. Among Min-Var portfolio estimators NNLSQ with regularization was the best performer with bias lower than that of sample and k-fold cross validation estimator by about 10%. The reason for that is that NNLSQ spectrum oscillates over population eigenvalues. Thus, with higher number of simulations negative bias cancels out positive bias leading to unbiased average. Other estimators, however, suffer from the same bias structure as does the sample estimator: either positive or negative bias on the left side of the spectrum and bias with a reverse sign on the right side.

We should note that we did not adjust λ value for every new simulation setup. We did that for clearer comparability but theoretically it should be respectively corrected.

Also, we found that usage of linear interpolation in NNLSQ case increases the bias on the left (more noisy) side of the spectrum. If one keeps eigenvalues with $z = 0$ as is and proceed with calculation of medians across simulations then the results for NNLSQ estimator will be better (on average) and comparable to those of LW estimator (187 bias for # simulations = 200 in the last test). This is because linear interpolation makes it more difficult for eigenvalues to oscillate : neighbors of the point push it to the area of the same sign and the further they are from the true eigenvalues, the harder it is to get interpolated values with opposite sign leading to a bias.

Further research

1. More research should be done on how to deal with infinite estimated eigenvalues.
2. Change the constrain in NNLSQ routine from non-negative to strictly positive.
3. Amend regularization - we should not penalize z for deviations from 0, because it leads to infinite \hat{d} . Rather, we could track deviations from sample eigenvalues for example.
4. Apply isotonic or kernel regression on \hat{d} values to decrease noise.