Covariance Estimator. The unbiased estimator for the covariance \( C = \frac{1}{n} \sum_{i=1}^{n} x_i x_i^T \) can be recovered as:

\[
\hat{C}_r = C - C_0, \\
\hat{C}_g = \frac{1}{m} \sum_{i=1}^{m} (S_i S_i^T)^{1/2} (S_i S_i^T)^{-1/2} d_i, \\
\hat{C}_l = \frac{1}{m} \sum_{i=1}^{m} (S_i S_i^T)^{-1/2} x_i x_i^T (S_i S_i^T)^{-1/2} d_i,
\]

where \( E[C] = C_0 \), \( C_g \), and \( C_l \) with \( b_{1:k} = \frac{1}{m+1} \).

Remark. In the recovery stage, at most m entries of \( S_i \) and \( d_i \) have to be calculated respectively.

Error Analysis. Let the estimator \( \hat{C}_r \) be defined as in Eq.(1) with the sampling probabilities \( p_{1:k} = P_{1:k} + (1-a) x_i \). Then, with probability at least \( 1 - \eta - \delta \) we have

\[
||C - \hat{C}_r||_F \leq \left( \frac{2\sigma^2 + a\log (2m)}{m} \right)^{1/2},
\]

where \( a \) and \( \sigma \) are as in Eq. (1), \( \eta \) is the desired estimation accuracy for \( C \), and \( \delta \) is the probability of failure when one or more entries of \( S_i \) are not sampled.

Remark. Both \( \eta \) and \( \delta \) decrease when the data size \( n \) increases, and a smaller \( \eta \) and \( \delta \) favor the accuracy of \( \hat{C}_r \). Also, \( \sigma \) should lie between 0 and 1 in order to achieve the smallest error bound.


Accuracy Comparisons

- \( X_1: \eta = 0.81, \eta \geq 0.55 \) (\( \eta = 0.55, \eta \geq 0.55 \)) for \( \hat{C}_g \) and \( \hat{C}_l \)
- \( X_2: \eta = 0.81, \eta \geq 0.55 \) (\( \eta = 0.55, \eta \geq 0.55 \)) for \( \hat{C}_g \) and \( \hat{C}_l \)
- \( X_3: \eta = 0.81, \eta \geq 0.55 \) (\( \eta = 0.55, \eta \geq 0.55 \)) for \( \hat{C}_g \) and \( \hat{C}_l \)


References