Chapter 3
Covariance Matrix Estimation

3.1 Introduction

Covariance matrix estimation allows the adaptation of Gaussian-based mutation operators to local solution space characteristics. The covariance matrix adaptation evolution strategy (CMA-ES) [1] is an example for a highly developed ES using covariance matrix estimation. Besides the cumulative path length control, it iteratively adapts the covariance matrix for the Gaussian mutation operator. In numerous artificial benchmark analyses and real-world applications, covariance matrix estimation has proven its strengths. But it is not restricted to the adaptive update mechanism of the CMA-ES. Efficient implementations allow the fast estimation of the covariance matrix with specialized methods based on the current evolutionary population or based on a training set of successful solutions.

In this chapter, we propose a covariance matrix estimation variant using a (1+1)-ES with Gaussian mutation and Rechenberg's adaptive step size control rule. The idea is to estimate the covariance matrix of the last N successful solutions corresponding to successful generations and to sample from the Gaussian distribution employing the estimated covariance matrix. For covariance matrix estimation, we employ maximum likelihood-based empirical covariance matrix estimation and the Ledoit-Wolf method [2]. The latter shrinks the empirical covariance matrix and is know to be a powerful extension in practice.

This chapter is structured as follows. Section 3.2 gives a brief introduction to covariance matrix estimation focusing on empirical estimation and the Ledoit-Wolf estimator, while Sect. 3.3 presents the covariance matrix estimation ES, which we abbreviate with COV-ES in the following. Related work is discussed in Sect. 3.4. The COV-ES is experimentally analyzed in Sect. 3.5. Conclusions are drawn in Sect. 3.6.

O. Kramer, *Machine Learning for Evolution Strategies*,
Studies in Big Data 20, DOI 10.1007/978-3-319-33383-0_3

3.2 Covariance Matrix Estimation

Covariance matrix estimation is the task to find the unconstrained and statistically interpretable parameters of a covariance matrix. It is still an actively investigated research problem in statistics. Given a set of points $\mathbf{x}_1, \ldots, \mathbf{x}_N \in \mathbb{R}^d$, the task is to find the covariance matrix $\mathbf{C} \in \mathbb{R}^{d \times d}$. The estimation of $\mathbf{C}$ has an important part to play in various fields like time series analysis, classical multivariate statistics and data mining. A common approach for estimating $\mathbf{C}$ is the maximum likelihood approach. The log-likelihood function is maximized by the sample covariance

$$\mathbf{S} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i^T . \tag{3.1}$$

This sample covariance is an unbiased estimator of the population covariance matrix and requires a sufficiently large data set. For covariance matrix estimation in an ES this means that the training set size N has to be large enough for a robust covariance matrix estimation. But as the distribution changes during the optimization run, in particular getting narrower when approximating the optimum, N also has an upper limit. Due to the curse of dimensionality problem the requirements on the populations size grow enormously with increasing problem dimension d. Probability distributions and the maximum likelihood estimation are introduced in detail by Bishop [3].

As the maximum likelihood method is not a good estimator of the eigenvalues of the covariance matrix, the shrinkage method has been introduced. In the Ledoit-Wolf covariance matrix estimation [2], the covariance matrix estimator obtained from Sharpe's single-index model [4] joins the sample covariance matrix in a weighted average. Assume, we have given the sample covariance matrix $\mathbf{S}$. Idea of the shrinkage methods is to compute a weighted covariance matrix from $\mathbf{S}$ and a shrinked but more structured matrix $\mathbf{F}$, which is the shrinkage target

$$\mathbf{C} = \phi \cdot \mathbf{S} + (1 - \phi) \cdot \mathbf{F} \tag{3.2}$$

with weight vector ϕ. $\mathbf{F}$ is also known as sample constant correlation matrix. It is defined via sample variances and average sample correlations, see [2]. The question comes up for the optimal weight ϕ^* between both matrices. For this sake, a loss is defined that measures the deviation of the shrinkage estimator $\mathbf{C}$ and the true covariance matrix Σ

$$L(\phi) = \|\mathbf{C} - \Sigma\|_F^2 \tag{3.3}$$

with Frobenius norm $\|\cdot\|_F^2$. The estimation risk is the corresponding expected value of the loss. From this loss the optimal value ϕ^* is proposed to be chosen as

$$\phi^* = \max\left\{0, \min\left\{\frac{\kappa}{T}, 1\right\}\right\} \tag{3.4}$$

with parameter κ. For a detailed derivation see [2]. We employ the covariance matrix estimators from the SCIKIT- LEARN library.

- The command `from sklearn.covariance import LedoitWolf` imports the Ledoit-Wolf covariance matrix estimator.
- Similarly, `from sklearn.covariance import EmpiricalCovariance` imports the empirical covariance matrix estimator for comparison.
- `LedoitWolf().fit(X)` trains the Ledoit-Wolf estimator with set `X` of patterns. The estimator saves the corresponding covariance matrix in attribute `covariance_`.
- `numpy.linalg.cholesky(C)` computes the Cholesky decomposition of `C`. The result is multiplied with a random Gaussian vector scaled by step size `sigma` with `numpy.dot(C_,sigma * np.random.randn(N))`.

3.3 Algorithm

The integration of the covariance matrix estimation into the ES optimization framework is described in the following. We employ a (1+1)-ES with Rechenberg rule for step size control. The population for the estimation process is based on the history of the best solutions that have been produced in the course of the optimization process.

Algorithm 3 shows the pseudocode of the covariance matrix estimation-based ES with Ledoit-Wolf covariance estimation. In the first phase, the start solution $\mathbf{x}$ is initialized, while the covariance matrix is initialized with the identity matrix $\mathbf{C} = \mathbf{I}$. In the main generational loop, an offspring solution $\mathbf{x}'$ is generated with

$$\mathbf{x}' = \mathbf{x} + \mathbf{z} \tag{3.5}$$

based on the Gaussian distribution using the Cholesky decomposition for computing $\sqrt{\mathbf{C}}$ of covariance matrix $\mathbf{C}$ for employing

$$\mathbf{z} \sim \sigma \cdot \sqrt{\mathbf{C}}\mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{3.6}$$

Step size σ is adapted with Rechenberg's 1/5th success rule. The offspring solution $\mathbf{x}'$ is accepted, if its fitness is superior, i.e., if $f(\mathbf{x}') \leq f(\mathbf{x})$. From the set of the last N successful solutions $\{\mathbf{x}_i\}_{i=1}^{N}$, a new covariance matrix $\mathbf{C}$ is estimated. This training set of successful solutions forms the basis of the covariance matrix estimation population. Novel mutations will be sampled from this estimated covariance matrix and will consequently be similar to the successful past solutions, which allows the local approximation of the solution space and a movement towards promising regions. For the covariance matrix estimation, we use the empirical estimation and the Ledoit-Wolf approach introduced in the previous section. These steps are repeated until a termination condition is fulfilled.

Algorithm 3 COV-ES

initialize $\mathbf{x}$
$\mathbf{C} = \mathbf{I}$
repeat
  adapt σ with Rechenberg
  $\mathbf{z} \sim \sigma \cdot \sqrt{\mathbf{C}}\mathcal{N}(\mathbf{0}, \mathbf{I})$
  $\mathbf{x}' = \mathbf{x} + \mathbf{z}$
  evaluate $\mathbf{x}' \rightarrow f(\mathbf{x}')$
  if $f(\mathbf{x}') \leq f(\mathbf{x})$ **then**
    replace $\mathbf{x}$ with $\mathbf{x}'$
    last N solutions $\rightarrow \{\mathbf{x}_i\}_{i=1}^{N}$
    covariance estimation $\{\mathbf{x}_i\}_{i=1}^{N} \rightarrow \mathbf{C}$
  end if
until termination condition

3.4 Related Work

The CMA-ES by Hansen and Ostermeier [1] is a successful evolutionary optimization method for real-parameter optimization of non-linear problems. It is successfully applied to various domains, e.g., to solving satisfiability in fuzzy logics [5], for tuning prediction models that serve as surrogates of expensive finite-element simulations [6], and to search for good degree distributions with different decoding behavior in luby transform code [7].

The covariance update in each step of the CMA-ES is similar to an approximation of the inverse Hessian matrix and requires a set of different equations [1]. Moreover, the approach is based on a derandomized step size control that takes into account how solutions move in the course of the optimization process. To avoid the computationally expensive complexity of $O(N^3)$ for the Cholesky decomposition in each iteration, Igel et al. [8] propose an alternative covariance matrix update equation. An efficient update rule that allows an update in $O(N^2)$ is proposed by Suttorp et al. [9], which also removes the need for matrix inversion. The CMA-ES is continuously improved. A computationally efficient variant for limited memory is introduced in [10], a variant capable of handling noisy objective functions has been introduced by Kruisselbrink et al. [6]. The CMA-ES is combined with local search by Caraffini et al. [11]. CMA-ES variants are proposed for multi-objective optimization [12–14] and constraint handling [15]. To avoid stagnation in local optima, the CMA-ES can be combined with restarts, which is experimentally analyzed on the CEC 2013 benchmark problems in [16]. A CMA-ES variant based on self-adaptation is the CMSA-ES by Beyer and Sendhoff [17]. It applies self-adaptive step size control, i.e., step sizes are inherited with the solutions they generated. The CMSA-ES employs the usual empirical covariance matrix update. Au and Leung [18] propose to cluster the eigenvalues of the covariance matrix of a CMA-ES and to sample search

points on a mirrored eigenspace spanned by eigenvectors. Krause and Glasmachers [19] propose a multiplicative update rule for the CMA-ES.

EDAs are similar to the COV-ES. As population-based approaches they estimate the distribution of the best solutions assuming a certain shape of the population distribution. An EDA variant based on Ledoit-Wolf estimation is the Shrinkage Estimation of Distribution Algorithm for the Multivariate Norm (SEDA) by Ochoa [20]. It works like a classic EDA without step size adaptation and ES-typical elements.

3.5 Experimental Analysis

In this section, we experimentally analyze the COV-ES by visualizing the estimation during evolutionary runs and by comparing the performance on a short benchmark test set. In the optimal case, the covariance matrix estimation approximates the local contour lines of the fitness function. This allows sampling new mutations according to the local fitness function conditions. Figure 3.1 shows visualizations of the Ledoit-Wolf estimation of the best solutions generated during the run of the ES on the Sphere function and on Rosenbrock. After 200 generations of the (1+1)-ES, the covariance matrix estimation based on the last $N = 100$ best solutions is illustrated. The plots visualize the contour lines of the fitness functions and the contour lines of the covariance matrix estimate. In particular on the Sphere function, the covariance matches the contour lines of the fitness function quite well. On Rosenbrock, we can observe that the search is turning around the curved landscape towards the optimum.

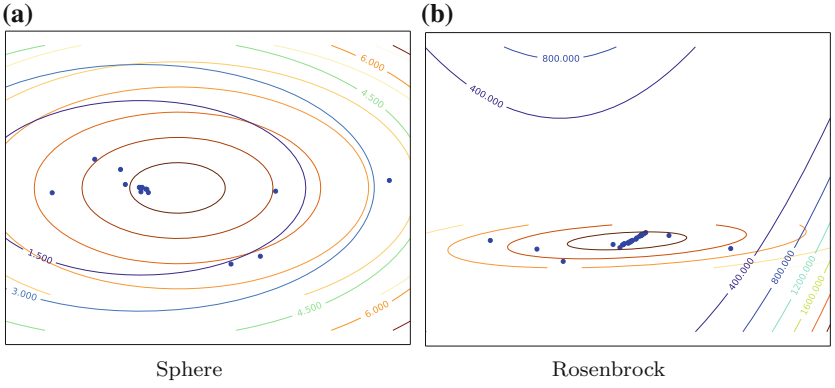

Fig. 3.1 Covariance matrix estimation of last $N = 100$ solutions of a (1+1)-ES after 200 generations. **a** On the Sphere function, Ledoit-Wolf allows a good adaptation of the covariance to the fitness contour lines. **b** On the Rosenbrock function, the covariance is adapting to the curved landscape towards the optimum

In the following, we compare the variants to the standard (1+1)-ES without covariance matrix estimation, i.e., with $\mathbf{C} = \mathbf{I}$ during optimization runs. The Rechenberg rule employs $T = 10$ and $\tau = 0.5$. For the covariance matrix estimation process, a population size of $N = 100$ is used. The ES terminate after 5000 fitness function evaluations. Table 3.1 shows the experimental analysis on the two functions Sphere and Rosenbrock with $d = 5$ and $d = 10$ dimensions. Each experiment is repeated 100 times and the mean values and corresponding standard deviations are shown. The last columns show the p-value of a Wilcoxon test [21] comparing the empirical covariance matrix estimation with the Ledoit-Wolf version.

The results show that the COV-ES outperforms the standard (1+1)-ES without covariance matrix estimation on Rosenbrock in both dimensions. While the Ledoit-Wolf estimation is able to adapt to the narrow valley when approximating the optimum, the standard (1+1)-ES fails. For $d = 5$, the empirical covariance matrix estimation ES variant is better than the classic (1+1)-ES, but is outperformed by the Ledoit-Wolf variant. For $d = 10$, the advantage of the covariance matrix estimation mechanism becomes even more obvious. The low p-values of the Wilcoxon test confirm the statistical significance of the result. On the Sphere function, the (1+1)-ES is slightly, but not significantly superior for $d = 5$, but shows significantly better results than both estimation variants for $d = 10$ (p-value 0.047). This is due to the fact that isotropic Gaussian mutation is the optimal setting on the symmetric curvature of the Sphere. The empirical covariance matrix estimation fails for $d = 10$, where Ledoit-Wolf is significantly superior. The ES with Ledoit-Wolf performs significantly worse on the Cigar with $d = 5$, but better for $d = 10$ dimensions, while no statistical difference can be observed on Griewank.

Figures 3.2 and 3.3 show comparisons of evolutionary runs of the ES with empirical covariance estimation and of the COV-ES on the four benchmark functions with $d = 10$ and a logarithmic scale of the fitness. Again, the ES terminate after 5000 fitness function evaluations. The plots show the mean evolutionary runs, while the

Table 3.1 Experimental comparison between (1+1)-ES and both COV-ES variants. Bold values indicate statistical significance with p-value of Wilcoxon test < 0.05

Problem		(1+1)-ES		Empirical		Ledoit-Wolf		Wilx.
	d	Mean	Dev	Mean	Dev	Mean	Dev	p-Value
Sphere	5	5.91e-31	6.54e-31	3.35e-30	2.97e-30	3.74e-30	3.54e-30	0.197
	10	7.97e-30	8.89e-30	0.177	0.237	4.33e-24	7.50e-24	**0.047**
Rosenbrock	5	0.286	0.278	6.88e-06	1.13e-05	1.69e-23	2.926-23	0.144
	10	0.254	0.194	5.138	4.070	1.23e-17	6.99e-18	**0.007**
Cigar	5	6.855	6.532	0.031	0.080	16.030	18.825	**0.007**
	10	7.780	11.132	2.35e5	4.78e5	17.544	28.433	**0.007**
Griewank	5	0.012	0.008	0.018	0.018	0.012	0.009	0.260
	10	0.010	0.010	0.020	0.014	0.008	0.005	0.144

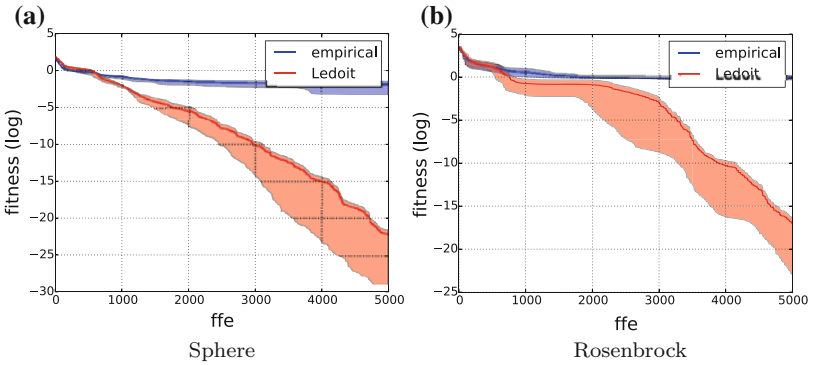

Fig. 3.2 Comparison of empirical covariance matrix estimation and Ledoit-Wolf estimation (**a**) on the Sphere function. The COV-ES allows a logarithmically linear approximation of the optimum. **b** Also on Rosenbrock, the COV-ES allows a logarithmically linear approximation of the optimum

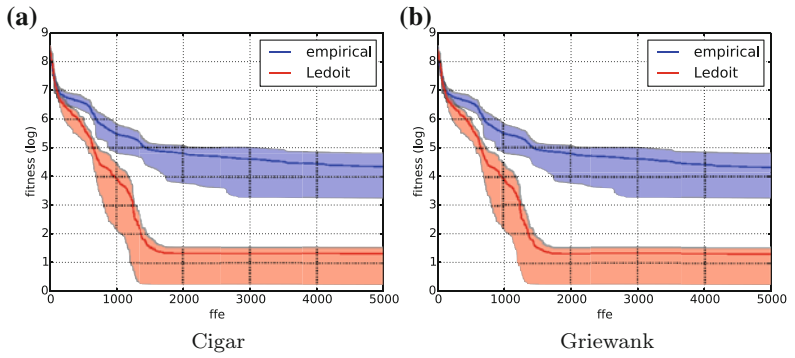

Fig. 3.3 **a** Although the smooth approximation of the optimum fails on the Cigar function, Ledoit-Wolf outperforms the empirical covariance matrix estimation. **b** Also on Griewank, the COV-ES outperforms the empirical covariance matrix estimation

upper and lower parts illustrate the best and worst runs. All other runs lie in the shadowed regions. The figures show that the Ledoit-Wolf variant is superior to the empirical variant on all four problems. Ledoit-Wolf allows a logarithmically linear development on most of the functions, particularly on the Sphere function and Rosenbrock. On Cigar and Griewank, the ES stagnates before reaching the optimum, which has already been shown in Table 3.1.

Last, we analyze the behavior of the CMA-ES with Ledoit-Wolf estimation w.r.t. various covariance matrix estimation training set sizes N and an increasing problem dimensionality on the Sphere function. The experimental results of the COV-ES with corresponding settings are shown in Table 3.2, where the mean fitness values and the standard deviations of 100 runs are shown with a termination of each run after 5000 fitness function evaluations. The best results are marked in bold.

Table 3.2 Analysis of covariance matrix estimation size N on fitness on Sphere function (Sp.) w.r.t. an increasing problem dimensionality d, and on Rosenbrock (Ro.) with $d = 10$

problem	$N = 20$	$N = 50$	$N = 100$
Sp. d=10	**2.83e-35** ± 9.1e-35	5.40e-29 ± 1.7e-28	2.30e-24 ± 5.6e-24
Sp. d=20	**1.39e-13** ± 1.6e-13	4.54e-09 ± 6.6e-09	1.41e-06 ± 2.5e-06
Sp. d=50	**5.59e-04** ± 1.1e-04	5.09e-02 ± 2.4e-02	5.49e-01 ± 2.6e-01
Ro. d=10	1.0017 ± 1.72	4.21e-08 ± 7.3e-08	**4.92e-18** ± 8.4e-18

We can observe that for all problem dimensions d, the best choice on the Sphere function is a low training set size N. There is a lower limit for the choice of N, i.e., values under $N = 20$ can lead to numerical problems. A larger N obviously slows down the optimization speed. This is probably due to the fact that older solutions are not appropriate to allow an estimate of the covariance matrix that can be exploited for good new mutations. The distributions become narrower when approximating the optimum. As expected, the optimization problem becomes harder for larger problem dimensions. The effect of perturbation of the optimization is weakened for large d.

3.6 Conclusions

In this chapter, we introduce an ES based on Ledoit-Wolf covariance matrix estimation. Due to changing distributions during evolutionary optimization processes, an adaptation of the probability distribution for the mutation operator is reasonable. In the exploration phase, the search is broadly scattered in solution space. During convergence phases, the search gets narrower. For Gaussian mutation, a covariance matrix can be employed that flexibly adapts to changing solution space conditions.

Due to efficient and easy-to-use implementations in machine learning libraries, there is no reason to employ alternative update rules that are more efficient to compute but less accurate. The Ledoit-Wolf method outperforms empirical covariance matrix estimation on the tested benchmark problems. Further, it allows smaller population sizes on high-dimensional solution spaces and thus a faster optimization process. In the ES we introduce in this Chapter, we employ the covariance matrix estimation process for a (1+1)-ES managing a training set of the last best solutions. The experimental results show that the evolutionary search can be accelerated significantly. Further, a small archive size turns out to be advantageous, probably due to the argument of changing distributions during the search.

The covariance matrix estimation with Ledoit-Wolf estimation allows a fast and easy modification of ES for the adaptation to certain fitness landscapes and solution spaces. The application of covariance matrix estimation to population-based ES is a straightforward undertaking. In each generation, the new parental population consisting of μ parents is subject to the covariance matrix estimation process.

References

1. Hansen, N., Ostermeier, A.: Adapting arbitrary normal mutation distributions in evolution strategies: The covariance matrix adaptation. In: International Conference on Evolutionary Computation, pp. 312–317 (1996)
2. Ledoit, O., Wolf, M.: Honey, i shrunk the sample covariance matrix. J. Portfolio Manag. **30**(4), 110–119 (2004)
3. Bishop, C.M.: Pattern Recognition and Machine Learning (Information Science and Statistics). Springer (2007)
4. Sharpe, W.F.: A simplified model for portfolio analysis. Manag. Sci. **9**(1), 277–293 (1963)
5. Brys, T., Drugan, M.M., Bosman, P.A.N., Cock, M.D., Nowé, A.: Solving satisfiability in fuzzy logics by mixing CMA-ES. In: Proceedings of the Genetic and Evolutionary Computation Conference, GECCO 2013, pp. 1125–1132 (2013)
6. Kruisselbrink, J.W., Reehuis, E., Deutz, A.H., Bäck, T., Emmerich, M.: Using the uncertainty handling CMA-ES for finding robust optima. In: Proceedings of the 13th Annual Genetic and Evolutionary Computation Conference, GECCO 2011, pp. 877–884. Dublin, Ireland, 12–16 July 2011
7. Chen, C., Chen, Y., Shen, T., Zao, J.K.: On the optimization of degree distributions in LT code with covariance matrix adaptation evolution strategy. In: Proceedings of the IEEE Congress on Evolutionary Computation, CEC 2010, pp. 1–8 (2010)
8. Igel, C., Suttorp, T., Hansen, N.: A computational efficient covariance matrix update and a (1+1)-CMA for evolution strategies. In: Proceedings of the Genetic and Evolutionary Computation Conference, GECCO 2006, pp. 453–460 (2006)
9. Suttorp, T., Hansen, N., Igel, C.: Efficient covariance matrix update for variable metric evolution strategies. Mach. Learn. **75**(2), 167–197 (2009)
10. Loshchilov, I.: A computationally efficient limited memory CMA-ES for large scale optimization. In: Proceedings of the Genetic and Evolutionary Computation Conference, GECCO 2014, pp. 397–404. Vancouver, BC, Canada, 12–16 July 2014
11. Caraffini, F., Iacca, G., Neri, F., Picinali, L., Mininno, E.: A CMA-ES super-fit scheme for the re-sampled inheritance search. In: Proceedings of the IEEE Congress on Evolutionary Computation, CEC 2013, pp. 1123–1130 (2013)
12. Rodrigues, S.M.F., Bauer, P., Bosman, P.A.N.: A novel population-based multi-objective CMA-ES and the impact of different constraint handling techniques. In: Proceedings of the Genetic and Evolutionary Computation Conference, GECCO 2014, pp. 991–998. Vancouver, BC, Canada, 12–16 July 2014
13. Santos, T., Takahashi, R.H.C., Moreira, G.J.P.: A CMA stochastic differential equation approach for many-objective optimization. In: Proceedings of the IEEE Congress on Evolutionary Computation, CEC 2012, pp. 1–6 (2012)
14. Voß, T., Hansen, N., Igel, C.: Improved step size adaptation for the MO-CMA-ES. In: Proceedings of the Genetic and Evolutionary Computation Conference, GECCO 2010, pp. 487–494 (2010)
15. Arnold, D.V., Hansen, N.: A (1+1)-CMA-ES for constrained optimisation. In: Proceedings of the Genetic and Evolutionary Computation Conference, GECCO 2012, pp. 297–304. Philadelphia, PA, USA, 7–11 July 2012
16. Loshchilov, I.: CMA-ES with restarts for solving CEC 2013 benchmark problems. In: Proceedings of the IEEE Congress on Evolutionary Computation, CEC 2013, pp. 369–376 (2013)
17. Beyer, H.G., Sendhoff, B.: Covariance matrix adaptation revisited—the cmsa evolution strategy. In: Proceedings of the 10th Conference on Parallel Problem Solving from Nature, PPSN X 2008, pp. 123–132 (2008)
18. Au, C., Leung, H.: Eigenspace sampling in the mirrored variant of (1, λ)-cma-es. In: Proceedings of the IEEE Congress on Evolutionary Computation, CEC 2012, pp. 1–8 (2012)
19. Krause, O., Glasmachers, T.: A CMA-ES with multiplicative covariance matrix updates. In: Proceedings of the Genetic and Evolutionary Computation Conference, GECCO 2015, pp. 281–288. Madrid, Spain, 11–15 July 2015

20. Ochoa, A.: Opportunities for expensive optimization with estimation of distribution algorithms. In: Tenne, Y., Goh, C.-K. (eds.) Computational Intelligence in Expensive Optimization Problems, pp. 193–218. Springer (2010)
21. Kanji, G.: 100 Statistical Tests. SAGE Publications, London (1993)