Covariance Matrix Estimation for Reinforcement Learning

Tomer Lancewicki[*]
Department of Electrical Engineering and Computer Science
University of Tennessee
Knoxville, TN 37996
tlancewi@utk.edu

Itamar Arel
Department of Electrical Engineering and Computer Science
University of Tennessee
Knoxville, TN 37996
itamar@eecs.utk.edu

Abstract

One of the goals in scaling reinforcement learning (RL) pertains to dealing with high-dimensional and continuous state-action spaces. In order to tackle this problem, recent efforts have focused on harnessing well-developed methodologies from statistical learning, estimation theory and empirical inference. A key related challenge is tuning the many parameters and efficiently addressing numerical problems, such that ultimately efficient RL algorithms could be scaled to real-world problem settings. Methods such as Covariance Matrix Adaptation - Evolutionary Strategy (CMAES), Policy Improvement with Path Integral (PI^2) and their variations heavily depends on the covariance matrix of the noisy data observed by the agent. It is well known that covariance matrix estimation is problematic when the number of samples is relatively small compared to the number of variables. One way to tackle this problem is through the use of shrinkage estimators that offer a compromise between the sample covariance matrix and a well-conditioned matrix (also known as the *target*) with the aim of minimizing the mean-squared error (MSE). Recently, it has been shown that a *Multi-Target Shrinkage Estimator* (MTSE) can greatly improve the single-target variation by utilizing several targets simultaneously. Unlike the computationally complex *cross-validation* (CV) procedure, the shrinkage estimators provide an analytical framework which is an attractive alternative to the CV computing procedure. We consider the application of shrinkage estimators in dealing with a function approximation problem, using the *quadratic discriminant analysis* (QDA) technique and show that a two-target shrinkage estimator generates improved performance. The approach paves the way for improved value function estimation in large-scale RL settings, offering higher efficiency and fewer hyper-parameters.

Keywords: covariance matrix estimation, path integral, classification uncertainty

[*] The authors are with the Machine Intelligence Lab at the University of Tennessee - http://mil.engr.utk.edu

1 Introduction

Reinforcement learning (RL) applied to real-world problems inherently involves combining optimal control theory and dynamic programming methods with learning techniques from statistical estimation theory [1, 2, 3, 4]. The motivation is achieving efficient value function approximation for the non-stationary iterative learning process involved, particularly when the number of state variables exceeds 10 [5]. Recent efforts in scaling RL address continuous state and/or action spaces by optimizing parametrized policies. For example, the *Policy Improvement with Path Integral* (PI^2) [5] combines a derivation from first principles of stochastic optimal control with tools from statistical estimation theory. It has been shown in [6] that PI^2 is a member of a wider family of methods which share probabilistic modeling concepts such as *Covariance Matrix Adaptation - Evolutionary Strategy* (CMAES) [7] and the *Cross-Entropy Methods* (CEM) [8]. The *Path Integral Policy Improvement with Covariance Matrix Adaptation* (PI^2 -CMA) [6] takes advantage on the PI^2 method by determining the magnitude of the exploration noise automatically [6]. The PI^2-SEQ [9] scheme applies PI^2 to sequences of motion primitives. One application of the PI^2-SEQ is concerned with object grasping under uncertainty [9, Sec. 5] while applying the experimental paradigm of [10]. The latter approach has illustrated that over time, humans adapt their reaching motion and grasp to the shape of the object position distribution, determined by the orientation of the main axis of its covariance matrix. Moreover, it has been shown that the PI^2 optimal control policy can be approximated through linear regression [11]. This connection allows the use of well-developed linear regression algorithms for learning the optimal policy. The aforementioned methods rely on accurate covariance matrix estimation of the multivariate data involved. Unfortunately, when the number of observations n is comparable to the number of state variables p the covariance estimation problem become more challenging. In such scenarios, the sample covariance matrix is not well-conditioned and is not necessarily invertible (despite the fact that those two properties are required for most applications). When $n \leq p$, the inversion cannot be computed at all [5, Sec. 2.2].

The same covariance problem arises in other related applications of RL. For example, in RL with Gaussian processes, the covariance matrix is regularized [12, Sec. 2]. However, although the regularization parameter plays a pivotal role, it is not clear how it should be set [12, Sec. 3]. Other related work [13] study the ability to mitigate potentially overconfident classifications by assessing how qualified the system is to make a judgment on the current test datum. It is well known that for a small ratio of training observations n to observation dimensionality p, conventional *Quadratic Discriminant Analysis* (QDA) classifier perform poorly, due to a highly variable class conditional sample covariance matrices. In order to improve the classifiers' performance, regularization is recommended, with the aim of providing an appropriate compromise between the bias and variance of the solution. While other regularization methods [14] define regularization coefficients by the computationally complicated *cross-validation* (CV) procedure, the shrinkage estimators studied in this paper provide an analytical solution, which is an attractive alternative to the CV procedure.

This paper elaborates on the *Multi-Target Shrinkage Estimator* (MTSE) [15] that addresses the problem of covariance matrix estimation when the number of samples is relatively small compared to the number of variables. MTSE offers a compromise between the sample covariance matrix and well-conditioned matrices (also known as *targets*) with the aim of minimizing the mean-squared error (MSE). Section 2 presents the MTSE and examine the squared biases of two diagonal targets. In Section 3, we conduct a careful experimental study and examine the two-target and one-target shrinkage estimator, as well as the *Lediot-Wolf* (LW) [16] method for different covariance matrices. We demonstrate an application for the *quadratic discriminant analysis* (QDA) classifier, showing that the *test classification accuracy rate* (TCAR) is higher when using the two-target, rather than one-target, shrinkage regularization. The QDA classifier is a fundamental component in DeSTIN [17] which is a deep learning system for spatiotemporal feature extraction. The DeSTIN architecture currently assumes diagonal covariance matrices, which is one of the targets examined in this paper. In our future research we intend to utilize the results shown in this paper in order to improve the DeSTIN architecture.

2 Multi-Target Shrinkage Estimation

Let $\{\mathbf{x}_i\}_{i=1}^{n}$ be a sample of independent identical distributed (i.i.d.) p-dimensional vectors drawn from a density having zero mean and covariance $\mathbf{\Sigma} = \{\sigma_{ij}\}$. The most common estimator of $\mathbf{\Sigma}$ is the sample covariance matrix $\mathbf{S} = \{s_{ij}\}$, defined as

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

and is unbiased, i.e., $E\{\mathbf{S}\} = \mathbf{\Sigma}$. The MTSE model [15] defined as

$$\hat{\mathbf{\Sigma}}(\boldsymbol{\gamma}) = \left(1 - \sum_{i=1}^{t}\gamma_i\right)\mathbf{S} + \sum_{i=1}^{t}\gamma_i\mathbf{T}_i, \tag{2}$$

where t is the number of the targets $\mathbf{T}_i$, $i = 1, \ldots, t$ and $\boldsymbol{\gamma} = [\gamma_1, \ldots, \gamma_t]^T$ is the vector of shrinkage coefficients. Our objective is therefore to find $\hat{\boldsymbol{\Sigma}}(\boldsymbol{\gamma})$ (2), which minimizes the MSE loss function

$$L(\boldsymbol{\gamma}) = E\left\{ \left\| \hat{\boldsymbol{\Sigma}}(\boldsymbol{\gamma}) - \boldsymbol{\Sigma} \right\|_F^2 \right\}. \tag{3}$$

The optimal shrinkage coefficient vector $\boldsymbol{\gamma}$ that minimize $L(\boldsymbol{\gamma})$ (3) can be found by using a strictly convex quadratic program [15]. In this paper, we use the two diagonal targets

$$\mathbf{T}_1 = \frac{\mathrm{Tr}(\mathbf{S})}{p}\mathbf{I}, \qquad \mathbf{T}_2 = \mathrm{diag}(\mathbf{S}). \tag{4}$$

Following the developments in [16, Sec. 2.2], the covariance matrix $\boldsymbol{\Sigma}$ can be written as $\boldsymbol{\Sigma} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^T$, where $\mathbf{V}$ and $\boldsymbol{\Lambda}$ are the eigenvector and eigenvalue matrices of $\boldsymbol{\Sigma}$, respectively. The eigenvalues of $\boldsymbol{\Sigma}$ are denoted as $\zeta_i, i = 1, \ldots, p$ in increasing order, i.e., $\zeta_1 \leq \zeta_2 \leq \ldots \leq \zeta_p$, and it is well known that $\sum_{i=1}^{p} \zeta_i = \mathrm{Tr}(\boldsymbol{\Sigma})$. As a result, the squared bias of $\mathbf{T}_1$ with respect to $\boldsymbol{\Sigma}$ can be written as

$$\left\| E\{\mathbf{T}_1\} - \boldsymbol{\Sigma} \right\|_F^2 = \left\| \frac{1}{p}\mathrm{Tr}(\boldsymbol{\Sigma})\mathbf{I} - \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^T \right\|_F^2 = \sum_{i=1}^{p} \left(\zeta_i - \bar{\zeta}\right)^2, \qquad \bar{\zeta} = \frac{\mathrm{Tr}(\boldsymbol{\Sigma})}{p} = \frac{1}{p}\sum_{i=1}^{p} \zeta_i \tag{5}$$

where $\bar{\zeta}$ is the mean of the eigenvalues $\zeta_i, i = 1, \ldots, p$. The above result shows that $\left\| E\{\mathbf{T}_1\} - \boldsymbol{\Sigma} \right\|_F^2$ is equal to the dispersion of the eigenvalues around their mean. Therefore, $\mathbf{T}_1$ becomes less suitable in describing $\boldsymbol{\Sigma}$ when the dispersion of the eigenvalues (5) increases. On the other hand, the expression of the squared bias of $\mathbf{T}_2$ with respect to $\boldsymbol{\Sigma}$ can be written as

$$\left\| E\{\mathbf{T}_2\} - \boldsymbol{\Sigma} \right\|_F^2 = \left\| \mathrm{diag}(\boldsymbol{\Sigma}) - \boldsymbol{\Sigma} \right\|_F^2 = \sum_{i \neq j} \sigma_{ij}, \tag{6}$$

which shows that it is equal to the off-diagonal entries in $\boldsymbol{\Sigma}$. Therefore, $\mathbf{T}_2$ becomes less suitable for describing $\boldsymbol{\Sigma}$ when the p variables of $\boldsymbol{\Sigma}$ are more highly correlated.

3 Experiments

In this section, we present an extensive experimental study of one-target and two-target shrinkage estimators. The estimators are affected by the squared bias and the variance of a target, when the latter depends on the number of data observations n. Therefore, we examine cases of different true covariance matrices $\boldsymbol{\Sigma}$ that result in different biases of $\mathbf{T}_1$ and $\mathbf{T}_2$. We then examine the estimator's performance as a function of n. In order to study the effect of the squared biases, we create a $p \times p$ covariance matrix $\boldsymbol{\Sigma}$ with determinant of one, i.e., $|\boldsymbol{\Sigma}| = 1$, according to two parameters. The first parameter is the condition number η, which is the ratio of the largest eigenvalue ζ_{max} to the smallest eigenvalue ζ_{min} of $\boldsymbol{\Sigma}$, i.e., $\eta = \frac{\zeta_{max}}{\zeta_{min}}$. In the experiments, the p eigenvalues of $\boldsymbol{\Sigma}$ denoted as $\zeta_i, i = 1, 2, \ldots, p$ are generated according to

$$\zeta_i = \zeta_{min}\left((\eta - 1)\frac{(i-1)}{(p-1)} + 1\right),\ i = 1, \ldots, p. \tag{7}$$

Then, the eigenvalue matrix $\boldsymbol{\Sigma}$ is defined as having elements $\zeta_i, i = 1, 2, \ldots, p$ in the matrix form

$$\boldsymbol{\Lambda}(\eta) = \mathrm{diag}(\zeta_1, \zeta_2, \ldots, \zeta_p). \tag{8}$$

The second parameter K, controls the rotation of $\boldsymbol{\Lambda}(\eta)$. Our approach is to select a set of orthonormal transformations, as in [18, Sec. 2.B]

$$\mathbf{E}(K) = \prod_{k=1}^{K} \mathbf{E}_k = \mathbf{E}_1\mathbf{E}_2 \ldots \mathbf{E}_K,\ \text{where each matrix } \mathbf{E}_k \text{is defined as } \mathbf{E}_k = \prod_{l=1}^{p-k} \mathbf{E}_{kl} = \mathbf{E}_{k1}\mathbf{E}_{k2} \ldots \mathbf{E}_{K(p-k)}. \tag{9}$$

The matrix $\mathbf{E}_{kl}$ is an orthonormal rotation of 45^0 in a two-coordinate plane for the coordinates k and $(p + 1 - l)$, i.e.,

$$\mathbf{E}_{kl} = I_{p \times p} + \boldsymbol{\Phi}(k, p + 1 - l), \tag{10}$$

where $\boldsymbol{\Phi}(i_k, j_k)$ is defined as

$$[\boldsymbol{\Phi}]_{ij} = \begin{cases} \frac{1}{\sqrt{2}} - 1 & \text{if } i = j = i_k \text{ or } i = j = j_k \\ \frac{1}{\sqrt{2}} & \text{if } i = i_k \text{ and } j = j_k \\ -\frac{1}{\sqrt{2}} & \text{if } i = j_k \text{ and } j = i_k \\ 0 & otherwise \end{cases}. \tag{11}$$

The parameter K is an integer value with the range $0 \leq K \leq p-1$, where $K=0$ indicates there is no rotation, and $K=p-1$ indicates full rotation, such that all the coordinates rotate with respect to each other at an angle of 45^0. Then, by using $\mathbf{\Lambda}(\eta)$ (8) and $\mathbf{E}$ (9), the covariance matrix is created by

$$\mathbf{\Sigma}(\eta, K) = \mathbf{E}(K)\,\mathbf{\Lambda}(\eta)\,\mathbf{E}^T(K)\,. \tag{12}$$

By employing the covariance matrix (12), the biases of $\mathbf{T}_1$ and $\mathbf{T}_2$ can be controlled independently for $\eta > 1$. The squared bias $\|E\{\mathbf{T}_1\} - \mathbf{\Sigma}\|_F^2$ is affected only by η, and increases as η does, when $\|E\{\mathbf{T}_1\} - \mathbf{\Sigma}\|_F^2 = 0$ for $\eta = 1$. The $\|E\{\mathbf{T}_2\} - \mathbf{\Sigma}\|_F^2$ is affected only by K, and increases as K does, when $\|E\{\mathbf{T}_2\} - \mathbf{\Sigma}\|_F^2 = 0$ for $K = 0$. It should be noted that if $\eta = 1$ then K has no impact while if η is near 1, then K could has minor impact. The shrinkage estimators used in the study are of the one-target variety with $\mathbf{T}_1$ and $\mathbf{T}_2$. In the figures that appear in this section, these estimators are denoted as T1 and T2, respectively. The LW estimator [16] is of the one-target shrinkage variety with $\mathbf{T}_1$, which uses a biased shrinkage coefficient estimator and is denoted as LW. Finally, the two-target shrinkage estimator appears in the figures as TT. We show that the two-target estimator can improve classification results compared with one-target estimators, when using the *quadratic discriminant analysis* (QDA) method. The purpose of the QDA is to assign observations to one of several $g = 1, \ldots, G$ groups with p-variate normal distributions

$$f_g(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^p\,|\mathbf{\Sigma}_g|}} \exp\left(-0.5\,(\mathbf{x} - \mathbf{m}_g)^T\,\mathbf{\Sigma}_g^{-1}\,(\mathbf{x} - \mathbf{m}_g)\right), \tag{13}$$

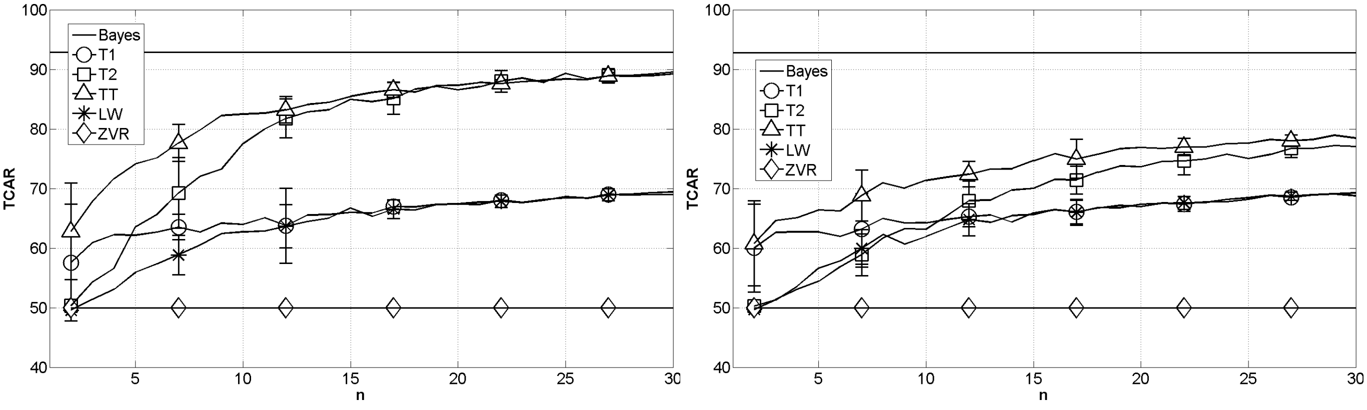
where $\mathbf{m}_g$ and $\mathbf{\Sigma}_g$ are the population mean vector and covariance matrix of the group g. An observation $\mathbf{x}$ is assigned to a class $\hat{g}$ according to

$$d_{\hat{g}}(\mathbf{x}) = \min_{1 \leq g \leq G} d_g(\mathbf{x})\,, \tag{14}$$

with

$$d_g(\mathbf{x}) = (\mathbf{x} - \mathbf{m}_g)^T\,\mathbf{\Sigma}_g^{-1}\,(\mathbf{x} - \mathbf{m}_g) + \ln|\mathbf{\Sigma}_g| - 2\ln\pi_g, \tag{15}$$

where π_g is the unconditional prior probability of observing a member from the group g. In our experiments, we classify two groups ($G = 2$), with observations generated from a normal distribution with zero mean and $\pi_1 = \pi_2$. The covariance matrix of the first group is the identity matrix $\mathbf{\Sigma}_1 = \mathbf{I}$, while that of the second group is the covariance matrix $\mathbf{\Sigma}_2(\eta, K) = \mathbf{\Sigma}(\eta, K)$ (12), which is generated on the basis of the previous experiments. The goal is to study the effectiveness of the shrinkage estimators when using QDA, by assigning observations to one of these two groups, based on the classification rule (14). We run our experiments for $n = 2, 3, \ldots, 30$. For each n, twenty sets of data of size n are produced.



(a)

(b)

Figure 1: QDA for (a) $\mathbf{\Sigma}_2(\eta, 0) = \mathbf{\Lambda}(\eta)$ with $\eta = 10$ and (b) an unrestricted $\mathbf{\Sigma}_2(10, K)$ with $K = 5$

We summarize for each experiment the average *test classification accuracy rate* (TCAR) with standard deviations (the bars in the figure) over the twenty replications for each n. For each group, 10^5 test observations were generated in order to exam the efficiency of the classifier. We provide the best TCAR, calculated by using (14), when the covariance matrices are known, denoted in the figures as Bayes. We also compare the results for a regularization [19, sec. 6], where the zero eigenvalues were replaced with a small number just large enough to permit numerically stable inversion. This has the effect of producing a classification rule based on Euclidean distance in the zero-variance subspace. We denote this procedure as the *zero-variance regularization* (ZVR). In all experiments, the TCAR of the two-target estimator is higher than the one-target variety. The LW estimator is inferior to its unbiased version when dealing with a small number of observations, and converges to its unbiased version as the number of observations increases. Fig. 1(a) presents the result when the covariance matrix is a diagonal matrix, i.e., $\mathbf{\Sigma}_2(\eta, 0) = \mathbf{\Lambda}(\eta)$, with $\eta = 10$, and therefore $\mathbf{T}_2$ is unbiased while $\mathbf{T}_1$ is biased. The target $\mathbf{T}_1$ provides a higher TCAR than $\mathbf{T}_2$ for small numbers of observations, and then $\mathbf{T}_2$ provides a better TCAR. In Fig. 1(b), the covariance matrix is unrestricted, i.e., $\mathbf{\Sigma}_2(10, K)$, with $K = 5$. The targets $\mathbf{T}_1$ and $\mathbf{T}_2$ are biased. The squared bias of $\mathbf{T}_1$ is not affected by K; whereas the higher the value of K, the higher the squared bias of $\mathbf{T}_2$, and therefore $\mathbf{T}_2$ loses its advantage over $\mathbf{T}_1$.

In conclusion, it has been shown that the *Multi-Target Shrinkage Estimator* (MTSE) [15] can greatly improve the single-target variation in the sense of *mean-squared error* (MSE) by utilizing several targets simultaneously. We consider the application of shrinkage estimator in the context of a function approximation problem, using the *quadratic discriminant analysis* (QDA) technique and show that a two-target shrinkage estimator generates improved performance. This is done by a careful experimental study which examines the squared biases of the two diagonal targets. Unlike the computationally complex *cross-validation* (CV) procedure; the shrinkage estimators provide an analytical solution which is an attractive alternative to the CV computing procedure, commonly used in the QDA. The approach paves the way for improved value function estimation in large-scale RL settings, offering higher efficiency and fewer hyper-parameters.

References

[1] P. Dayan and G. E. Hinton, "Using expectation-maximization for reinforcement learning," *Neural Computation*, vol. 9, no. 2, pp. 271–278, 1997.

[2] M. Ghavamzadeh and Y. Engel, "Bayesian actor-critic algorithms," in *Proceedings of the 24th international conference on Machine learning*. ACM, 2007, pp. 297–304.

[3] M. Toussaint and A. Storkey, "Probabilistic inference for solving discrete and continuous state markov decision processes," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 945–952.

[4] N. Vlassis, M. Toussaint, G. Kontes, and S. Piperidis, "Learning model-free robot control by a monte carlo em algorithm," *Autonomous Robots*, vol. 27, no. 2, pp. 123–130, 2009.

[5] E. Theodorou, J. Buchli, and S. Schaal, "A generalized path integral control approach to reinforcement learning," *J. Mach. Learn. Res.*, vol. 11, pp. 3137–3181, Dec. 2010.

[6] F. Stulp and O. Sigaud, "Path integral policy improvement with covariance matrix adaptation," in *Proceedings of the 29th International Conference on Machine Learning (ICML)*, 2012.

[7] N. Hansen and A. Ostermeier, "Completely derandomized self-adaptation in evolution strategies," *Evolutionary Computation*, vol. 9, no. 2, pp. 159–195, June 2001.

[8] S. Mannor, R. Y. Rubinstein, and Y. Gat, "The cross entropy method for fast policy search," in *ICML*, 2003, pp. 512–519.

[9] F. Stulp, E. Theodorou, and S. Schaal, "Reinforcement learning with sequences of motion primitives for robust manipulation," *IEEE Transactions on Robotics*, vol. 28, no. 6, pp. 1360–1370, Dec 2012.

[10] V. N. Christopoulos and P. R. Schrater, "Grasping objects with environmentally induced position uncertainty," *PLoS computational biology*, vol. 5, no. 10, 2009.

[11] F. Farshidian and J. Buchli, "Path integral stochastic optimal control for reinforcement learning," in *The 1st Multidisciplinary Conference on Reinforcement Learning and Decision Making (RLDM2013)*, 2013.

[12] G. Chowdhary, M. Liu, R. Grande, T. Walsh, J. How, and L. Carin, "Off-policy reinforcement learning with gaussian processes," *IEEE/CAA Journal of Automatica Sinica*, vol. 1, no. 3, pp. 227–238, 2014.

[13] H. Grimmett, R. Paul, R. Triebel, and I. Posner, "Knowing when we don't know: Introspective classification for mission-critical decision making," in *2013 IEEE International Conference on Robotics and Automation (ICRA)*, May 2013, pp. 4531–4538.

[14] P. J. Bickel and E. Levina, "Regularized estimation of large covariance matrices," *The Annals of Statistics*, vol. 36, no. 1, pp. pp. 199–227, 2008.

[15] T. Lancewicki and M. Aladjem, "Multi-target shrinkage estimation for covariance matrices," *IEEE Transactions on Signal Processing*, vol. 62, no. 24, pp. 6380–6390, Dec 2014.

[16] O. Ledoit and M. Wolf, "A well-conditioned estimator for large-dimensional covariance matrices," *Journal of Multivariate Analysis*, vol. 88, no. 2, pp. 365 – 411, 2004.

[17] S. Young, J. Lu, J. Holleman, and I. Arel, "On the impact of approximate computation in an analog destin architecture," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 5, pp. 934–946, May 2014.

[18] G. Cao, L. Bachega, and C. Bouman, "The sparse matrix transform for covariance estimation and analysis of high dimensional signals," *IEEE Transactions on Image Processing*, vol. 20, no. 3, pp. 625–640, 2011.

[19] J. H. Friedman, "Regularized discriminant analysis," *Journal of the American Statistical Association*, vol. 84, no. 405, pp. 165–175, 1989.