FIRST-ORDER OPTIMAL SEQUENTIAL SUBSPACE CHANGE-POINT DETECTION

Liyan Xie[*] *George V. Moustakides*[†] *Yao Xie*[*]

[*]Georgia Institute of Technology, School of Industrial and Systems Engineering, Atlanta, GA, USA.
[†]Rutgers University, Department of Computer Science, New Brunswick, NJ, USA.

ABSTRACT

We consider the sequential change-point detection problem of detecting changes that are characterized by a subspace structure. Such changes are frequent in high-dimensional streaming data altering the form of the corresponding covariance matrix. In this work we present a Subspace-CUSUM procedure and demonstrate its first-order asymptotic optimality properties for the case where the subspace structure is unknown and needs to be simultaneously estimated. To achieve this goal we develop a suitable analytical methodology that includes a proper parameter optimization for the proposed detection scheme. Numerical simulations corroborate our theoretical findings.

1. INTRODUCTION

Detecting changes in the distribution of high-dimensional streaming data is a fundamental problem in various applications such as swarm behavior monitoring [1], sensor networks, and seismic event detection. In various scenarios, the change can be represented as a linear subspace which is captured through a change in the covariance structure.

Given a sequence of samples $x_1, x_2, \ldots, x_t$, $t = 1, 2, \ldots$, where $x_t \in \mathbb{R}^k$ and k is the signal dimension, there may be a change-point time τ where the distribution of the data stream changes. Our goal is to detect this change as quickly as possible using on-line techniques. We are particularly interested in the structured change that occurs in the signal covariance. We study two related settings, the *emerging subspace*: meaning that the change is a subspace emerging from a noisy background, and the *switching subspace*: meaning that the change is a switch in the direction of the subspace. The emerging subspace problem can arise from coherent weak signal detection from seismic sensor arrays, and the switching subspace detection can be used for principal component analysis for streaming data. In these settings, the change can be shown to be equivalent to a low-rank component added to the original covariance matrix.

Classical approaches to covariance change detection usually consider generic settings without assuming any structure. The CUSUM statistics can be derived if the pre-change and post-change distributions are known. For the multivariate case, the Hotelling T^2 control chart is the traditional way to detect the covariance changes. The determinant of the sample covariance matrix was also used in [2] to detect change of the determinant of the covariance matrix. A multivariate CUSUM based on likelihood functions of multi-variate Gaussian is studied in [3] but it only considers the covariance change from Σ to $c\Sigma$ for a constant c. Offline change detection of covariance change from Σ_1 to Σ_2 is studied in [4] using the Schwarz information criterion [5], where the change-point location must satisfy certain regularity condition to ensure the existence of the maximum likelihood estimator. Recently, [6] studies the hypothesis test to detect a shift in the off-diagonal sub-matrix planted in the covariance matrix using the likelihood ratios.

In this paper, we propose the Subspace-CUSUM procedure by combing the CUSUM statistic with subspace estimation and proper parameter optimization. We prove that the resulting detector is first-order asymptotically optimal in the sense that the ratio of its expected detection delay with the corresponding of the optimum CUSUM (that has complete knowledge of the pre- and post-change statistics) tends to 1 as the average run length tends to infinity.

The rest of this paper is organized as follows. Section 2 details on the two problems of emerging and switching subspace. Section 3 presents the Subspace-CUSUM procedure. Section 4 considers the asymptotic analysis of the proposed scheme along with parameter optimization and proof of first-order asymptotic optimality. Finally, in Section 5 we present simulation results that corroborate our theoretical findings.

2. SUBSPACE CHANGE-POINT DETECTION

Both settings, emerging and switching subspace, can be shown to be related to the so-called *spiked covariance matrix* [7]. For simplicity, we consider the rank-one spiked covariance matrix problem, which is given by

$$\Sigma = \sigma^2 I_k + \theta u u^\intercal,$$

where I_k denotes an identity matrix of size k; θ the signal strength; $u \in \mathbb{R}^{k\times 1}$ represents a basis for the subspace $\|u\| = 1$ and σ^2 the noise power. We can define the Signal-to-Noise Ratio (SNR) as $\rho = \theta/\sigma^2$.

[*]Work supported by the US National Science Foundation under Grant CCF 1650913, CMMI 1538746, and CCF 1442635.

[†]Work supported by the US National Science Foundation under Grant CIF 1513373, through Rutgers University.

In the *emerging subspace* problem the sequentially observed data are as follows

$$\begin{aligned} x_t &\overset{\text{iid}}{\sim} \mathcal{N}(0,\sigma^2 I_k), & t&=1,2,\ldots,\tau, \\ x_t &\overset{\text{iid}}{\sim} \mathcal{N}(0,\sigma^2 I_k+\theta uu^\intercal), & t&=\tau+1,\tau+2,\ldots \end{aligned} \tag{1}$$

where τ is the unknown change-point that we would like to detect as soon as possible. We assume that the subspace u is unknown since it represents anomaly or new information.

In the *switching subspace* problem the data satisfy

$$\begin{aligned} x_t &\overset{\text{iid}}{\sim} \mathcal{N}(0,\sigma^2 I_k+\theta u_1u_1^\intercal), & t&=1,2,\ldots,\tau, \\ x_t &\overset{\text{iid}}{\sim} \mathcal{N}(0,\sigma^2 I_k+\theta u_2u_2^\intercal), & t&=\tau+1,\tau+2,\ldots \end{aligned} \tag{2}$$

where u_1 and u_2 are the pre- and post-change subspaces. We assume that u_1 is completely known since it describes the statistical behavior under nominal conditions while u_2 is considered unknown since, as before, it expresses an anomaly.

The switching subspace problem (2) can be easily reduced into the emerging subspace problem (1). Indeed if we select any orthonormal matrix $Q \in \mathbb{R}^{(k-1)\times k}$ that satisfies

$$Qu_1 = 0, \quad QQ^\intercal = I_{k-1},$$

and project the observed data onto the space that is orthogonal to u_1 namely $y_t = Qx_t \in \mathbb{R}^{k-1}$, then y_t is a zero-mean random vector with covariance matrix $\sigma^2 I_{k-1}$ before the change and $\sigma^2 I_{k-1}+\tilde{\theta}uu^\intercal$ after the change where $u = Qu_2/\|Qu_2\|$, and

$$\tilde{\theta} = \theta\,\|Qu_2\|^2 = \theta[1-(u_1^\intercal u_2)^2].$$

The data in (2) under this transformation becomes

$$\begin{aligned} y_t &\overset{\text{iid}}{\sim} \mathcal{N}(0,\sigma^2 I_{k-1}), & t&=1,2,\ldots,\tau, \\ y_t &\overset{\text{iid}}{\sim} \mathcal{N}(0,\sigma^2 I_{k-1}+\tilde{\theta}uu^\intercal), & t&=\tau+1,\tau+2,\ldots \end{aligned} \tag{3}$$

which is the emerging subspace problem in (1). We need however to emphasize that by projecting the observations onto a lower dimensional space we lose information, suggesting that the two versions of the problem *are not equivalent*. In particular the optimum detector for the transformed data in (3) and the one of the original data in (2) *do not coincide*. This can be easily verified by computing the corresponding CUSUM tests and their (optimum) performance. Despite this difference, it is clear that with this result we are going to present next, and by adopting the transformed version (3), we offer a computationally simple method to solve the original problem (2). Therefore, from now on, our analysis will focus solely on detecting τ with the ccorresponding observations following the model depicted in (1).

3. SUBSPACE CUSUM

The CUSUM test [8, 9], when the observations are i.i.d. before and after the change, is known to be exactly optimum [10] in the sense that it solves a very well defined constrained optimization problem introduced in [11]. If $f_\infty(x), f_0(x)$ denote the pre- and post-change probability density function (pdf) of the observations respectively then the CUSUM statistic S_t and the corresponding CUSUM stopping time T_{C} are defined [10] as follows

$$S_t = (S_{t-1})^+ + \log\frac{f_0(x_t)}{f_\infty(x_t)}, \;\; T_{\text{C}} = \inf\{t>0: S_t \ge b\}, \tag{4}$$

where $(x)^+ = \max\{x,0\}$ and b denotes a constant threshold. We must of course point out that application of CUSUM is only possible if we have exact knowledge of the pre- and post-change pdfs.

For the data model depicted in (1) the log-likelihood ratio takes the special form

$$\log\frac{f_0(x_t)}{f_\infty(x_t)} = \frac{1}{2\sigma^2}\frac{\rho}{1+\rho}\Big\{(u^\intercal x_t)^2 - \sigma^2\Big(1+\frac{1}{\rho}\Big)\log(1+\rho)\Big\}.$$

The multiplicative factor $\rho/[2\sigma^2(1+\rho)] > 0$ can be omitted since it only performs a constant scaling of the test statistic. We can therefore define the CUSUM test statistic using the following recursion

$$S_t = (S_{t-1})^+ + (u^\intercal x_t)^2 - \sigma^2\left(1+\frac{1}{\rho}\right)\log(1+\rho). \tag{5}$$

Using a simple argument based on Jensen's inequality, we can claim that the increment in (5) has a negative average under the nominal measure and a positive average under the alternative. Due to this property, the CUSUM statistic S_t oscillates near 0 before the change, and increases with a linear trend after the change.

Since in our case we assume that the vector u is unknown we propose the following alternative to (5) with u replaced by any estimate $\hat{u}_t$

$$\mathcal{S}_t = (\mathcal{S}_{t-1})^+ + (\hat{u}_t^\intercal x_t)^2 - d. \tag{6}$$

Quantity d is a constant that we would like to select properly so that the increment of $\mathcal{S}_t$ mimic the main property of the increment of the CUSUM statistic S_t, that is, have a negative mean under nominal and a positive mean under the alternative probability measure. This will require

$$\mathbb{E}_\infty[(\hat{u}_t^\intercal x_t)^2] < d < \mathbb{E}_0[(\hat{u}_t^\intercal x_t)^2]. \tag{7}$$

The *proposed* CUSUM-like stopping time is then defined as

$$\mathcal{T}_{\text{C}} = \inf\{t>0: \mathcal{S}_t \ge b\}. \tag{8}$$

To be able to apply (6) we need to specify d and of course the estimate $\hat{u}_t$. Regarding the latter we propose a sliding window of size w and form the sample covariance matrix

$$\Sigma_t = \textstyle\sum_{i=t+1}^{t+w} x_i x_i^\intercal,$$

using the observations $\{x_{t+1},\dots,x_{t+w}\}$ that lie in *the future* of t. Then $\hat{u}_t$ is simply the unit-norm eigenvector corresponding to the largest eigenvalue of Σ_t. The usage of observations from the future might seem somewhat awkward but it is always possible by properly delaying the data. The main advantage of this idea is that it provides estimates $\hat{u}_t$ that are *independent* from x_t. Of course employing observations from times after t affects the actual performance of our scheme. In particular, if with (8) we stop at time $\mathcal{T}_{\mathrm{C}}=t$ this implies that we used data from times up to $t+w$ and, consequently, $t+w$ is the true time we stop and not t.

The independence between $\hat{u}_t$ and x_t allows for the simple computation of the two expectations in (7). However, for this computation to be possible, especially under the alternative regime, it is necessary to be able to describe the statistical behavior of our estimate $\hat{u}_t$. We will assume that the window size w is sufficiently large so that Central Limit Theorem type approximations are possible for $\hat{u}_t$ and we will consider that $\hat{u}_t$ is actually Gaussian with mean u (the correct vector) and (error) covariance matrix that can be specified, analytically, of being size $1/w$ [12, 13]. Explicit formulas will be given in the Appendix.

Lemma 1. *Adopting the Gaussian approximation for $\hat{u}_t$ we have the following two mean values under the pre- and post-change regime:*

$$\mathbb{E}_\infty[(\hat{u}_t^\intercal x_t)^2]=\sigma^2,$$
$$\mathbb{E}_0[(\hat{u}_t^\intercal x_t)^2]=\sigma^2(1+\rho)\left[1-\frac{k-1}{w\rho}\right].$$

Proof. The proof is given in the Appendix. □

Lemma 1 also suggests that the window size w and the drift d must satisfy

$$\sigma^2<d<\sigma^2(1+\rho)\left(1-\frac{k-1}{w\rho}\right). \tag{9}$$

Necessary condition for this to be true is that $w>(k-1)(1+\rho)/\rho^2$. Actually this constraint is required for the Gaussian approximation to make sense. But in order for the approximation to be efficient we, in fact, need w to be significantly larger than the lower bound. We can see that when the SNR is high ($\rho\gg 1$) then with relatively small window size we can obtain efficient estimates. When on the other hand SNR is low ($\rho\ll 1$) then far larger window sizes are necessary to guarantee validity of the Gaussian approximation.

4. ASYMPTOTIC ANALYSIS

In this section we will provide performance estimates for the optimum CUSUM test (that has all the information regarding the data) and the Subspace-CUSUM test proposed in the previous section. This will allow for the optimum design of the two parameters w,d and for demonstrating that the resulting detector is asymptotically optimum.

In sequential change detection there are two quantities that play vital role in the performance of a detector: a) the average run length (ARL) and b) the expected detection delay (EDD). ARL measures the average period between false alarms while EDD the (worst-case) average detection delay. It is known that CUSUM minimizes the latter while keeps the former above a prescribed level. Let us first compute these two quantities for the case of CUSUM given in (4).

4.1. Asymptotic performance

From [**?**, Pages 396–397] we have that the test depicted in (4) has the following performance

$$\mathbb{E}_\infty[T_{\mathrm{C}}]=\frac{e^b}{\mathsf{K}}\big(1+o(1)\big),\quad \mathbb{E}_0[T_{\mathrm{C}}]=\frac{b}{\mathsf{I}_0}\big(1+o(1)\big), \tag{10}$$

where b is the constant threshold; K is of the order of a constant with its exact value being unimportant for the asymptotic analysis; finally I_0 is the Kullback-Leibler information number $\mathsf{I}_0=\mathbb{E}_0\{\log\left[f_0(x)/f_\infty(x)\right]\}$. We recall that the worst-case average detection delay in CUSUM is equal to $\mathbb{E}_0[T_{\mathrm{C}}]$. This is the reason we consider the computation of this quantity. If now, we impose the constraint that the ARL is equal to $\gamma>1$ and for the asymptotic analysis that $\gamma\to\infty$, then we can compute the threshold b that can achieve this false alarm performance namely $b=(\log\gamma)\big(1+o(1)\big)$. Substituting this value of the threshold in EDD we obtain

$$\mathbb{E}_0[T_{\mathrm{C}}]=\frac{\log\gamma}{\mathsf{I}_0}\big(1+o(1)\big). \tag{11}$$

Applying this formula in our problem we end up with the following optimum performance

$$\mathbb{E}_0[T_{\mathrm{C}}]=\frac{2\log\gamma}{\rho-\log(1+\rho)}\big(1+o(1)\big). \tag{12}$$

For the performance computation of Subspace-CUSUM, since the increment $(\hat{u}_t^\intercal x)^2-d$ in (6) is not a log-likelihood, we cannot use (11) directly. To compute the ARL of $\mathcal{T}_{\mathrm{C}}$ we first find $\delta_\infty>0$ from the solution of the equation

$$\mathbb{E}_\infty[e^{\delta_\infty[(\hat{u}_t^\intercal x_t)^2-d]}]=1 \tag{13}$$

and then we note that $\delta_\infty[(\hat{u}_t^\intercal x)^2-d]$ is the log-likelihood ratio between the two pdfs $f_0=\exp\{\delta_\infty[(\hat{u}_t^\intercal x)^2-d]\}f_\infty$ and f_∞. This allows us to compute the threshold b asymptotically as $b=(\log\gamma)\big(1+o(1)\big)/\delta_\infty$ after assuming that $w=o(\log\gamma)$. Similarly we can find a $\delta_0>0$ and define $\tilde{f}_\infty=\exp\{-\delta_0[(\hat{u}_t^\intercal x_t)^2-d]\}f_0$ so that $\delta_0[(\hat{u}_t^\intercal x_t)^2-d]$ is the log-likelihood ratio between f_0 and $\tilde{f}_\infty$ leading to $\mathbb{E}_0[\mathcal{T}_{\mathrm{C}}]=b\big(1+o(1)\big)/(\mathbb{E}_0[(\hat{u}_t^\intercal x_t)^2]-d)$ with the dependence on δ_0 being in the $o(1)$ term. Substituting b we obtain

$$\mathbb{E}_0[\mathcal{T}_{\mathrm{C}}]=\frac{\log\gamma}{\delta_\infty\big(\mathbb{E}_0[(\hat{u}_t^\intercal x_t)^2]-d\big)}\big(1+o(1)\big)+w, \tag{14}$$

where the last term w is added because we use data from the future of t as we explained before. Parameter δ_∞, defined in

(13), is directly related to d. We show in the Appendix that d can be expressed in terms of δ_∞ as follows

$$d = -\frac{1}{2\delta_\infty}\log(1-2\sigma^2\delta_\infty). \qquad (15)$$

After using Lemma 1 and (15) we obtain the following expression for the EDD:

$$\mathbb{E}_0[\mathcal{T}_{\rm C}] = \frac{\log\gamma(1+o(1))}{\sigma^2\delta_\infty(1+\rho)\left(1-\frac{k-1}{w\rho}\right)+\frac{1}{2}\log(1-2\sigma^2\delta_\infty)} + w. \qquad (16)$$

4.2. Parameter optimization and asymptotic optimality

Note that in the previous equation we have two parameters δ_∞ and w and the goal is to select them so as to minimize the EDD. Therefore if we first fix the window size w we can minimize over δ_∞ (which is equivalent to minimizing with respect to the drift d). We observe that the denominator is a concave function of δ_∞ therefore it exhibits a single maximum. The optimum δ_∞ can be computed by taking the derivative and equating to 0 which leads to a particular δ_∞. Substituting this optimal value we obtain the following minimum EDD:

$$\mathbb{E}_0[\mathcal{T}_{\rm C}] = \frac{2\log\gamma(1+o(1))}{(1+\rho)\left(1-\frac{k-1}{w\rho}\right)-1-\log\left[(1+\rho)\left(1-\frac{k-1}{w\rho}\right)\right]} + w. \qquad (17)$$

Equ. (17) involves only the target ARL level γ and the window size w. If we keep w constant it is easy to verify that the ratio of the EDD of the proposed scheme over the EDD of the optimum CUSUM tends, as $\gamma \to \infty$, to a quantity which is strictly greater than 1. In order to make this ratio tend to 1 and therefore establish asymptotic optimality we need to select the window size w as a function of γ. Actually we can perform this selection optimally by minimizing (17) with respect to w for given γ. The following proposition identifies the optimum window size.

Proposition 1. *For each ARL level γ, the optimal window size that minimizes the corresponding EDD is given by*

$$w^* = \sqrt{\log\gamma}\cdot\frac{\sqrt{2(k-1)}}{\rho-\log(1+\rho)}\big(1+o(1)\big),$$

resulting in an optimal drift

$$d^* = \frac{\sigma^2(1+\rho)\left(1-\frac{k-1}{w^*\rho}\right)}{(1+\rho)\left(1-\frac{k-1}{w^*\rho}\right)-1}\log\left[(1+\rho)\left(1-\frac{k-1}{w^*\rho}\right)\right].$$

Using these optimal parameter values it is straightforward to establish that the corresponding Subspace-CUSUM is first-order asymptotically optimum. This is summarized in our next theorem.

Theorem 1. *As the ARL level $\gamma \to \infty$, the corresponding EDD of the Subspace-CUSUM procedure $\mathcal{T}_{\rm C}$ with the two parameters d and w optimized as above satisfies*

$$\lim_{\gamma\to\infty}\frac{\mathbb{E}_0[\mathcal{T}_{\rm C}]}{\mathbb{E}_0[T_{\rm C}]} = 1. \qquad (18)$$

Proof. As we pointed out, the proof is straightforward. Indeed if we substitute the optimum d and w and then take the ratio with respect to the optimum CUSUM performance depicted in (12) we obtain

$$\frac{\mathbb{E}_0[\mathcal{T}_{\rm C}]}{\mathbb{E}_0[T_{\rm C}]} = 1 + \sqrt{\frac{k-1}{2\log\gamma}} + o(1) \to 1,$$

which proves the desired limit. Even though the ratio tends to 1, we note that $\mathbb{E}_0[\mathcal{T}_{\rm C}] - \mathbb{E}_0[T_{\rm C}] = \Theta(\sqrt{\log\gamma}) \to \infty$. This is corroborated by our simulations (see Fig. 1, red curve). □

5. NUMERICAL EXAMPLES

We present simulations to illustrate the satisfactory performance of Subspace-CUSUM. For comparison, we consider two other detection procedures: one uses the largest eigenvalue of the sample covariance matrix Σ_t as the test statistic while the other is the exact CUSUM assuming all parameters are known (ideal but unrealistic case).

The threshold for each detection procedure is determined through Monte-Carlo simulation so they all have the same ARL. Fig. 1 depicts the EDD versus ARL with the latter under a logarithmic scale. Parameters are selected as follows: $k = 5$, $\theta = 1$, $\sigma^2 = 1$ and window length $w = 20$. Exact CUSUM (black) is compared against Subspace-CUSUM (green) and largest eigenvalue scheme (blue). We see that Subspace-CUSUM has much smaller EDD than the largest eigenvalue procedure while Subspace-CUSUM with optimized window size w (red) is uniformly more efficient. We also consider EDD versus ARL for different w and with numerically optimized w so as to minimize the detection delay for each ARL level. The results appear in Fig. 2, which demonstrate that indeed the optimal w increases with ARL.

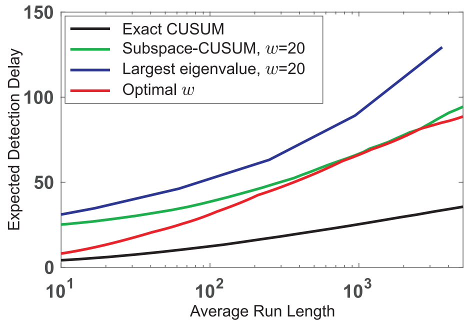

Fig. 1. Comparison of the largest eigenvalue procedure and CUSUM procedures.

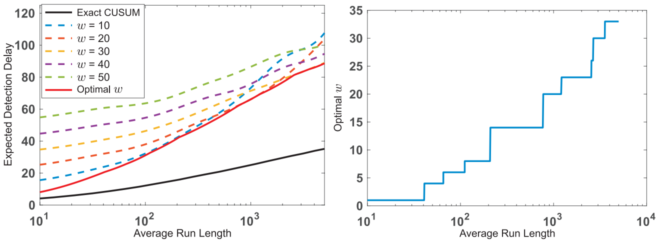

Fig. 2. Left: Minimal EDD (red) among window sizes w from 1 to 50; Right: Corresponding optimal window size w.

6. REFERENCES

[1] Matthew Berger, Lee M Seversky, and Daniel S Brown, "Classifying swarm behavior via compressive subspace learning," in *Robotics and Automation (ICRA), 2016 IEEE International Conference on*. IEEE, 2016, pp. 5328–5335.

[2] Frank B Alt, "Multivariate quality control," *Encyclopedia of Statistical Sciences*, 2004.

[3] John D Healy, "A note on multivariate cusum procedures," *Technometrics*, vol. 29, no. 4, pp. 409–412, 1987.

[4] Jie Chen and AK Gupta, "Statistical inference of covariance change points in gaussian model," *Statistics*, vol. 38, no. 1, pp. 17–28, 2004.

[5] Gideon Schwarz et al., "Estimating the dimension of a model," *The annals of statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[6] Ery Arias-Castro, Sébastien Bubeck, Gábor Lugosi, et al., "Detection of correlations," *The Annals of Statistics*, vol. 40, no. 1, pp. 412–435, 2012.

[7] Iain M Johnstone, "On the distribution of the largest eigenvalue in principal components analysis," *Annals of statistics*, pp. 295–327, 2001.

[8] Ewan S Page, "Continuous inspection schemes," *Biometrika*, vol. 41, no. 1/2, pp. 100–115, 1954.

[9] David Siegmund, *Sequential analysis: tests and confidence intervals*, Springer Science & Business Media, 1985.

[10] George V Moustakides, "Optimal stopping times for detecting changes in distributions," *The Annals of Statistics*, pp. 1379–1387, 1986.

[11] Gary Lorden, "Procedures for reacting to a change in distribution," *The Annals of Mathematical Statistics*, pp. 1897–1908, 1971.

[12] Theodore Wilbur Anderson, "Asymptotic theory for principal component analysis," *The Annals of Mathematical Statistics*, vol. 34, no. 1, pp. 122–148, 1963.

[13] Debashis Paul, "Asymptotics of sample eigenstructure for a large dimensional spiked covariance model," *Statistica Sinica*, pp. 1617–1642, 2007.

A. APPENDIX

Proof of Lemma 1. Using the independence between $\hat{u}_t$ and x_t we can write

$$\mathbb{E}[(\hat{u}_t^\intercal x_t)^2] = \mathbb{E}\big[\hat{u}_t^\intercal \mathbb{E}[x_t x_t^\intercal]\hat{u}_t\big]. \tag{19}$$

Consequently, under the nominal regime

$$\mathbb{E}_\infty[(\hat{u}_t^\intercal x_t)^2] = \mathbb{E}_\infty[\hat{u}_t^\intercal \sigma^2 I_k \hat{u}_t] = \sigma^2,$$

with the last equality being true because $\hat{u}_t$ is of unit norm.

Under the alternative regime we are going to use Central Limit Theorem arguments [12, 13] that describe the statistical behavior of the estimator. We have that

$$\sqrt{w}(\omega_t - u) \to \mathcal{N}\left(0, \frac{1+\rho}{\rho^2}(I_k - uu^\intercal)\right)$$

where the limit is in distribution as $w \to \infty$ and ω_t denotes the *un-normalized* eigenvector. For large w we can write $\omega_t = u + v_t$ where

$$v_t \sim \mathcal{N}\left(0, \frac{1+\rho}{w\rho^2}(I_k - uu^\intercal)\right).$$

Our estimator $\hat{u}_t$ is related to ω_t through the normalization process $\hat{u}_t = \omega_t/\|\omega_t\|$, and if we use this in (19) after recalling that under the alternative $\mathbb{E}_0[x_t x_t^\intercal] = \sigma^2(I_k + \rho uu^\intercal)$ and using repeatedly the fact that u and v_t are orthogonal, we have

$$\begin{aligned}
&\mathbb{E}_0[(\hat{u}_t^\intercal x_t)^2] = \sigma^2 \mathbb{E}_0\left[\hat{u}_t^\intercal (I_k + \rho uu^\intercal)\hat{u}_t\right] \\
&\quad = \sigma^2(1 + \rho\mathbb{E}_0[(\hat{u}_t^\intercal u)^2]) = \sigma^2\left(1 + \rho\mathbb{E}_0\left[\frac{1}{1+\|v_t\|^2}\right]\right) \\
&\approx \sigma^2\left(1 + \rho\mathbb{E}_0\left[1 - \|v_t\|^2\right]\right) = \sigma^2(1+\rho)\left(1 - \frac{k-1}{w\rho}\right).
\end{aligned}$$

For the approximate equality we used the fact that to a first order approximation we can write $1/(1+\|v_t\|^2) \approx 1 - \|v_t\|^2$ because $\|v_t\|^2$ is of the order of $1/w$ while the approximation error is of higher order. This completes the proof. □

Proof of Proposition 1. Let us first evaluate the expectation in (13) to demonstrate the relationship between δ_∞ and d depicted in (15). Using standard computations involving Gaussian random vectors we can write

$$\begin{aligned}
&\mathbb{E}_\infty[e^{\delta_\infty[(\hat{u}_t^\intercal x_t)^2 - d]}] = e^{-\delta_\infty d}\mathbb{E}_\infty\left[\mathbb{E}_\infty[e^{\delta_\infty(\hat{u}_t^\intercal x_t)^2}|\hat{u}_t]\right] \\
&\quad = e^{-\delta_\infty d}\mathbb{E}_\infty\left[\int e^{\delta_\infty x_t^\intercal(\hat{u}_t\hat{u}_t^\intercal)x_t} \cdot \frac{e^{-x_t^\intercal x_t/(2\sigma^2)}}{\sqrt{(2\pi)^k \sigma^{2k}}}dx_t\right] \\
&\qquad\qquad = \frac{e^{-\delta_\infty d}}{\sqrt{1 - 2\sigma^2\delta_\infty}}.
\end{aligned}$$

To compute the integral we used the standard technique of "completing the square" in the exponent and with proper normalization generate an alternative Gaussian pdf which integrates to 1. The interesting observation is that the result of the integration does not actually depend on $\hat{u}_t$.

If we use the optimum value for d in terms of δ_∞ then as we argued in the text we obtain for EDD the expression appearing in (16). We can now fix w and optimize EDD with respect to δ_∞. This is a straightforward process since it amounts in maximizing the denominator. Taking the derivative and equating to 0 yields the optimum δ_∞

$$\delta_\infty^* = \frac{1}{2\sigma^2}\left(1 - \frac{1}{(1+\rho)\left(1 - \frac{k-1}{w\rho}\right)}\right).$$

Substituting this value in (16) produces (17).

The next step consists in minimizing (17) with respect to w. Again taking the derivative and equating to 0 we can show that the optimum window size is the w^* depicted in the proposition. □