2017 IEEE 18th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)

Date of Conference: 3-6 July 2017

Entropy-Based Covariance Determinant Estimation

Ferran De Cabrera, *Student Member, IEEE*, Jaume Riba, *Senior Member, IEEE*, and Gregori Vázquez, *Senior Member, IEEE*
Signal Theory and Communications Department, Technical University of Catalonia (UPC)
{ferran.de.cabrera, jaume.riba, gregori.vazquez}@upc.edu

DOI: 10.1109/SPAWC.2017.8227751

***Abstract*—An information-theoretic approach is described to estimate the determinant of the covariance matrix of a random vector sequence (a common task in a wide range of estimation and detection problems in signal processing for communications). The method is based on a prior entropy-based processing of the data using kernels and offers robustness against small-entropy contamination. The trade-off between optimality, accuracy and robustness is analyzed, along with the impact of the relative kernel bandwidth and data size.**

***Index Terms*—Rényi entropy, Information Potential, U-Statistics, Hadamard Ratio, Kernel Methods, Cramér-Rao Bound (CRB), Spectrum Sensing, Cognitive Radio.**

I. Introduction

The need for robust signal processing arises in those applications where the distributional assumptions on the data do not hold in practice [1]. The goal is to develop methods capable of trading-off some efficiency at the nominal model to gain resistance against the effects of deviations. In particular, robust estimation of the covariance matrix of a vector sequence has been a research topic for decades. In the specific context of communications, efficient spectrum sensing algorithms for cognitive radio systems should be robust against impulsive mad-made noises that are present in practical communications systems (see [2] and references therein). A possible way to address this issue is to assume that the underlying distribution is some heavy-tailed elliptical distribution (see [3] and the seminal works [4] and [5]). In some situations, only the determinant of the covariance matrix is needed. For example, the Hadamard ratio, i.e. the determinant of the sample covariance matrix over the product of its diagonal elements, is the Generalized Likelihood Ratio Test (GLRT) of whether or not a composite covariance matrix is block diagonal in the case of Gaussian data [6]. For spectrum sensing using uncalibrated multiantenna secondary receivers in the context of cognitive radio, detecting the block-diagonal structure of a covariance matrix becomes a crucial task. Moreover, the Hadamard ratio is the core of the generalized coherence [7], a natural generalization of the magnitude-squared coherence (MSC) statistic that is widely used for non-parametric detection of a common signal on two noisy communications channels.

In a different research direction, the estimation of entropy, mutual information and divergence (jointly with their many different variants), have found numerous applications [8] apart from their prominent role in information theory. For example, information theory descriptors can be used in machine learning as non-parametric cost functions for the design of adaptive systems [9]. In this context, universal estimates, i.e. those which do not assume knowledge of the statistical properties of the observed data, are usually required. Most of these estimates are based on non-parametric density estimates. In particular, Parzen density estimation (see [9] and references therein) clearly links information theory with kernel methods, which have become established techniques in the last fifteen years to perform nonlinear signal processing. This link is evident in the case of estimating Rényi second-order entropy (a generalized notion of Shannon differential entropy that still satisfies important properties of the former) from Parzen density estimates.

In this paper, kernel methods are proposed as a tool for robust estimation. The main motivation is that the entropy depends mainly on the probabilities of the events and not on the magnitude of them. The focus is on those cases where the observed random signal is contaminated by other signal showing much less entropy than the former although probably a higher variance. This is typical in practice in those applications where large-valued impulsive outliers or abrupt changes on the mean could be observed. The main novelty with respect to other robust methods for estimating the covariance matrix is that here the interest is to estimate its determinant and not the overall matrix, opening the possibility of estimating the determinant from the entropy, extracted directly from the data.

II. Information potential estimation

The Information Potential (IP) [9] (the argument of the log in the second order Rényi entropy) of a continuous M-dimensional random vector $\mathbf{x}$ with Probability Density Function (p.d.f.) $f(\mathbf{x})$ is defined as

$$V = \int f^2(\mathbf{x}) d\mathbf{x} \tag{1}$$

In the case that the samples are distributed as $\mathcal{CN}(0, \mathbf{\Sigma})$ (referred to as nominal conditions) it can be easily shown that

$$V = (2\pi)^{-M} |\mathbf{\Sigma}|^{-1} \tag{2}$$

i.e. the IP is inversely proportional to the determinant $|\mathbf{\Sigma}|$ of the covariance matrix and insensitive to the mean (a property inherent to any entropy measure).

If a Parzen density estimator with Gaussian kernel function is used to estimate the p.d.f. from N i.i.d. samples of $\mathbf{x}$ ($\{\mathbf{x}_i\}_{1 \le i \le N}$) the resulting IP estimator becomes:

$$\hat{V} = \frac{1}{N^2} \sum_{1 \le i \le N} \sum_{1 \le j \le N} k_{\mathbf{W}} (\mathbf{x}_i - \mathbf{x}_j) = \frac{1}{N} + \frac{\hat{U}}{(2\pi)^M |\mathbf{W}|}$$

This work has been supported by projects COMPASS, TEC2013-47020-C2-2-R and WINTER, TEC2016-76409-C2-1-R (AEI/FEDER, UE).

978-1-5090-3009-5/17/$31.00 ©2017 IEEE

where $k(\mathbf{z}) = e^{-\mathbf{z}^H \mathbf{W}^{-1} \mathbf{z}}$ is a Gaussian kernel with $\mathbf{W}$ a diagonal (possibly data-dependent) kernel matrix, and

$$\hat{U} = \frac{2}{N(N-1)} \sum_{1 \le i < j \le N} k\left(\mathbf{x}_i - \mathbf{x}_j\right) \tag{3}$$

is a U-statistics [10] [11] (i.e. an unbiased estimate of $E[k\left(\mathbf{x}_i - \mathbf{x}_j\right)]$) computed from the data, very similar to the kernelized energy detection proposed in [2] in the context of cognitive radio. Although a direct computation of $\hat{U}$ would imply a complexity of $O(N^2)$, well-known techniques such as the incomplete Cholesky decomposition [12][9], which exploits the reduced Gram matrix band structure, can be used to achieve a computation complexity of $O(N)$, just the same as the sample covariance.

The following result (see Appendix VII-1) will be used for the analysis of $\hat{U}$: if $\mathbf{u} \sim \mathcal{CN}(\bar{\mathbf{u}}, \mathbf{C})$, $\mathbf{v} \sim \mathcal{CN}(\bar{\mathbf{v}}, \mathbf{C})$, and $E\left[\mathbf{u}\mathbf{v}^H\right] = \gamma \mathbf{C}$, then

$$E\left[k(\mathbf{u})\right] = |\mathbf{W}| \left|\mathbf{W} + \mathbf{C}\right|^{-1} \tag{4}$$

$$E\left[k(\mathbf{u})k(\mathbf{v})\right] = \frac{|\mathbf{W}|^2}{\left|\mathbf{W} + (1 - |\gamma|)\,\mathbf{C}\right| \left|\mathbf{W} + (1 + |\gamma|)\,\mathbf{C}\right|} \tag{5}$$

Defining $\mathbf{z} = \mathbf{x}_i - \mathbf{x}_j \sim CN\left(0, 2\mathbf{\Sigma}\right)$, we have from (4)

$$\bar{U} = E\left[\hat{U}\right] = E\left[k(\mathbf{z})\right] = |\mathbf{W}| \left|\mathbf{W} + 2\mathbf{\Sigma}\right|^{-1} \tag{6}$$

and (see Appendix VII-2):

$$\sigma^2_{\hat{U}} = \frac{aN(N-1)(N-2) + bN(N-1)/2}{\left(N(N-1)/2\right)^2} \tag{7}$$

with

$$a = \frac{|\mathbf{W}|^2}{|\mathbf{W} + \mathbf{\Sigma}| \left|\mathbf{W} + 3\mathbf{\Sigma}\right|} - \frac{|\mathbf{W}|^2}{\left|\mathbf{W} + 2\mathbf{\Sigma}\right|^2} \tag{8}$$

$$b = \frac{|\mathbf{W}|^2}{|\mathbf{W}| \left|\mathbf{W} + 4\mathbf{\Sigma}\right|} - \frac{|\mathbf{W}|^2}{\left|\mathbf{W} + 2\mathbf{\Sigma}\right|^2} \tag{9}$$

Note that, for any finite value of a and b, $\hat{U}$ will be consistent given the impact of b in (7) becomes asymptotically negligible. However, b cannot be neglected (as proposed in [9]) to characterize $\sigma^2_{\hat{U}}$ because it tends to zero more slowly than a for $|\mathbf{W}| \to 0$, as seen in (9), and small bandwidth values will precisely be more adequate to gain robustness against contamination.

III. Covariance determinant estimation

The direct approach would be estimating the covariance matrix as the $\hat{\mathbf{\Sigma}}_S = \frac{1}{N-1} \sum_{i=1}^{N} \left(\mathbf{x}_i - \bar{\mathbf{x}}\right) \left(\mathbf{x}_i - \bar{\mathbf{x}}\right)^H$, where $\bar{\mathbf{x}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i$ is the sample mean. The U-statistics [11] version of it is

$$\hat{\mathbf{\Sigma}}_S = \frac{2}{N(N-1)} \sum_{1 \le i < j \le N} \frac{1}{2} \left(\mathbf{x}_i - \mathbf{x}_j\right) \left(\mathbf{x}_i - \mathbf{x}_j\right)^H, \tag{10}$$

which has the advantage of avoiding any explicit estimation of the sample mean, and highlighting the link with (3). The determinant would be finally estimated as $\hat{D}_K = \left|\hat{\mathbf{\Sigma}}_S\right|$.

Alternatively, a second possibility is proposed: estimating first $\hat{U}$ and then estimating the determinant directly as $\hat{D}_K = g_{\mathbf{W}}(\hat{U})$, where $g_{\mathbf{W}}(.)$ is a kernel-dependent monotonic decreasing function. Using this rationale, and focusing to the univariate case for clarity and space reasons, we obtain from (6) the following monotonic relationship between $\bar{U}$ and the variance Σ:

$$\bar{U} = \left(W/\Sigma\right) \left(W/\Sigma + 2\right)^{-1} \tag{11}$$

from which, assuming that a sample moment $\hat{U}$ has been obtained from the U-statistics in Eq. (3), we can apply the method of moments to estimate $\hat{\Sigma}$ as

$$\hat{\Sigma} = g_W\left(\hat{U}\right) = \left(\hat{U}^{-1} - 1\right) W/2. \tag{12}$$

IV. Performance analysis

Using Jensen's inequality in Eq. (12):

$$E\left[\hat{\Sigma}\right] - \Sigma \ge \left(\bar{U}^{-1} - 1\right) W/2 - \Sigma = 0, \tag{13}$$

which means that the bias of $\hat{\Sigma}$ is strictly positive. However, as $\hat{U}$ is consistent and unbiased, it converges in probability to $\bar{U}$, which implies that $\hat{\Sigma}$ converges in probability to Σ, i.e., $\hat{\Sigma}$ is asymptotically unbiased.

Defining the relative bandwidth as $w = W/\Sigma$, the relative variance can be written as (see Appendix VII-3):

$$\bar{\sigma}^2_{\hat{\Sigma}} = \sigma^2_{\hat{\Sigma}}/\Sigma^2 \approx \sigma^2_{\hat{U}} w^{-2} (w+2)^4 \left(1 + 3\bar{U}^{-2}\sigma^2_{\hat{U}}\right)/4 \tag{14}$$

where $\sigma^2_{\hat{U}}$ is defined in (7) with constants a and b in (8)&(9) given by:

$$a = w^2 \left((w+1)^{-1}(w+3)^{-1} - (w+2)^{-2}\right) \tag{15}$$

$$b = w^2 \left(w^{-1}(w+4)^{-1} - (w+2)^{-2}\right). \tag{16}$$

Note that when $w \to 0$ the variance of the proposed estimator tends to infinity, irrespective of the fact that $a \to 0$ and $b \to 0$. The reason is that b goes to zero as $O(w)$ (instead of $O(w^2)$) and this is why we didn't neglect it in (7).

A. Asymptotic performance and CRB

For any $w > 0$, we have from (7) that $lim_{N\to\infty} N\sigma^2_{\hat{U}} = 4a$. Therefore, using (14), (15) and (16), we can state that

$$1 \le lim_{N\to\infty} N\bar{\sigma}^2_{\hat{\Sigma}} = \frac{(w+2)^2}{(w+1)(w+3)} \le \frac{4}{3} \tag{17}$$

with lower and upper-bounds achieved for $w \to 0$ and $w \to \infty$, respectively. The previous equation quantifies the asymptotic penalty on the estimator variance as a function of the kernel bandwidth. Note that the sample mean estimator of variance ($\hat{\Sigma}_S$) in the nominal conditions is efficient and fulfills that $lim_{N\to\infty} N\bar{\sigma}^2_{\hat{\Sigma}_S} = 1$. Therefore, the previous equation shows in particular that the proposed estimator is asymptotically efficient as the kernel size tends to infinity ($w \to \infty$). As w decreases, the asymptotic variance of the proposed estimator is increased with respect to the CRB, but never more than $4/3$ (the maximum asymptotic penalty).

B. Threshold effect

From (7) it is clear that the asymptotic analysis assu that:

$$bN(N-1)/2 < aN(N-1)(N-2)/L$$

where $L \gg 1$. Using Eqs. (15) and (16), (18) can be e stated as:

$$N > 2\left(L\frac{(w+1)(w+3)}{w(w+4)}+1\right),$$

which establishes the condition for the asymptotic analys be valid. The previous equation shows the interplay betwee and w. In particular, the lower is the relative kernel bandw w, the higher should be the value of N to guarantee the estimator reaches the asymptotic regime. For small values of w violating the condition, the estimator variance will be highly amplified. The condition is also useful to determine the minimum value of w that can be used as a function of N. If we fix, for example, $L = 10$ and assume very small w, we obtain that a rough value of the minimum allowable relative kernel size is

$$w_{min} \approx 15/N \tag{20}$$

We can then assure that for $w > 15/N$ the estimator variance in nominal conditions will not be more amplified than (roughly) a factor of $4/3$ with respect to the CRB[1].

C. Robustness

Let us assume an ε-contaminated additive model [13] given by

$$x_{\epsilon i} = x_i + z_i y_i. \tag{21}$$

The contamination rate is determined by the zero-one process z_i, defined by $P(z_i = 1) = \varepsilon$, y_i is a white contamination process (independent of x_i) representing the outlier, and $P(y_i = Y_k) = p_k$ with $k = 1, \ldots, K$. It is noted that this model, which embraces also the continuous case as $K \to \infty$, is assumed only for concreteness and for providing insights later on. In essence, we are modeling a contamination characterized with a heavy-tailed distributions, which are those most susceptible to have a huge impact on the sample covariance.

The sample variance is biased as follows (see Appendix VII-4 for details)

$$E\left[\hat{\Sigma}_S\right] = E\left[\left|x_{\varepsilon i} - x_{\varepsilon j}\right|^2\right]/2 = \Sigma + \varepsilon\left(\sigma_y^2 + \mu_y^2\left(1-\varepsilon\right)\right) \tag{22}$$

where μ_y and σ_y^2 are the mean and variance of y_i, respectively. The variance is therefore overestimated with an additive bias term proportional to the mean, variance and rate of the contamination process.

However, the p.d.f. of the contaminated data can be written as a weighted sum of shifted replicas of the original one:

$$f_\varepsilon(x) = (1-\varepsilon)f(x) + \varepsilon\sum_{k=1}^{K} p_k f(x - Y_k). \tag{23}$$

[1]It is worth noting that, in the real data case, w_{min} turns out to be inversely proportional to N^2 (the proof is omitted), thus improving the efficiency/robustness trade-off and giving more tolerance in fixing w.

$\hat{\Sigma}[0] = \hat{\Sigma}_S$; $\Delta = 1$; $q = 1$; $0 < \delta \ll 1$

while $\Delta > \delta$

$W[q] = 15\hat{\Sigma}[q]/N$

$\hat{U}[q] = \frac{2}{N(N-1)}\sum_{1\le i<j\le N} k_{W[q]}\left(x_i - x_j\right)$

$\hat{\Sigma}[q] = \left(\frac{1}{\hat{U}[q]} - 1\right)\frac{W[q]}{2}$

$\Delta = \left|\hat{\Sigma}[q] - \hat{\Sigma}[q-1]\right|/\hat{\Sigma}[q]$

$q \leftarrow q + 1$

end

Fig. 1. Iterative procedure for determining the kernel bandwidth, W.

Using (1), we obtain the following lower and upper bound for the IP (see Appendix VII-5):

$$V\,\mathrm{Col}\left(zy\right) \le V_\varepsilon \le V \tag{24}$$

where

$$V_\varepsilon = \int f_\varepsilon^2(x)dx \tag{25}$$

is the IP of the contaminated data and

$$\mathrm{Col}\left(zy\right) = (1-\varepsilon)^2 + \varepsilon^2\sum_{k=1}^{K} p_k^2 \le 1 \tag{26}$$

is the collision probability [14] of the additive contamination. The IP is therefore underestimated, which leads to overestimating as well the resulting variance. However, the impact of the contamination is now much smaller. Note that the IP of the contaminated data is lower-bounded in a multiplicative manner by the collision probability $\mathrm{Col}\left(zy\right)$ (see the left hand inequality in (24)). This probability depends solely on the contamination rate ε and on the probabilities p_k associated to the additive outlier values. Remarkably, the values Y_k of the contamination process have now *no* impact on the IP. The main advantage is then that the impact on the estimation is governed solely by the collision probability of the outliers values, and *not* by how large the outliers values are. This proves why small kernel bandwidths are interesting to achieve robustness.

D. Kernel bandwidth determination

Kernel bandwidth W operates as an scale parameter that depends on the data dynamic range. As the variance is precisely the parameter we want to estimate, the possibility of using an iterative method to estimate the bandwidth from the data arises naturally, as summarized in Fig. 1. The sample variance is first estimated, which is known to be optimal in nominal conditions but inflated in the presence of contamination. This value is used to fix the bandwidth W to a conservative value as a function of the available number of samples. Using this value, we estimate the entropy-based variance which is used to fix the relative kernel bandwidth for the next iteration, and this procedure is repeated until no significant relative change (δ) of the estimated variance value is observed.

V. Numerical results

Fig. 2 shows the normalized variance of $\hat{U}$ in nominal conditions a function of w for increasing values of N, analytical

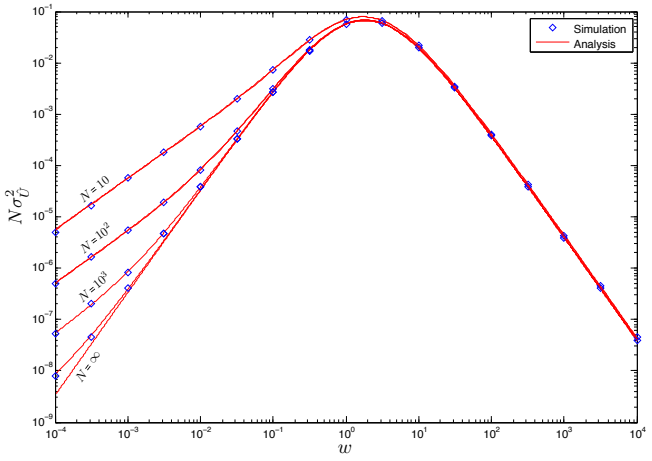

Fig. 2. Normalized variance of $\hat{U}$ as a function of the relative kernel bandwidth w for different values of the data size, N.

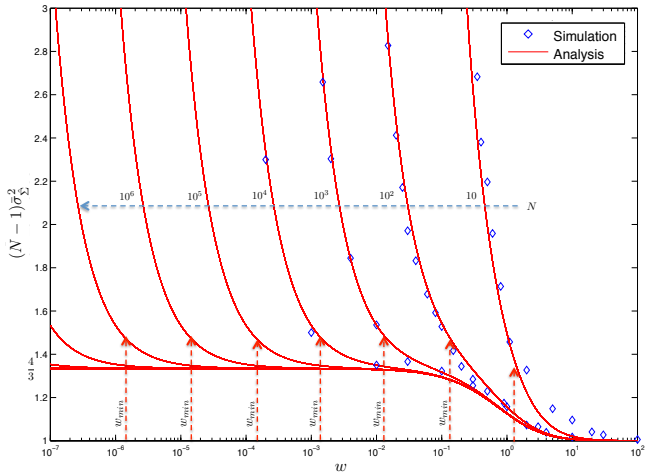

Fig. 3. Variance amplification with respect to the CRB as a function of the relative kernel bandwidth w for different values of the data size, N.

(Eqs. (7), (15) and (16)) and numerical. The influence of b in (7) is manifested as a variance penalty with respect to the asymptotic case ($N \to \infty$), and this occurs for moderate and small w.

Fig. 3 shows the variance of the estimator in nominal conditions as a function of w for increasing values of N, analytical (Eqs. (14), (15) and (16)) and numerical. According to (17), as N is increased, the maximum penalty with respect to the CRB tends to $4/3$ when using moderately small relative kernel bandwidths. The larger is N, the larger is the margin for the use of small kernel bandwidths, which are those interesting for achieving robustness without trading-off too much the estimator accuracy in nominal conditions. Moreover, the existence of this margin for large N is what provides insensitivity to the used outlier model assumptions. Therefore, the larger is N, the less critical is to fix the adequate kernel bandwidth in order to achieve sufficiently high accuracy in nominal conditions. The threshold phenomenon is confirmed (indicated by dashed vertical arrows). For example, for $N = 10^5$, w can be as small as $w_{min} = 1.5 \times 10^{-4}$ for the purpose of improving robustness, without scarifying significantly the performance in nominal conditions.

Fig. 4 shows the robustness of the proposed entropy-based variance estimator in the presence of outliers. The relative bias of the variance estimate is shown as a function of the variance of a zero-mean binary outlier process, for two different values of the contamination rate, ε. At each point, the average number of iterations required by the algorithm is shown. While for small contamination a pair of iterations roughly suffices, more iterations are needed in the case of large-valued outliers and contaminations, specially for moderate N where the determination of the kernel bandwidth at every iteration becomes more critical. While the sample variance exhibits a non-robust behavior with an unbounded relative bias as the outlier variance increases, the proposed estimator exhibits a floor. The asymptotic value obtained from an analytic computation of the IP is indicated as dashed (red) horizontal lines, whose floor for large outliers is given by $1/v_\varepsilon$ (see (26)).

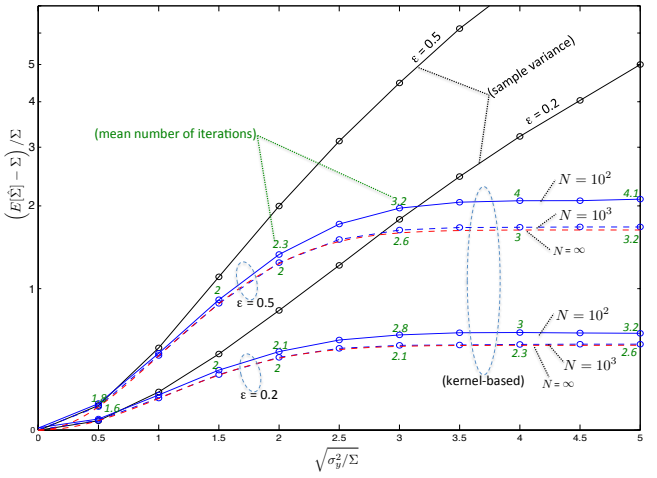

Fig. 4. Relative bias of the variance estimators vs. relative outlier variance, for different contamination rates, ε, and data sizes, N.

VI. Conclusions

We considered the problem of estimating the determinant of the covariance matrix without explicitly estimating that matrix. We have shown that by first estimating an entropy-based measure and applying the Gaussian assumption in a second stage, we get robustness to the overall estimate in the presence of outliers typically associated to man-made noise in communications systems. The resulting estimator is affected by the collision probability of the contamination (insensitive to their values), instead of its variance. A procedure for estimating the kernel bandwidth from the data has been provided, understanding its interplay with the data size and the department from optimality. An open research line is the extension to robust generalized coherence estimation.

VII. Appendices

1) Proof of (4) & (5):

$$E\left[k_{\mathbf{W}}(\mathbf{u})\right] = \pi^{-M}\left|\mathbf{C}\right|^{-1}\int e^{-\mathbf{u}^H\mathbf{W}^{-1}\mathbf{u}}e^{-\mathbf{u}^H\mathbf{C}^{-1}\mathbf{u}}d\mathbf{u}$$

$$= \pi^{-M} |\mathbf{C}|^{-1} \int e^{-\mathbf{u}^H(\mathbf{W}^{-1}+\mathbf{C}^{-1})\mathbf{u}} d\mathbf{u}$$

$$= |\mathbf{C}|^{-1} \left|(\mathbf{W}^{-1}+\mathbf{C}^{-1})\right|^{-1} = |\mathbf{W}|\,|\mathbf{W}+\mathbf{C}|^{-1}$$

Defining $\mathbf{z} = [\mathbf{u}^T, \mathbf{v}^T]^T$, we have $E\left[k_{\mathbf{W}}(\mathbf{u})k_{\mathbf{W}}(\mathbf{v})\right] = E\left[k_{\tilde{\mathbf{W}}}(\mathbf{z})\right]$, with $\tilde{\mathbf{W}} = \mathbf{I}_M \otimes \mathbf{W}$, which yields (5).

2) a and b in (7): Defining $\mathbf{d}_{i,j} = \mathbf{x}_i - \mathbf{x}_j$, the variance of $\hat{U}$ in (3) can be expressed as:

$$\sigma^2_{\hat{U}} = E\left[\hat{U}^2\right] - \bar{U}^2 = (N(N-1)/2)^{-2} \times$$

$$\sum_{i,j,i',j'} \left(E\left[k_{\mathbf{W}}\left(\mathbf{d}_{i,j}\right) k_{\mathbf{W}}\left(\mathbf{d}_{i',j'}\right)\right] - \left(E\left[k_{\mathbf{W}}\left(\mathbf{d}_{i,j}\right)\right]\right)^2 \right)$$

with $i < j$ and $i' < j'$. In the summation we have:

- $N(N-1)/2$ terms with $i = i'$ and $j = j'$, all equal to $b = E\left[k^2_{\mathbf{W}}(\mathbf{u})\right] - \left(E\left[k_{\mathbf{W}}(\mathbf{u})\right]\right)^2$ with $\mathbf{C_u} = 2\mathbf{\Sigma}$. For $E\left[k^2_{\mathbf{W}}(\mathbf{u})\right]$, use (5) with $\gamma = 1$, which yields the first term of (9). For the term $E\left[k_{\mathbf{W}}(\mathbf{u})\right]$ we use use (4) to obtain the second term of (9);
- $(N(N-1)/2)(N-2)$ terms with $i = i'$ and $j \neq j'$;
- $(N(N-1)/2)(N-2)$ terms with $j = j'$ and $i \neq i'$.

Therefore, we have $N(N-1)(N-2)$ terms all equal to $a = E\left[k_{\mathbf{W}}(\mathbf{u})\, k_{\mathbf{W}}(\mathbf{v})\right] - \left(E\left[k_{\mathbf{W}}(\mathbf{u})\right]\right)^2$ with $\mathbf{C_u} = \mathbf{C_v} = 2\mathbf{\Sigma}$ and $\mathbf{C_{uv}} = \mathbf{\Sigma}$. For $E\left[k_{\mathbf{W}}(\mathbf{u})\, k_{\mathbf{W}}(\mathbf{v})\right]$ use (5) with $\gamma = 1/2$, which yields the first term of (8). The second term has been proved before. The remaining terms are such that $i \neq i'$ and $j \neq j'$ and there are zero as $k_{\mathbf{W}}(\mathbf{u})$ and $k_{\mathbf{W}}(\mathbf{v})$ are independent.

3) Small perturbation analysis for $\sigma^2_{\hat{\Sigma}}$: Defining $dU = \hat{U} - \bar{U}$ (with $|dU| \ll \bar{U}$) and assuming normality,

$$\hat{\Sigma} = \left(\left(\bar{U} + dU\right)^{-1} - 1\right) W/2$$

$$\hat{\Sigma} \approx \left(\left(\bar{U}^{-1} - 1\right) - \bar{U}^{-2} dU + \bar{U}^{-3} dU^2\right) W/2 \approx \bar{\Sigma} + d\Sigma$$

$$\sigma^2_{\hat{\Sigma}} \approx \left(\bar{U}^{-4}\sigma^2_{\hat{U}} + 3\bar{U}^{-6}\sigma^4_{\hat{U}}\right) W^2/4$$

4) Impact of contamination to sample variance:

$$\hat{\Sigma}_S = \frac{2}{N(N-1)} \sum_{1 \leq i < j \leq N} |x_{\varepsilon i} - x_{\varepsilon j}|^2 /2$$

$$E\left[\hat{\Sigma}_S\right] = E\left[|x_{\varepsilon i} - x_{\varepsilon j}|^2\right]/2$$

$$E\left[\hat{\Sigma}_S\right] = \left(E\left[|x_i - x_j|^2\right] + E\left[|z_i y_i - z_j y_j|^2\right]\right)/2$$

$$= \Sigma + E\left[|z_i|^2\right] E\left[|y_i|^2\right] - |E\left[z_i\right]|^2 |E\left[y_i\right]|^2$$

$$= \Sigma + \varepsilon\left(\sigma^2_y + \mu^2_y\right) - \varepsilon^2 \mu^2_y = \Sigma + \varepsilon\left(\sigma^2_y + \mu^2_y\left(1 - \varepsilon\right)\right)$$

5) Impact of contamination to IP: Let us define

$$f_\varepsilon(x) = \sum_{k=0}^{K} \tilde{p}_k f(x - Y_k)$$

with $\tilde{p}_k = 1 - \varepsilon$ for $k = 0$ and $\tilde{p}_k = \varepsilon p_k$ for $1 \leq k \leq K$. For the upper-bound:

$$f_\varepsilon(x) = \sum_{k=0}^{K} \left(\sqrt{\tilde{p}_k}\right)\left(\sqrt{\tilde{p}_k} f(x - Y_k)\right)$$

$$f_\varepsilon(x) \leq \sqrt{\sum_{k=0}^{K} \left(\sqrt{\tilde{p}_k}\right)^2 \sum_{k=0}^{K} \left(\sqrt{\tilde{p}_k} f(x - Y_k)\right)^2}$$

$$= \sqrt{\sum_{k=0}^{K} \tilde{p}_k f^2(x - Y_k)}$$

$$V_\varepsilon \leq \int \left(\sum_{k=0}^{K} \tilde{p}_k f^2(x - Y_k)\right) dx$$

$$= \sum_{k=0}^{K} \tilde{p}_k \left(\int f^2(x - Y_k) dx\right) = \left(\sum_{k=0}^{K} \tilde{p}_k\right) V = V$$

For the lower-bound:

$$V_\varepsilon = \sum_{k=0}^{K} \sum_{k'=0}^{K} \tilde{p}_k \tilde{p}_{k'} g(Y_{k'} - Y_k)$$

$$0 \leq g(z) = \int f(x+\tau) f(x) dx \leq g(0) = V$$

$$V_\varepsilon \geq \sum_{k=0}^{K} \tilde{p}^2_k g(0) = \left(\sum_{k=0}^{K} \tilde{p}^2_k\right) V$$

REFERENCES

[1] A. M. Zoubir, V. Koivunen, Y. Chakhchoukh, and M. Muma, "Robust estimation in signal processing: A tutorial-style treatment of fundamental concepts," *IEEE Signal Process. Magazine*, vol. 29, no. 4, pp. 61–80, 2012.

[2] A. Margoosian, J. Abouei, and K. N. Plataniotis, "An accurate kernelized energy detection in Gaussian and non-gaussian/impulsive noises," *IEEE Trans. Signal Process*, vol. 63, no. 1, pp. 5621–5636, 2015.

[3] Y. Sun, P. Babu, and D. P. Palomar, "Regularized robust estimation of mean and covariance matrix under heavy-tailed distributions," *IEEE Trans. Signal Process*, vol. 63, no. 12, pp. 3096–3109, June 2015.

[4] D. E. Tyler, "A distribution-free M-estimator of multivariate scatter," *Ann. Statistic.*, vol. 15, no. 1, pp. 234–251, 3 1987.

[5] J. T. Kent and D. E. Tyler, "Maximum likelihood estimation for the wrapped Cauchy distribution," *J. Appl. Statist.*, vol. 15, no. 2, pp. 247–254, 1988.

[6] D. Ramírez, J. Vía, I. Santamaría, and L. Scharf, "Detection of spatially correlated Gaussian time series," *IEEE Trans. Signal Process*, vol. 58, no. 10, pp. 5006–5015, 2010.

[7] D. Cochran, H. Gish, and D. Sinno, "A geometric approach to multiple-channel signal detection," *IEEE Trans. Signal Process*, vol. 43, no. 9, pp. 2049–2057, 1995.

[8] Q. Wang, S. R. Kulkarni, and S. Verdú, *Universal estimation of Information measures for analog sources*. Foundations and trends in Communications and Information Theory, 2009, no. 5:3.

[9] J. C. Principe, *Information Theoretic Learning: Rényi's Entropy and Kernel Perspectives*. NewYork: Springer, 2010.

[10] A. W. van der Vaart, *Asymptotic Statistics*. Cambridge University Press, Cambridge., 1998, vol. 3.

[11] R. J. Serfling, *Approximation Theorems of Mathematical Statistics*. New York, NY, USA: Wiley, 1980, vol. (Chapter 5).

[12] G. H. Golub and C. F. V. Loan, *Matrix Computations*. Johns Hopkins University Press, 1996.

[13] R. Maronna, R. Martin, and V. Yohai, *Robust Statistics: Theory and Methods*. New York, NY, USA: Wiley, 2006.

[14] S. Fehr and S. Berens, "On the conditional Rényi entropy," *IEEE Trans. Inf. Theory*, vol. 60, no. 11, p. 6801, 6810 2014.