Minimum Support ICA Using Order Statistics. Part I: Quasi-range Based Support Estimation*

Frédéric Vrins and Michel Verleysen**

Université catholique de Louvain, Machine Learning Group,
Place du Levant, 3, 1380 Louvain-la-Neuve, Belgium
{vrins, verleysen}@dice.ucl.ac.be

Abstract. The minimum support ICA algorithms currently use the extreme statistics difference (also called the *statistical range*) for support width estimation. In this paper, we extend this method by analyzing the use of (possibly averaged) differences between the $N-m+1$-th and m-th order statistics, where N is the sample size and m is a positive integer lower than $N/2$. Numerical results illustrate the expectation and variance of the estimators for various densities and sample sizes; theoretical results are provided for uniform densities. The estimators are analyzed from the specific viewpoint of ICA, i.e. considering that the support widths and the pdf shapes vary with demixing matrix updates.

1 Introduction

Recently, new contrasts for ICA have been developed for the separation of bounded sources, based on the fact that the output support width varies with the mixing coefficients; the independent components are recovered one by one, by finding directions in which the outputs have a minimum support convex hull measure [1, 3]. Such approach benefits from some advantages: on one hand the contrast is extremely simple and free of spurious maxima [1]; and on the other hand, its optimization can be easily handled, leading to interesting results in terms of speed and residual crosstalk.

The support estimation of a pdf f_X has been extensively studied in statistics and econometrics. Nevertheless, most methods require resampling techniques or tricky tuning parameters, and are thus not really appropriated to ICA algorithms. For instance, if the support $\Omega(X)$ of the output is (a, b), existing ICA methods currently use the *range* approximation to estimate the (Lebesgue) measure of the support $\mu[\Omega(X)] : b - a \simeq R(X) \triangleq \max_{i,j}[x_i - x_j]$, $1 \leq i, j \leq N$ where the x_j can either be considered as iid realizations of the common random variable (r.v.) X, or as a samples of a stationary stochastic process constituted of a sequence of N independent r.v. X_j, all sharing the same pdf f_X.

An extended form of this estimator will be considered here, using order statistics differences. The study is motivated by the idea that the extreme statistics

* The authors are grateful to D. Erdogmus for having inspirited this work by fruitful discussion on expectation and variance of cdf differences and m-spacings.

** Michel Verleysen is Research Director of the Belgian F.N.R.S.

J. Rosca et al. (Eds.): ICA 2006, LNCS 3889, pp. 262–269, 2006.

are not necessarily reliable. Let $x_{(1)} \leq \cdots \leq x_{(N)}$ be a rearranged version of the observed sample set $\mathcal{X}_N = \{x_1, ..., x_N\}$; each of the $x_{(j)}$ can be seen as a realization of a r.v. $X_{(j)}$. Obviously, the $X_{(j)}$ are not independent and do not share the same pdf. Both $x_{(j)}$ and $X_{(j)}$ are called the j-th order statistic of $\mathcal{X}_N$. This appellation is not related to the *(higher) order statistics*, frequently used in the ICA community. The *order statistics*, as defined in this paper, have already been used in the BSS context in [4] (see also [10], [9] and references therein). These ordered variates can be used to define the range $R_1(X) = X_{(N)} - X_{(1)}$, or the (symmetric) *quasi-ranges* (QR): $R_m(X) = X_{(N-m+1)} - X_{(m)}$, with $m < \lfloor N/2 \rfloor$. Such QR could be also used to estimate the quantity $b - a$. However, even if $R_m(X)$ is a generalization of $R(X) = R_1(X)$, both estimators only involve two sample points. In order to include more points in the estimation, we also compare $R_m(X)$ to $\langle R_m(X) \rangle \triangleq 1/m \sum_{i=1}^{m} R_i(X)$.

The QR-based support estimation is analyzed in Section 2, for various pdf and sample sizes. Specific phenomena are discussed in Section 3, keeping in mind that the pdf of X vary with time in ICA applications, due to the iterative updates of the demixing matrix. Note that the performance analysis of ICA algorithms using the above estimators is discussed in a separated paper [2].

2 Density, Bias and Variances of the QR

A large attention has been paid to order statistics and QRs in the statistic literature. For instance, the pdf $f_{R_m(X)}$ of $R_m(X)$ for $\Omega(X) = (a, b)$ has been established in [8]. If F_X denotes the cdf of X, the computation of $f_{X_{(j)}}$ yields

$$f_{R_m(X)}(r) = \frac{N!}{((m-1)!)^2(N-2m)!} \int_{-\infty}^{\infty} F_X^{m-1}(x) \left[F_X(x+r) - F_X(x)\right]^{N-2m} \times f_X(x) f_X(x+r) \left[1 - F_X(x+r)\right]^{m-1} dx \ .$$

It can be seen that the density $f_{R_m(X)}$ is a function of $f_X = F'_X$, as well as of N and m. Although, the above theoretical expression is of few use in practice; for most parent densities f_X, no analytical expression can be found for simple functions of $R_m(X)$, such as expectation and moments. Dealing with $f_{\langle R_m(X) \rangle}(r)$ is even worst, since $f_{\langle R_m(X) \rangle}$ depends on the joint density of $R_1(X), \cdots, R_m(X)$. A more reasonable way to compute the expectation and variances of $R_m(X)$ and $\langle R_m(X) \rangle$ is to prefer numerical simulations to theoretical developments that are valid for a single density only; this is done in Section 2.1. However, for comparison purposes, the exact expressions of $\mathrm{E}[R_m(X)]$, $\mathrm{VAR}[R_m(X)]$, $\mathrm{E}[\langle R_m(X) \rangle]$ and $\mathrm{VAR}[\langle R_m(X) \rangle]$ aer given in Section 2.2 in the case where f_X is the uniform pdf.

2.1 Empirical Expectation and Variance of QRs

Let us note U, L, T and V white r.v. having uniform, linear, triangular and 'V'-shape densities, respectively. We note the empirical expectations and variances

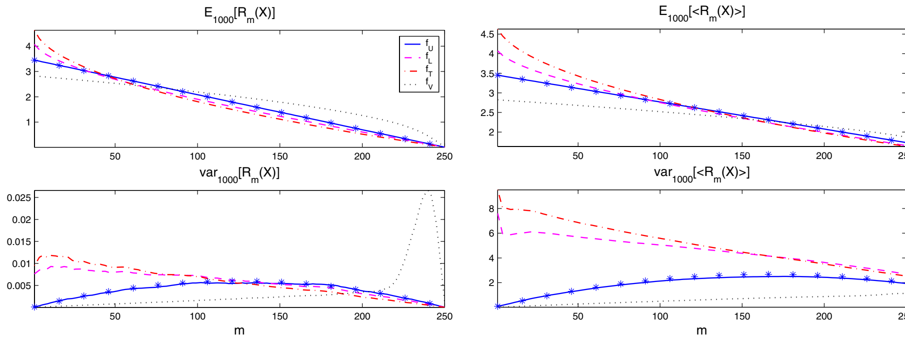

Fig. 1. Empirical expectations and variances of $R_m(X)$ (left) and $\langle R_m(X)\rangle$ (right) for $N = 500$ (1000 trials). The theoretical curves for the uniform case are labelled ' $*$ '.

of estimators taken over t trials as $\mathrm{E}_t[\cdot]$ and $\mathrm{VAR}_t[\cdot]$. The evolution of these quantities with respect to m is shown in Fig. 1. Three particular effects have to be emphasized.

- *Effect of m and f_X*: the estimation error increases with m for fixed N at a rate depending on f_X, and m has thus to be chosen small enough in comparison to N (the true support measures of the white r.v. are $\mu[\Omega(T)] = 2\sqrt{6} > \mu[\Omega(L)] = 3/2\sqrt{8} > \mu[\Omega(U)] = 2\sqrt{3} > \mu[\Omega(V)] = 2\sqrt{2}$). The support of V and U can be estimated with a low variance, even for a small m, contrarily to T and L. For instance, the variance of the estimators decreases with m for linear and triangular r.v., while this behavior cannot be observed for U or V; the variance of the estimators increases when unreliable points (i.e. corresponding to a low value of the pdf) are taken into account. The shape of $\mathrm{VAR}_t[R_m(U)]$ and $\mathrm{VAR}_t[\langle R_m(U)\rangle]$ are more surprising, but they have been confirmed by the analytical equations given in Section 2.2.
- *Effect of N*: it can be reasonably understood, though not visible on Fig. 1, that $R_m(X)$ and $\langle R_m(X)\rangle$ are asymptotically unbiaised estimators of $b - a$ if b and a are not isolated points, that is if the support $\Omega(X)$ includes some neighborhoods of b and a. Similarly, $\lim_{N\to\infty} \mathrm{VAR}[R_m(X)] = 0$ (for m fixed); We conjecture that the latter limit holds for $\langle R_m(X)\rangle$, with fixed m. Note that the convergence rate depends of f_X. These properties can be easily confirmed when X is uniform (see next section).
- *$R_m(X)$ vs $\langle R_m(X)\rangle$*: the error of $R_m(X)$ increases at a higher rate than the one of $\langle R_m(X)\rangle$ for increasing m and fixed N; this is a consequence of the regularization due to the average in $\langle R_m(X)\rangle$: $\Pr[\langle R_m(X)\rangle \geq R_m(X)] = 1$.

The above simulation results indicate that $\langle R_m(X)\rangle$ should be preferred to $R_m(X)$ for support estimation; for a small m compared to N, both the error and the variance are improved. The choice of m is difficult, though : it must be small enough to ensure a small error, but not too small if one desires to estimate the support of e.g. f_T or f_L or of noisy data; an optimal value of m depends of the unknown density.

2.2 Exact Expectation and Variance of QRs for Uniform Pdf

In this section, contrarily to the remaining of the paper, U denotes a normalized uniform r.v. with support equal to $(0, 1)$, in order to simplify the mathematical developments, that are sketched in the Appendix.

Using the expression of $f_{R_m(X)}$ given in Section 2, it can be shown that $\mathrm{E}[R_m(U)] = (N-2m+1)/(N+1)$ and $\mathrm{VAR}[R_m(U)] = 2m(N-2m+1)/((N+2)(N+1)^2)$. Simple manipulations directly show that $R_m(U)$ is an asymptotically unbiaised estimator of the support measure, monotonously increasing with limit $b-a$, when m is kept fixed.

The expectation of $\langle R_m(U)\rangle$ can directly be derived from $\mathrm{E}[R_m(U)]$; if m is fixed, $\langle R_m(U)\rangle$ is asymptotically unbiased. On the contrary if we set $m = \lfloor N/k \rfloor$, $k \in \mathbb{Z}$, the asymptotic bias equals $1/k$. Such bias can be cancelled if $m(N)$ increases at a lower rate that N; this is e.g. the case if $m(N) = \lfloor \sqrt{N/k} \rfloor$. Regarding the variance of $\langle R_m(U)\rangle$, we have

$$\mathrm{VAR}[\langle R_m(U)\rangle] = \frac{-3m^3 + 2m^2(N-2) + 3mN + (N+1)}{3m(N+2)(N+1)^2} \ .$$

Using ad-hoc scaling coefficients, the related quantities can be obtained for *white* r.v. (no more confined in $(0, 1)$). The theoretical curves (labelled using '$*$') are superimposed to the associated empirical ones if Fig. 1.

3 Estimating the Mixture Support

The above discussion gives general results regarding the estimation of the support convex hull measure of a one-dimensional r.v. Let us now focus on the support estimation of a single output (deflation scheme) of the 2-dimensional ICA application; the support varies with the associated row of transfer matrix. For the ease of understanding, we constrain the sources to share the same pdf. The instantaneous noise-free ICA mixture scheme, under whiteness constraint, leads to the following expression for an output:

$$Z_X(\phi + \varphi) = \cos(\phi + \varphi) S_1 + \sin(\phi + \varphi) S_2 \ , \tag{1}$$

where the S_i are the sources, and ϕ and φ are resp. the mixing-whitening and demixing angles. The subscript X means that the sources follow the pdf f_X. We define $\theta = \phi + \varphi$ as the input-output transfer angle.

The minimum support ICA approach updates the angle φ to minimize the objective function $\mu[\Omega(Z_X(\theta))]$. Since it has been shown that this cost function is concave in a given quadrant, a gradient-based method leads to $\theta = k\pi/2$, with $k \in \mathbb{Z}$. In practice however, $\mu[\Omega(Z_X(\theta))]$ has to be estimated, for example using the proposed form of estimators. The following subsections points out two phenomena that have to be considered.

3.1 The Mixing Effect

Fig. 2 shows the surface of the empirical expectation of the error ϵ, defined as

$$\epsilon(X, N, m, \theta) = \mu[\Omega(Z_X(\theta))] - \langle R_m(Z_X(\theta)) \rangle \ , \tag{2}$$

when f_X is a triangular or 'V'-shape density, θ ranges from 0 to $\pi/2$, and N from 2 to 500.

In addition to the bias dependency on f_X, we can observe what we call the 'mixing effect': the error increases for θ going from 0 or $\pi/2$ to $\pi/4$. This phenomenon can be explained as follows. The pdf of a sum of r.v. is obtained by convoluting the pdfs of the summed r.v. Therefore, the tails of resulting pdfs will be less sharp than the source pdfs. For instance, the pdf of a sum of two normalized uniform r.v. is triangular. The mixing effect phenomenon can now be understood, since for fixed N and m, the support measure of a pdf with sharp tails is better estimated than of a pdf with smoothly decreasing tails. The main consequence of this phenomenon is that the empirical contrast is more 'flat' than the true one with respect to the transfer angle.

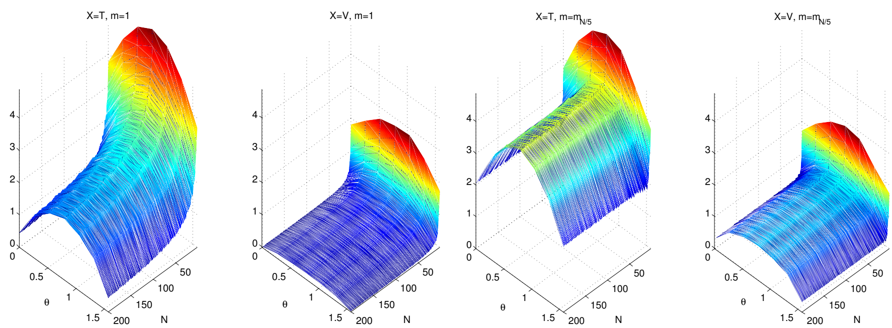

Fig. 2. Empirical error $\mathrm{E}_{100}\left[\epsilon(X, N, m, \theta)\right]$ for various source pdf f_X, with N ranging from 2 to 200, and $m_{N/5} \triangleq \max(1, \lfloor N/5 \rfloor)$

3.2 The Large-m Given N Effect

The mixing effect emphasizes that the support estimation quality depends of θ: the support of $Z_X(\pi/4)$ is always more under-estimated than the one of $Z_X(k\pi/2)$ by using QR or averaged QR estimators. This results from the fact that the output density depends of the transfer angle. In section 2.1, the effect of m on the expectation and variance of the estimators is shown to depend of the density f_X. In the ICA application it thus depends of θ: the bias increases with m, at a rate depending of $f_{Z_X(\theta)}$, i.e. of θ. This is a tricky point, since even if $\mu[\Omega(Z_X(\pi/4))] > \mu[\Omega(Z_X(k\pi/2))]$, this inequality evaluated using the support measure approximations can be violated. In this scenario, occurring for m greater than a threshold value $m^\dagger$, the contrast optimum will be obtained for

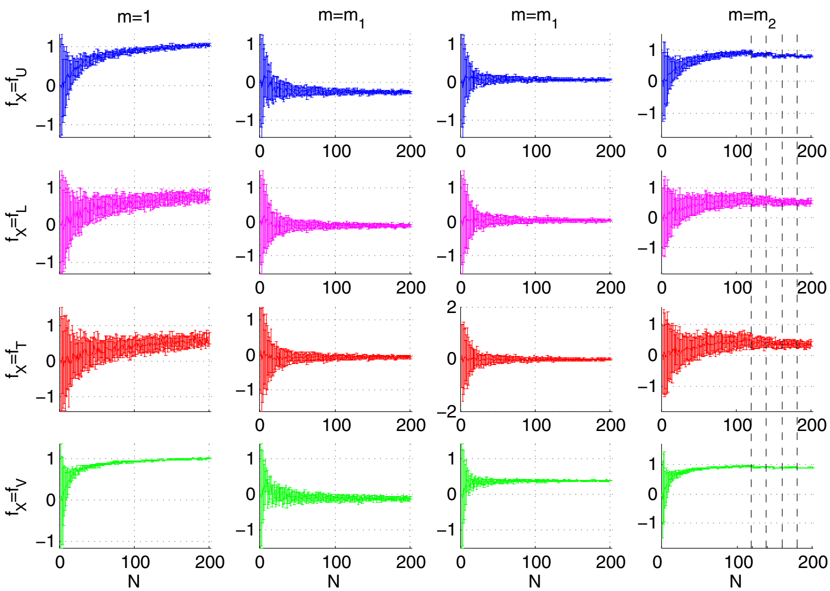

Fig. 3. Evolution of error means for various pdf f_X, N ranging from 2 to 200 and $m_1 = \max(1, \lfloor N/5 \rfloor)$,$m_2$ is given by eq. (3). $\mathrm{E}_{100}\left[\mathcal{E}_1(X)\right] = \mathrm{E}_{100}\left[\langle\mathcal{E}_1(X)\rangle\right]$ (first col.); $\mathrm{E}_{100}\left[\mathcal{E}_{m_1}(X)\right]$ (second col.); $\mathrm{E}_{100}\left[\langle\mathcal{E}_{m_1}(X)\rangle\right]$ (third col.); and $\mathrm{E}_{100}\left[\langle\mathcal{E}_{m_2}(X)\rangle\right]$ (last col.). The width of the curves reflects twice the empirical variance of $\mathcal{E}(X)$ and $\langle\mathcal{E}(X)\rangle$.

$\theta = \pi/4$ rather than $\theta \in \{0, \pi/2\}$, i.e. the algorithm will totally fail. For example, when dealing with $\langle R_m(Z_U(\theta))\rangle$ and two 500-sampled sources, $m^\dagger \simeq 100$. If $R_m(Z_U(\theta))$ is considered, $m^\dagger \simeq 40$. Indeed, the pdf of $Z_U(\pi/4)$ is triangular, and we see on Fig. 1 that the estimation of the support of a white triangular r.v. is lower than the estimated support of a white uniform r.v. for these values of N and $m > m^\dagger$. These values of $m^\dagger$ obviously decrease with decreasing N.

Fig. 3 illustrates the quantities $\mathcal{E}_m(X) \triangleq R_m(Z_X(\pi/4)) - R_m(Z_X(0))$ and $\langle\mathcal{E}_m(X)\rangle \triangleq \langle R_m(Z_X(\pi/4))\rangle - \langle R_m(Z_X(0))\rangle$: negative values of $\mathcal{E}_m(X)$ and $\langle\mathcal{E}_m(X)\rangle$ obtained for $m > m^\dagger$ clearly indicate that the optima of the empirical contrast, i.e. the corresponding estimators will lead to wrong source separation solutions. The last column shows the result obtained by using (3); the vertical dashed lines indicated that m has been incremented. This comment suggests to pay attention to the choice of m: it must be small enough by comparison to N to ensure a small error and $m < m^\dagger$, but greater than one for regularization purposes. Therefore, if $\overline{\alpha}$ denotes the nearest integer to α, we suggest to take m according to the rule

$$\max\left(1, \Re\left(\overline{\left[\left(\frac{N-18}{6.5}\right)^{0.65} - 4.5\right]}\right)\right) . \tag{3}$$

Though the above rule of the thumb will not be detailed here, we mention that rule (3) results from a distribution-free procedure for choosing m and valid for all θ and all source pdfs; the method is detailed in [2]. A positive point is that the critical value $m^\dagger$ does not seem to be sensitive to the number of sources, probably due to the compensation of two effects: even if the tails of $f_{Z_X(\theta)}$ tend to decrease exponentially when many sources are mixed since $f_{Z_X(\theta)}$ tend to be Gaussian-shape (inducing a large under-estimation of the support), large-value sample points can be observed due to the summed r.v. (so that the estimated mixture support should be larger than the estimated source support).

4 Conclusion and Future Work

In this paper, we have investigated the use of the quasi-ranges $R_m(X)$ and averaged symmetric quasi-ranges $\langle R_m(X)\rangle$ for support width minimization approaches to ICA. Note that the computation of the *true* QR requires the knowledge of the order statistics of X, that are unknown here; in this paper, the i-th order statistic $X_{(i)}$ was approximated by the i-th largest observed value $x_{(i)}$ of X. This work is motivated by the fact that extreme statistics can be unreliable. It is shown that m has to be chosen small in comparison to N, but greater than one to make the variance of the estimators decrease for several kinds of pdf. The main advantage of the averaged QR is that it takes $2m$ points into account. From both the expectation and variance points of view, the averaged QR has better performances than the simple QR. We have shown that an excessive value m with given N could lead the related ICA algorithms to totally fail; from this point of view too, the averaged QR has to be preferred to the QR. Future work should focus on a study involving specific noise, as well as a comparison with existing endpoint estimators.

References

1. Vrins, F., Jutten, C. & Verleysen, M. (2005) SWM : a Class of Convex Contrasts for Source Separation. In proc. ICASSP, *IEEE Int. Conf. on Acoustics, Speech and Sig. Process.*: V.161-V.164, Philadelphia (USA).
2. Vrins, F. and & Verleysen, M. (2006) Minimum Support ICA Using Order Statistics. Part II: Performance Analysis. In Proc. ICA, *Int. Conf. on Ind. Comp. Anal. and Blind Sig. Sep.*, Charleston SC (USA).
3. Cruces, S. & Duran, I. (2004) The Minimum Support Criterion for Blind Source Extraction: a Limiting Case of the Strengthened Young's Inequality. In proc. ICA, *Int. Conf. on Ind. Comp. Anal. and Blind Sig. Sep.*: 57–64, Granada (Spain).
4. Pham, D.-T. (2000) Blind Separation of Instantaneous Mixture of Sources Based on Order Statistics. *IEEE Trans. Sig. Process.*, **48**(2): 363–375.
5. Ebrahimi, N., Soofi, E.S, Zahedi, H. (2004) Information Properties of Order Statistics and Spacings. *IEEE Trans. Info. Theory*, **50**(1): 177–183.
6. Spiegel, M.R. (1974) "Mathematical Handbook of Formulas and Tables." McGraw-Hill, New York.
7. Papadatos, N. (1999) Upper Bound for the Covariance of Extreme Order Statistics From a Sample of Size Three. *Indian J. of Stat.*, Series A, Part 2 **61**: 229–240.
8. David, H.A. (1970). "Order Statistics." Wiley series in probability and mathematical statistics, Wiley, New York.

9. Even, J. (2003) "Contributions à la Séparation de Sources à l'Aide de Staitistiques d'Ordre." PhD Thesis, Univ. J. Fourier, Grenoble (France).
10. Blanco, Y. & Zazo, S. (2004) An Overview of BSS Techniques Based on Order Statistics: Formulation and Implementation Issues. In proc. ICA, *Int. Conf. on Indep. Comp. An. and Blind Sig. Sep.*: 73–80, Granada (Spain).

Appendix. Details of Some Results for Uniform Densities

• Expectation and Variance of $R_m(U)$

It is known e.g. from [5] that the pdf of the i-th order statistic a uniform r.v. is $f_{U_{(i)}}(u) = \frac{N!}{(i-1)!(N-i)!}[F_U(u)]^{i-1}[1 - F_U(u)]^{N-i}f_U(u)$. By using basic properties of expectation and $\Omega(U) = (0,1)$, we have $\mathrm{E}[U_{(i+1)} - U_{(i)}] = \frac{1}{N+1}\frac{N!}{i!(N-i-1)!}\int_0^1 x^i(1-x)^{N-i-1}dx$, where the integral equals $\frac{i!(N-i-1)!}{N!}$ [6]. It comes that $\sum_{i=m}^{N-m}\mathrm{E}[U_{(i+1)} - U_{(i)}] = \frac{N-2m+1}{N+1}$. Similar development as above on the i-th order statistic of a uniform variable on $(0,1)$ leads to $\mathrm{VAR}[U_{(i)}] = \frac{i(N-i+1)}{(N+2)(N+1)^2} = \mathrm{VAR}[U_{(N-i+1)}]$.

Since $\mathrm{CORR}[U_i, U_j]$ $(1 \leq i < j \leq N)$ is know from [7], we obtain

$$\mathrm{COV}[U_{(i+p)}, U_{(i)}] = \frac{i(N+1-i-p)}{(N+2)(N+1)^2} \quad . \tag{4}$$

We find $\mathrm{VAR}[U_{(i+p)} - U_{(i)}] = \mathrm{VAR}[U_{(i+p)}] + \mathrm{VAR}[U_i] - 2\mathrm{COV}[U_{(i+p)}, U_{(i)}]$, which equals $\frac{p(N+1-p)}{(N+2)(N+1)^2}$. The results enounced in Section 2.2 comes when setting $i = m$ and $p = N - 2m + 1$.

• Expectation and Variance of $\langle R_m(U)\rangle$

We obviously have $\mathrm{E}[\langle R_m(X)\rangle] = \frac{1}{m}\sum_{p=1}^{m}\frac{N-2p+1}{N+1} = \frac{N-m}{N+1}$.

The computation of $\mathrm{VAR}[\langle R_m(U)\rangle]$ is more tricky. Observe first that:

$$m^2\mathrm{VAR}[\langle R_m(U)\rangle] = \sum_{p=1}^{m}\mathrm{VAR}[R_p(U)] + 2\sum_{1\leq i<j\leq m}\mathrm{COV}\left[R_i(U), R_j(U)\right] \quad . \tag{5}$$

Using eq. (4), we find: $\mathrm{COV}_{i<j}\left[R_i(U), R_j(U)\right] = 2i\frac{N+1-2j}{(N+2)(N+1)^2}$.

We have, using basic properties:

$$\sum_{p=1}^{m}\mathrm{VAR}[\langle R_p(U)\rangle] = \frac{(N+1)m(m+1) - 2/3m(m+1)(2m+1)}{(m+2)(m+1)^2} \quad , \tag{6}$$

and

$$\sum_{1\leq i<j\leq m}\mathrm{COV}\left[R_i(U), R_j(U)\right]] = \frac{m(m-1)}{6(N+2)(N+1)^2}\left\{-3m^2 + m(2N-3) + 2N\right\}.$$

which leads to the results presented in Section 2.2.