CONCOMITANT OF ORDERED MULTIVARIATE NORMAL DISTRIBUTION WITH APPLICATION TO PARAMETRIC INFERENCE

Eric Chaumette and François Vincent

University of Toulouse/Isae-Supaero, 10 Avenue Edouard Belin, 31400 Toulouse, France.
(eric.chaumette@isae.fr, francois.vincent@isae.fr)

ABSTRACT

In statistics, the concept of a concomitant, also called the induced order statistic, arises when one sorts the members of a random sample according to corresponding values of another random sample. Indeed, multivariate order statistics induced by the ordering of linear combinations of the components arises naturally in many instances. As a contribution, we provide a general second-order statistical prediction of concomitant of order statistics for multivariate normal distribution, generalizing earlier works. We exemplify its usefulness in parametric inference via two examples related to deterministic and Bayesian estimation.

Index Terms— Multivariate normal distribution, Order statistics, Concomitants, Parametric Inference, Mean Square Error

1. INTRODUCTION

The ordered values of a sample of observations are called the order statistics of the sample: if \( \mathbf{\theta}^T = (\theta_1, \ldots, \theta_M) \) is a vector of \( M \) real valued random variables, then \( \theta_{(M)} = (\theta_{(1)}, \ldots, \theta_{(M)})^T \) denotes the vector of order statistics induced by \( \theta \) where \( \theta_{(1)} \leq \theta_{(2)} \leq \ldots \leq \theta_{(M)} \) [1]. Order statistics and extreme values are among the most important functions of a set of random variables in probability and statistics. There is natural interest in studying the highs and lows of a sequence, and the other order statistics help in understanding concentration of probability in a distribution. Order statistics are also useful in statistical inference, where estimates of parameters can be based on some suitable functions of the order statistics vector (robust location estimates, detection of outliers, censored sampling, characterizations and goodness of fit). [1] or be implicitly ordered as in maximum likelihood estimation for parametric inference (see Section 3 and [2]). Since there is no direct extension of order concept to multivariate random variables, the extension of procedure based on order statistics to such situations is inapplicable. However, if we consider a random sample arising from a bivariate distribution \( \{(s_1, \theta_1), \ldots, (s_M, \theta_M)\} \), ordering of the values recorded on the first variable \( s \) generates a set of random variables associated with the corresponding \( \theta \) variate [3]. These random variables obtained due to the ordering of the \( \theta \)'s are known as the concomitants of order statistics \( \mathbf{s}_{(M)} \) and are denoted \( \mathbf{s}_{(M)} = (s_{(1)}, \ldots, s_{(M)}) \). Hence the general concept of a concomitant in statistics, also called the induced order statistic, arising when one sorts the members of a random sample according to corresponding values of another random sample [1]. In that perspective, a generalization of the bivariate case, where the sample \( \{\theta_1, \theta_2, \ldots, \theta_M\} \) consists of \( M \) multivariate random variables, is obtained by resorting to a linear combination of the form \( \mathbf{s}^T = (\theta_1^T \mathbf{a}_1, \ldots, \theta_M^T \mathbf{a}_M) \). Then the ordering of the sample \( \mathbf{s} \), i.e. \( \mathbf{s}_{(M)} \), induces the associate ordering of random vectors \( \theta_{(M)} \), i.e. \( \Theta_{(M)} = \{\theta_{(1)}, \ldots, \theta_{(M)}\} \). Multivariate order statistics induced by the ordering of linear combinations of the components arises naturally in many instances. For example, in the evaluation of the performance of students in a course, the final grade may be a weighted average of the scores in a mid-term test and the final examination. Other interesting examples arise in hydrology while analyzing extreme lake levels [4], in biological selection problem [5], ocean engineering [6], development of structural designs [7]. Therefore, the need to characterize the order statistics and their concomitants has led to a large body of work summarized in [1][8][9]. A fairly general second-order statistical prediction of concomitants of ordered multivariate normal distribution has been given in [4] and [10] for the situation in which the random vectors \( \theta_m \) are independent. Unfortunately the situation where vectors \( \theta_m \) are independent is not the common situation in many instances of the setting under consideration (see Section 3).

Therefore, as a contribution, we provide the most general second-order statistical prediction of order statistics and their concomitants for multivariate normal distribution, whatever they are dependent or independent. These closed forms generalize the earlier work from [4] and [10]. We exemplify their usefulness in parametric inference. Indeed, the asymptotic performance analysis of the mean square error (MSE) of maximum likelihood estimators (MLEs) can be refined by the study of concomitants of ordered estimates (generalizing the single unknown parameter case addressed in [2]). In Kalman filtering for linear discrete state-space models, concomitants of ordered estimates can be used to monitor the range of the states vector.

2. STATISTICAL PREDICTION OF CONCOMITANTS OF ORDERED MULTIVARIATE NORMAL DISTRIBUTION

Let us consider the observation of \( M \) random Gaussian vectors with \( P \) components: \( \{\theta_m\}_{m=1}^M \). The vector gathering the \( PM \) Gaussian random variables is denoted by \( \mathbf{\Theta} = \{\theta_1, \ldots, \theta_M\} \) in \( \mathcal{M}_{RE}(P, M) \), and:

\[ E[\mathbf{g}(\mathbf{y})] = \int g(y) p(y) dy \]
\[ v_\Theta = (\theta_1^T, \ldots, \theta_M^T), \quad v_\Theta \sim N_{PM} (\mu_{v_\Theta}, C_{v_\Theta}), \quad (1) \]
\[
\mu_{v_\Theta} = \begin{pmatrix} \mu_{\theta_1} \\ \vdots \\ \mu_{\theta_M} \end{pmatrix}, \quad C_{v_\Theta} = \begin{bmatrix} C_{\theta_1, \theta_1} & \cdots & C_{\theta_1, \theta_M} \\ \vdots & \ddots & \vdots \\ C_{\theta_M, \theta_1} & \cdots & C_{\theta_M, \theta_M} \end{bmatrix},
\]

As in [4], let us consider the following \( M \)-dimensional vector:
\[
s = (s_1^T, a, s_2^T) = \Theta^T a, \quad a \in M_P (P, 1), \quad (2a)\]
then, the concomitants of \( s_{(M)} = (s_{(1)}, \ldots, s_{(M)}) \) are defined as:
\[
\Theta_{[m]} = [\theta_{[1]}, \ldots, \theta_{[M]}] \quad | \quad (m) = \theta_{[m]} \iff s_{(m)} = s_{m}. \quad (2b)\]

Let \( \{d_1, \ldots, d_M\} \) be the \( M \)-dimensional unit basis vectors; then:
\[
\theta_{[m]} = \text{vec} (\Theta_{[d_m]}), S_m = d_m^T \otimes I_P, \quad (3a)\]
\[
E [\theta_{[m]}] = S_m E [v_\Theta], C_{\theta_{[m]}, \theta_{[m]}} = S_m C_{v_\Theta} S_m^T. \quad (3b)\]

In other words, the first and second order statistical prediction of \( \theta_{[\cdot]} \), \( \theta_{[M]} \) \( \Theta \), derive from the first and second order statistical prediction of \( v_\Theta \), which is introduced in this section. First, note that \( s_{(M)} \in \text{Per}(s) \), where \( \text{Per}(s) = \{ s_i = P_i s, i = 1, \ldots, M \} \) is the collection of random vectors \( s_i \), corresponding to the \( M \) different permutations of the components of \( s \). Here \( P_i \in M_{M \times M} \) are permutation matrices with \( P_i \neq P_j \) for all \( i \neq j \). Let \( \Delta \in M_{(M-1) \times M} \) be the difference matrix such that \( (s_{(s_1)}, s_{(s_2)}, \ldots, s_{(M-1)})^T, \text{i.e.}, the nth row of } \Delta = d_{m+1}^T - d_m^T, m = 1, \ldots, M-1. \)

Let \( S_i = \{ s : \Delta s_i \geq 0 \} \) where \( s_i \sim N_M (\mu_s, C_s) \), \( \mu_s = P_i \mu_s, C_s = P_i C_s P_i^T \). As the set of events \( \{S_i\}_{i=1}^M \) is a partition of \( \mathbb{R}^M \), whatever the vector of real valued functions \( f(\cdot) \), by the theorem of total probability we have:

\[
E \left[ f \left( v_{\Theta|I} \right) \right] = \sum_{i=1}^M E \left[ f \left( v_{\Theta|I} \right) \mid S_i \right] \mathcal{P}(S_i), \quad (4a)\]

where \( \Theta_i = \Theta P_i^T. \) However, from a computational point of view, it is wiser to express (4a) as:

\[
E \left[ f \left( v_{\Theta|I} \right) \right] = \sum_{i=1}^M E \left[ f \left( v_{\Theta|I} \right) \mid U_i \right] \mathcal{P}(U_i), \quad (4b)\]

where \( U_i = \{ u : u_i \geq -\Delta_i \mu_s \} \) and \( u_i = -\Delta_i (s - \mu_s) \sim \mathcal{N}_{M-1} (0, \Delta_i C_s \Delta_i^T), \Delta_i = \Delta P_i. \) As (1):

\[
\xi_i = u_i \mid v_{\Theta|I} \sim \mathcal{N}_{M-1+MP} \left( \mu_{\xi_i}, C_{\xi_i} \right), \quad (5a)\]
\[
\mu_{\xi_i} = \mu_{v_{\Theta|I}}, \quad C_{\xi_i} = \begin{bmatrix} C_{u_1, v_{\Theta|I}} & C_{u_2, v_{\Theta|I}} \\ C_{u_1, v_{\Theta|I}}^T & C_{v_{\Theta|I}} \end{bmatrix}, \quad (5b)\]

where \( v_{\Theta|I} = (P_i \otimes I_P) v_\Theta, \) therefore:

\[
E \left[ f \left( v_{\Theta|I} \right) \mid U_i \right] = E \left[ f \left( v_{\Theta|I} \right) \mid u_i \right] \mid U_i\right], \quad (4b)\]
and (4b) can be finally rewritten as:

\[
E \left[ f \left( v_{\Theta|I} \right) \right] = \sum_{i=1}^M E \left[ f \left( v_{\Theta|I} \right) \mid u_i \right] \mathcal{P}(U_i), \quad (6)\]

In particular:
\[
\mu_{v_{\Theta|I}} = E \left[ v_{\Theta|I} \right] = \sum_{i=1}^M E \left[ v_{\Theta|I} \mid u_i \right] \mathcal{P}(U_i), \quad (7a)\]
\[
C_{v_{\Theta|I}} = \sum_{i=1}^M E \left[ \left( v_{\Theta|I} - E \left[ \mid u_i \right] \right) \mid u_i \right] \mathcal{P}(U_i), \quad (7b)\]

where:
\[
E \left[ v_{\Theta|I} \mid u_i \right] = \mu_{v_{\Theta|I}} + C_{u_i, v_{\Theta|I}}^{-1} u_i, \quad (7c)\]
\[
C_{v_{\Theta|I}} \mid u_i = C_{v_{\Theta|I}} - C_{u_i, v_{\Theta|I}} C_{u_i, v_{\Theta|I}}^{-1} C_{u_i, v_{\Theta|I}}^T, \quad (7d)\]

Then a smart exploitation of (7a-c) yields (12):

\[
\mu_{v_{\Theta|I}} = \sum_{i=1}^M \left( \mu_{u_i} + C_{u_i, v_{\Theta|I}} \mu_{v_{\Theta|I}} \right) \mathcal{P}_i, \quad (8a)\]
\[
C_{v_{\Theta|I}} = \sum_{i=1}^M \left( C_{v_{\Theta|I}} \mid u_i + \mu_{v_{\Theta|I}} \mu_{v_{\Theta|I}}^T \right) \mathcal{P}_i, \quad (8b)\]

where:
\[
P_i = \mathcal{P}(U_i), \mathcal{P}(U_i) = E \left[ \mid u_i \right] \mid U_i \right], \quad (8c)\]
\[
C_{v_{\Theta|I}} \mid u_i = C_{v_{\Theta|I}} - C_{u_i, v_{\Theta|I}} C_{u_i, v_{\Theta|I}}^{-1} C_{u_i, v_{\Theta|I}}^T, \quad (8d)\]

As shown in [2], \( \{P_i, \mathcal{P}_i, \mathcal{P}_i \}_{i=1}^M \) can be computed by resorting to algorithms proposed by Genz [1] for numerical evaluation of multivariate normal distributions and moments over domains included in \([-10, 10]^M\). Note that the use of (3b) in conjunction with (8a-b) yields a generalization of (2) obtained for \( P = 1 \) and \( a \equiv a = 1 \).

The correctness of expressions (8a) and (8b) can be checked (see Appendix) by inspection of the case where the column vectors of matrix \( \Theta \) are i.i.d., which has been addressed in [4].

In the two sources case, \( \Theta = \{\theta_1, \theta_2\} \in M_{L_2} (P, 2) \) and:

\[
\mu_{v_{\Theta|I}} = \begin{bmatrix} \mu_{\theta_1} \\ \mu_{\theta_2} \end{bmatrix}, \quad C_{v_{\Theta|I}} = \begin{bmatrix} C_{11} \pm C_{12} \\ C_{12} \pm C_{22} \end{bmatrix}, \quad (9)\]

Moreover \( \mu_{\theta_1} = -\mu_{\theta_2}, P_1 + P_2 = 1, e_1 - e_2 = E \left[ u_1 \right] = 0, \quad R_1 + R_2 = E \left[ u_1^2 \right] = \frac{\sigma^2}{\sigma^2 + \alpha^2} (C_1 + C_2 - 2C_{12}), a \), leading to:

\[
C_{\theta_m} = E \left[ \mid \theta_m \mid \right] - E \left[ \mid \theta_m \mid \right] E \left[ \mid \theta_m \mid \right]^T, \quad (9)\]
\[
E \left[ \mid \theta_m \mid \right] = \mu_m + (-1)^m \left( \mu_2 - \mu_1 \right) P_2 + \frac{\alpha^2 + \beta^2}{\sigma^2} e_2, \quad \mu_2 = \mu_1 + \mu_m + \mu_m P_2, \quad (10)\]

where \( a_1 = (C_{12} - C_1) \alpha \) and \( a_2 = (C_{12}^T - C_2) \alpha \).
3. APPLICATION TO PARAMETRIC INFERENCE

3.1. Maximum likelihood estimation

The ongoing success of ML estimators (MLEs) originates from the fact that, under reasonably general conditions on the probabilistic observation model [13][14], the MLEs are, in the limit of large sample support, Gaussian distributed and efficient. Additionally, if the observation model is Gaussian, some additional asymptotic regions of operation yielding, for a subset of MLEs, Gaussian distributed and efficient estimates, have also been identified at finite sample support [15][16][17][18][19]. However, many estimation problems are actually unidentifiable unless they are regularized by imposing the ordering of some unknown parameters. For illustration purposes, let us consider $L$ independent observations of the linear model [11]:

$$y(l) = H(\Theta)x(l) + v(l), \quad 1 \leq l \leq L,$$

where $y(l)$ is the vector of samples of size $N$, $M$ is the number of signal sources, $x(l)$ is the vector of complex sources for the $M$ sources for the $l^{th}$ observation, $\Theta = [\theta_1 \ldots \theta_M]$, $H(\Theta) = [h(\theta_1) \ldots h(\theta_M)]$ and $h(\cdot)$ is a vector of $N$ parametric functions depending on a vector of $P$ unknown parameters $\Theta \in \Omega \subset \mathbb{R}^P$, $v(l)$ is complex noise, and $\Omega$ is a Gaussian conditional model [11][15]:

$$\mathbb{E}[\theta_i \mid x_i] \equiv E[\theta_i, x_i] = \hat{\theta}_i,$$

where $\hat{\theta}_i$ is an analytical solution of $\theta_i$. For instance, if (10) is a Gaussian conditional model [11][15][16][17][18][19]. However, many estimation problems are actually unidentifiable unless they are regularized by imposing the ordering of the unknown parameters $\{\theta_{m}\}_{m=1}^{M}$. A straightforward ordering of $\{\theta_{m}\}_{m=1}^{M}$ arises in the computation of MLEs which requires a multidimensional non linear optimization for which analytical solutions are in general not available. For instance, if (10) is a Gaussian conditional model [11][15]:

$$\hat{\Theta} = [\hat{\theta}_1 \ldots \hat{\theta}_M] = \arg \max_{\Theta} \prod_{l=1}^{L} y(l)^H H(\Theta) y(l),$$

where $H_A = A (A^H A)^{-1} A^H$. Therefore one has to resort to numerical search techniques, generally compatible with computer programming, such as the conversion of a $P M$-dimensional search grid over $\Omega^M$ into a 1-dimensional search grid. For instance, in Matlab, this is done by the sub2ind.m function which returns the linear index equivalents to the specified subscripts for each dimension of an $N$-dimensional array. For example, if $P = 2$ and $\Theta \in \Omega = [a_1, b_1] \times [a_2, b_2]$, one can generate a rectangular grid over $\Omega$ [11]:

$$G = \left\{ \left( \frac{a_1 + i_1 \delta_1}{2}, \frac{a_2 + i_2 \delta_2}{2} \right) \mid 0 \leq i_1 \leq I_1 \quad 0 \leq i_2 \leq I_2 \right\},$$

and convert each $\Theta \in G$ into an equivalent linear search index $s = i_1 + (I_1 + 1) i_2$. Thereby, in practice (11) becomes $s = \arg \max_{s} \prod_{l=1}^{L} y(l)^H H(\Theta(s)) y(l)$ and the issue of model identifiability is solved by ordering $s$ yielding $s(\Delta)$. If $\delta_1$ and $\delta_2$ are small enough, then $s \approx \Theta^T a - s_0$ where $a^T = (1/\delta_1, (I_1 + 1)/\delta_2)$ and $s_0 = a_1 + (I_1 + 1) a_2/\delta_2$. Then, since the ordering does not depend on $s_0$, $\hat{\Theta} = \Theta(\hat{s}(\Delta))$ are induced order statistic of $\hat{s}(\Delta)$ (2b), that is concomitants of $\hat{s}(\Delta)$. Therefore, the asymptotic performance analysis of the MLE of MLEs is refined by the study of concomitant of ordered multivariate normal distribution.

For illustration purposes, let us consider a radar system consisting of a 1-element antenna array receiving scaled, time delayed, and Doppler-shifted echoes of a known complex bandpass signal $e(t) e^{-j2\pi f_c t}$, where $f_c$ is the carrier frequency. A standard observation model of a radar antenna receiving a pulse train of $I$ pulses of duration $\delta t_0$ and bandwidth $B$, with a pulse repetition interval $\delta t$ is given by (10) where $[20] L = 1$, $N = [\delta t/B]$, $\Theta^T = (\tau, \omega)$, $h(\Theta) = \psi(\tau) \otimes \phi(\tau)$, $\psi(\omega) = 1, \ldots, e^{j2\pi \omega ((1-1)\delta t)}$, $\phi(\tau) = (e^{-\tau}, \ldots, e^{j2\pi \delta t/B \tau})$, $\tau$ and $\omega$ denoting the delay and the Doppler-shift associated to a target. The MLEs of $\Theta$ are asymptotically efficient and Gaussian, and for 2 targets [21]:

$$C_{\Theta} = \text{CRB}_{\Theta} = 2 \text{Re} \left\{ J(\Theta) \circ \left( \left(x_1^T x_1^* \otimes I_{2 \times 2} \right) \right)^{-1} \right\},$$

where $J(\Theta)$ is given in [22]. We consider a high resolution scenario in terms of $\Theta$, that is a small Doppler-Shift $\delta \omega = 1/(12f_c)$ ($I = 8$) and a small delays difference $\delta \tau = 1/(8B)$ ($\delta t_0 = 32/B$); $e(t)$ is a linear chirp. Figure (1) displays the empirical and theoretical MSE to CRB ratio (shrinkage factor) averaged over the two targets for both the delay and Doppler shift. The empirical MSE are assessed with $10^6$ Monte-Carlo trials from the normally distributed vector associated with the asymptotic behavior of $\nu_o \sim N(\nu_o, \text{CRB}_{\nu_o})$. The theoretical MSE is computed from (9). The match between theoretical and empirical results provides an empirical proof of the exactness of $C_{\nu_o\Theta}$ and $E \left[ \nu_o^2 \right]$ given in (8a-8b).

3.2. Monitoring of the states range of Kalman filters

We consider the class of real linear discrete state-space (LDSS) models represented with the state and measurement equations:

$$x_k = F_{k-1} x_{k-1} + w_{k-1}, \quad y_k = H_k x_k + v_k,$$

where the time index $k \geq 1$, $x_k$ is the $M$-dimensional state vector, $y_k$ is the $N$-dimensional measurement vector and the model matrices $F_k$ and $H_k$ are known. The process noise sequence $\{w_k\}$ and the measurement noise sequence $\{v_k\}$, as well as the initial state $x_0$ are Gaussian random vectors. $x_{k} \equiv \tilde{x}_{k} | y_{1}, \ldots, y_{k}$ denotes an estimate of $x_k$ based on measurements up to and including time $k$. If $\{w_k\}$ and $\{v_k\}$ are uncorrelated, then the minimum MSE
estimator \( \hat{x}_{k|k} \) for LDSS models has a recursive predictor/corrector format, aka the Kalman filter (KF) \[23][24, \ldots \[m\], \theta \[m\] = \delta_{m\theta} \]

which include \[4, (1)(2)(17)\].

Finally, remembering that \( \text{vec} (AXB) = (B^T \otimes A) \text{vec} (X) \) and \( (A \otimes B)(C \otimes D) = AC \otimes BD \), one obtains from (3a-b):

Finally, remembering that vec \( (AXB) = (B^T \otimes A) \text{vec} (X) \) and \( (A \otimes B)(C \otimes D) = AC \otimes BD \), one obtains from (3a-b):

Finally, remembering that vec \( (AXB) = (B^T \otimes A) \text{vec} (X) \) and \( (A \otimes B)(C \otimes D) = AC \otimes BD \), one obtains from (3a-b):
5. REFERENCES

[22] T. Menni, E. Chaumette, P. Larzabal, "Reparameterization and constraints for CRB: duality and a major inequality for system analysis and design in the asymptotic region", in Proc. IEEE ICASSP, 2012