Statistics and Applications {ISSN 2454-7395 (online)}
Special Issue in Memory of Prof. C R Rao
Volume 22, No. 3, 2024 (New Series), pp 471–483
http://www.ssca.org.in/journal

On Retrieving Multivariate Data Sets from Their Moments

Serge B. Provost[1], S. Ejaz Ahmed[2] and Zhaoqi Yang[1]

[1]*Department of Statistical and Actuarial Sciences, The University of Western Ontario, London, Canada*
[2]*Department of Mathematics and Statistics, Brock University, Saint Catharines, Canada*

Received: 12 June 2024; Revised: 15 September 2024; Accepted: 23 September 2024

Abstract

This paper introduces several methodologies that solve the inverse problem of recovering a multivariate sample from subsets of its associated marginal and joint integer moments. These results rely in part on their univariate counterpart, which is examined in some detail. It is also explained that some of them also apply to complex-valued data sets. Several illustrative examples are presented.

Key words: Inverse problem; Multivariate samples; Joint moments; Complex-valued observations.

AMS Subject Classifications: 62H05; 11P70; 47A57.

1. Introduction

Evidently, one can readily evaluate sample moments from a given data set. The problem being considered herein, which consists of retrieving a sample of multivariate observations from certain of its marginal and joint sample moments, can be regarded as an inverse problem.

Inverse problems generally involve determining certain causes from some effects. They currently constitute a rich field of research. For instance, they appear in the Mathematics Subject Classification index in connection with quantum theory, optics, harmonic analysis, trigonometry, linear operators, and electromagnetic theory. Inverse problems of various nature have, for example, also found applications in geophysics (Zhdanov, 2015), acoustics (Klyuchinskiy *et al.* 2020), image processing (Zou *et al.* 2021), astronomy (Escárate *et al.* 2023), system identification (Blanken and Oomen, 2020), language processing (Nakanishi, 2024), machine learning (Koffer *et al.* 2023), signal processing (Giovannelli and Idier, 2015) and tomography (Mohamad-Djafari, 2013).

The results introduced in this paper imply that a certain number of marginal and joint moments actually hold all the information that is contained in a given data set since the latter

Corresponding Author: Serge B. Provost
Email: provost@stats.uwo.ca

can be entirely retrieved from the former. Accordingly, such moments constitute sufficient statistics. To some extent, this remark provides a justification for making use of moment-based statistical methodologies such as the density function estimation techniques advocated in Provost and Zheng (2015), Provost and Ha (2016), Jin *et al.* (2016), Zareamoghaddam *et al.* (2017), Kang *et al.* (2019), Provost *et al.* (2020) and Provost and Zang (2024).

The problem of recovering a univariate sample of size n from its first n moments is considered in Section 2 where the applicability of the result is discussed. The case of bivariate observations and their sample moments is addressed in Section 3 where generalizations to complex-valued and multivariate data sets are explored. All the results and their extensions are illustrated by means of numerical examples. Lastly, some concluding remarks are offered in Section 4.

2. A theorem relating a univariate data set to its moments

In this section, we state a result that was established in Provost *et al.* (2020), explain that it holds in the complex domain, and discuss related considerations. Two numerical examples are provided as well.

Theorem 1: A data set of size n can be recovered from the first n moments of the sample.

The proof of this result is given in the Appendix for the sake of completeness. The following example illustrates the steps to follow when applying Theorem 1.

Example 1: Let $n = 5$ and the sample be $\{1.2,\ 3.4,\ 6.7,\ 8.1,\ 11.9\}$. The moments of orders zero to five are $1, 6.26, 53.022, 511.6790, 5301.7767, 57492.260726$ and, for $j = 0, 1, 2, 3, 4, 5$, the e_j's as defined in the Appendix, are $1, 31.3, 357.29, 1814.543, 3910.731, 2634.91704$. According to equation (1), the resulting polynomial is then $-2634.91704+3910.731\,x-1814.543\,x^2 +357.29\,x^3 - 31.3\,x^4 + x^5$, its five roots being $\{1.2, 3.4, 6.7, 8.1, 11.9\}$.

We note that the proof of Theorem 1 remains valid in the complex domain. It should also be observed that any loss of precision can be avoided by making use of fractions.

Example 2: Let $n = 3$ and the sample be $\{2.4 + 5.1\,i,\ 6.7 - 9.5\,i, 11.8 + 1.4\,i\}$, that is, $\{\frac{12}{5} + \frac{51\,i}{10}, \frac{67}{10} - \frac{19\,i}{2}, \frac{59}{5} + \frac{7\,i}{5}\}$ in fractional form. The moments of orders zero, one, two and three are $1, \frac{209}{30} - \ i, \frac{2389}{100} - \frac{1163\,i}{50}$, and $-\frac{56531}{1500} + \frac{38517\,i}{1000}$, and for $j = 0, 1, 2, 3$, the e_j's as defined in the Appendix are $1, \frac{209}{10} - 3\,i, \frac{17807}{100} - \frac{2781\,i}{100}, \frac{93192}{125} + \frac{56127\,i}{250}$. The polynomial, $x^3 - \left(\frac{209}{10} - 3\,i\right) x^2 + \left(\frac{17807}{100} - \frac{2781\,i}{100}\right) x - \left(\frac{93192}{125} + \frac{56127\,i}{250}\right)$, is then obtained from equation (1) and, as expected, its three roots are $\{\frac{12}{5} + \frac{51\,i}{10}, \frac{67}{10} - \frac{19\,i}{2}, \frac{59}{5} + \frac{7\,i}{5}\}$.

Since there exists a one-to-one correspondence between the observations and their associated empirical distribution function, the following corollary to Theorem 1 holds.

Corollary 1: Given a simple random sample of size n from a continuous distribution, its empirical distribution function F_n is uniquely specified by the first n sample moments.

In light of the strong law of large numbers, for every fixed x, the empirical distribution function $F_n(x)$ will converge almost surely to the underlying distribution function $F(x)$. Moreover, given a simple random sample of size n, the Glivenko-Cantelli theorem states that

$$\sup_{x\in\Re}|F_n(x)-F(x)|$$

tends to zero almost surely, and that the convergence of $F_n(x)$ to $F(x)$ is uniform. However, as was aptly pointed out by Ričardas Zitikis, a colleague of the first author, a contradiction would ensue if one were to let n tend to infinity in Corollary 1 as this result would then imply that, given the integer moments of a random variable, its distribution could be specified uniquely. This is clearly not the case since there exists distinct distributions whose integer moments are all identical.

Consider for example the following density functions:

$$f_1(x)=\frac{1}{4}e^{-\sqrt{|x|}},\quad x\in\Re,$$

and

$$f_2(x)=\frac{1}{4}e^{-\sqrt{|x|}}\Big(\cos(\sqrt{|x|})+1\Big),\quad x\in\Re,$$

which are plotted in Figure 1.

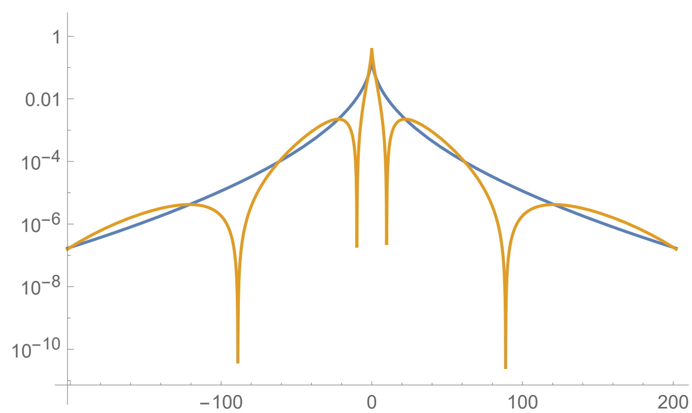

Figure 1: Plots of $f_1(x)$ and $f_2(x)$ on a logarithmic scale for $-200<x<200$

Although these two distributions are clearly distinct, their k^{th} moment,

$$m_1(k)=\frac{1}{2}\Big((-1)^k+1\Big)\,\Gamma(2k+2)$$

and

$$m_2(k)=\frac{1}{2}\Big((-1)^k+1\Big)\,\Gamma(2k+2)\Big(1-\frac{\sin(k\,\pi/2)}{2^{k+2}}\Big),$$

happen to coincide for $k=0,1,2,\ldots.$

To summarize, in the limit, F_n can specify the underlying population distribution function. However, as previously illustrated, a population distribution function F may not be uniquely specified by an infinite sequence of its integer moments. Thus, Corollary 1 cannot be extended beyond finite values of n.

It should also be pointed out that moment-based methodologies lend themselves to the modeling of massive data sets since only a moderate number of moments are needed to apply such techniques, as opposed to other approaches such as those based on likelihoods for which all the observations are required. Actually, ample information can generally be secured from a fairly limited number of moments, whereas each data point contains an equal amount of information that is inversely proportional to the sample size. Moreover, once a new set of observations, $\{x_{n_1+1}, \ldots, x_n\}$, becomes available in addition to an initial dataset, $\{x_1, \ldots, x_{n_1}\}$, there is no need to make use of each of the n_1 original data points to compute the moments since the h^{th} updated moment will then be $(n_1 m_h + \sum_{i=n_1+1}^{n} x_i^h)/n$ where m_h denotes the h^{th} sample moment as evaluated from the initial data set.

3. On recovering multivariate samples from their moments

The four propositions introduced in this section enable one to retrieve bivariate sets of observations from some of their marginal and joint moments—or those of their component-wise ranks, the observations on each variable being assumed to be distinct. It is explained that each of the proposed methodologies also apply to multivariate data sets and that two of them hold in the complex domain. Several numerical examples are provided.

Proposition 1: A bivariate sample $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ can be retrieved from the first n marginal moments of the first variable, that is,

$$m_{1,0}, \ldots, m_{n-1,0}, m_{n,0},$$

in conjunction with the following bivariate sample moments:

$$m_{0,1}, m_{1,1}, \ldots, m_{n-1,1},$$

where $m_{j,k}$ denotes the moment of orders j and k, which is equal to $\sum_{i=1}^{n} x_i^j \, y_i^k / n$.

Proof: In light of Theorem 1, the observations on the first variable, namely, $x_1, \ldots, x_n$ can be retrieved from the given marginal moments. The remainder of the proof relies on a representation of the joint moments that involves a Vandermonde matrix.

It is assumed that the following joint moments are known:

$$m_{j,1} = \frac{1}{n} \sum_{1=1}^{n} x_i^j \, y_i, \quad j = 0, \ldots, n-1.$$

This system of equations can be equivalently expressed as follows:

$$\frac{1}{n} \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \\ x_1^2 & x_2^2 & \cdots & x_n^2 \\ \vdots & \vdots & \ddots & \vdots \\ x_1^{n-1} & x_2^{n-1} & \cdots & x_n^{n-1} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} m_{0,1} \\ m_{1,1} \\ m_{2,1} \\ \vdots \\ m_{n-1,1} \end{pmatrix}$$

where the above matrix is a Vandermonde matrix, which is nonsingular since the x_i's are assumed to be nonidentical. Note that the vector of y_j's which is the unique solution of this linear system, enables one to pair each of them appropriately with the corresponding x_i. □

Remark 1: Given the definition of $m_{j,k}$, it is apparent that the order of the n bivariate sample points is immaterial. Thus, in applications, it suffices to set a certain order for the x_i's, and the y_j's to be associated with these x_i's will be properly ordered in the solution vector of the linear system.

Additionally, we note that the $2n$ moments that are specified in Proposition 1 are jointly sufficient statistics, since they provide enough information to recover the entire bivariate sample of ordered observations—which, incidentally, requires $3n-1$ pieces of information, namely, the observations on each variable and the ranks of $n-1$ observations on the second component relative to those on the first.

Example 3: Let the sample be $\{(1,7),(2,2),(5,3)\}$. Given the marginal moments on the first variable, one can determine that the observations on the first variable are 1, 2 and 5. Additionally, let the joint moments of orders (0,1), (1,1) and (2,1), that is, $m_{0,1}=4$ $m_{1,1}=26/3$ and $m_{2,1}=30$, be available. The solution of the following system, which is (7,2,3), yields the values of the y_j's to be associated with the x_i's:

$$\frac{1}{3}\begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 5 \\ 1^2 & 2^2 & 5^2 \end{pmatrix}\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}=\begin{pmatrix} 4 \\ 26/3 \\ 30 \end{pmatrix}.$$

As is the case for univariate observations, complex-valued bivariate or multivariate observations can also be recovered. This can be readily achieved by initially implementing Theorem 1 and then, solving a linear system of equations involving complex values.

Example 4: Let $\{(2.4+5.1\,i,7.3-1.8\,i),(6.7-9.5\,i,2.2),(11.8+1.4\,i,9.8\,i)\}$ be the sample to recover. Note that if the first three marginal moments of the first variables are given, one can retrieve the three observations on the first component, which happens to be the univariate data set utilized in Example 2. Now, assume that the joint moments of orders (0,1), (1,1) and (2,1), namely, $m_{0,1}=19/6+(8\,i)/3$, $m_{1,1}=231/25+(851\,i)/20$ and $m_{2,1}=-(105469/600)+(640219\,i)/1500$ are available. As expected, the solution of the linear system,

$$\frac{1}{3}\begin{pmatrix} 1 & 1 & 1 \\ 12/5+51\,i/10 & 67/10-19\,i/2 & 59/5+7\,i/5 \\ -81/4+612\,i/25 & -1134/25-1273\,i/10 & 3432/25+826\,i/25 \end{pmatrix}\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}$$

$$=\begin{pmatrix} 19/6+8\,i/3 \\ 231/25+851\,i/20 \\ -105469/600+640219\,i/1500 \end{pmatrix}$$

is $\{y_1,y_3,y_4\}=\{73/10-(9\,i)/5,11/5,(49\,i)/5\}$.

A trivariate observation vector $(x_i,y_i,z_i),\ i=1,\ldots,n$, can be similarly recovered if, in addition to the the first n marginal moments of the first variable from which the x_i's can be specified, one knows $m_{0,1,0},m_{1,1,0},\ldots,m_{n-1,1,0}$ which will yield the y_j's associated with the x_i's, as well as $m_{0,0,1},\ldots,m_{0,n-1,1}$ which will then yield the z_k's associated with the y_j's. By proceeding in like fashion, Proposition 1 can extended to sets of multivariate observations.

Example 5: Let the sample be $\{(2,4,6),(7,3,1),(5,6,3)\}$. Given the marginal moments on the first variable, we can determine that the observations on that variable are 2, 5 and 7 and, in light of Remark 1, we may let $\{x_1, x_2, x_3\} = \{2, 7, 5\}$ (or any other permutation thereof). The joint moments of orders (0,1,0), (1,1,0) and (2,1,0) are $m_{0,1,0} = 13/3$ $m_{1,1,0} = 59/3$ and $m_{2,1,0} = 313/3$, and the joint moments of orders (0,0,1), (0,1,1) and (0,2,1) are $m_{0,0,1} = 10/3$ $m_{0,1,1} = 15$ and $m_{0,2,1} = 71$. The solutions of the systems of equations,

$$\frac{1}{3}\begin{pmatrix} 1 & 1 & 1 \\ 2 & 7 & 5 \\ 2^2 & 7^2 & 5^2 \end{pmatrix}\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 13/3 \\ 59/3 \\ 313/3 \end{pmatrix}$$

and

$$\frac{1}{3}\begin{pmatrix} 1 & 1 & 1 \\ 4 & 3 & 6 \\ 4^2 & 3^2 & 6^2 \end{pmatrix}\begin{pmatrix} z_1 \\ z_2 \\ z_3 \end{pmatrix} = \begin{pmatrix} 10/3 \\ 15 \\ 71 \end{pmatrix},$$

yield the values of the y_j's to be paired with the x_i's, that is, $\{y_1, y_2, y_3\} = \{4, 3, 6\}$, and then those of the z_k's to be paired with the y_j's namely, $\{z_1, z_2, z_3\} = \{6, 1, 3\}$).

Proposition 1 can also be extended as follows: Given the marginal moments of the first variable up to order n, any additional set of n joint moments that does not include any of the first variable marginal moments can be utilized to recover the sample. The resulting system of equations can be solved by making use of an array of computing packages. This flexibility in the selection of joint moments also applies in the case of multivariate observations.

Example 6: Let the sample be $\{(2,4),(5,6),(7,3)\}$. Given the first three marginal moments on the first variable which are $\{14/3, 26, 476/3\}$, it can be determined from Theorem 1 that the observations on that variable are 2, 5 and 7. Now, assume that the joint moments of orders (0,1), (1,2) and (2,3), namely, $m_{0,1} = 13/3$, $m_{1,2} = 275/3$ and $m_{2,3} = 6979/3$ are available. It then suffices to solve of system, $\{y_1 + y_2 + y_3 = 13,\ 2y_1^2 + 5y_2^2 + 7y_3^2 = 275,\ 4y_1^3 + 25y_2^3 + 49y_3^3 = 6979\}$ to obtain the corresponding values for the second variable, that is, (4,6,3).

Proposition 2: A bivariate sample of size n can be retrieved from the first n marginal sample moments of each variable, that is, $m_{i,0},\ i = 1, \ldots, n,$ and $m_{0,j},\ j = 1, \ldots, n,$ where $m_{i,j}$ denotes the sample moment of orders i and j, *in conjunction with* the ranks of the observations within each variable—or equivalently those of the corresponding pseudo-observations.

Pseudo-observations are the component-wise ranks of the data points divided by n. Note that all the pseudo-observations originating from a given sample can be secured from the associated empirical copula, as originally defined by Deheuvels (1979).

Proof: As previously explained, the data on each variable can be retrieved from the marginal moments by appealing to Theorem 1. Then, given the ranks of the observations on each variable, the observations can be appropriately paired. □

Example 7: Let the original sample be {(1,7), (2,2), (5,3)}. First, it can be determined from the first three marginal moments of each variable that the observations on the first and second variables are respectively {1, 2 ,5}, and {2, 3, 7}. If in addition, it is known that the ranks of the observations on each component are $[r_1, s_1] = [1, 3], [r_2, s_2] = [2, 1]$ and

$[r_3, s_3] = [3, 2]$, then, it can readily be determined that the sample points are (1,7), (2,2) and (5,3).

This approach can be directly extended to sets of multivariate observations.

Example 8: Consider the following sample of trivariate observations: {(2, 5, 7), (3, 4, 8),(1, 3, 6)}. Given the first three marginal moments of each of the three variables, it can be determined from Theorem 1 that the observations on the first, second and third components are {1, 2, 3}, {3, 4, 5} and {6, 7, 8}, respectively. If it is also known that the ranks of these component-wise observations are $[r_1, s_1, t_1] = [2, 3, 2]$, $[r_2, s_2, t_2] = [3, 2, 3]$, and $[r_3, s_3, t_3] = [1, 1, 1]$, it can then be readily determined that the sample points are (2, 5, 7), (3, 4, 8) and (1, 3, 6).

Proposition 3: A random sample of size n arising from a continuous bivariate distribution can be retrieved from the first n marginal moments of each variable, that is, $m_{i,0},\ i = 1, \ldots, n$ and $m_{0,j},\ i, j = 1, \ldots, n,$ *in conjunction with* the joint moments, $m^*_{0,1}, \ldots, m^*_{n-1,1}$, of the ranks of the observations.

Proof: In light of Theorem 1, the observations on each variable, namely, $x_1, \ldots, x_n$, and $y_1, \ldots, y_n$, can be recovered from the marginal moments. The remainder of the proof relies on a representation of the joint moments of the ranks that involves a Vandermonde matrix. Let again r_i and s_i denote the ranks of the observations with respect to the first and second variables. By assumption, the joint moments, $m^*_{0,1}, \ldots, m^*_{n-1,1}$, of the ranks are known with, in general,

$$m^*_{j,k} = \frac{1}{n} \sum_{i=1}^{n} s_i^j\, r_i^k, \ \ j = 0, \ldots, n-1.$$

Note that $m^*_{0,1} = (n+1)/2$. This system of equations can be equivalently expressed as follows:

$$\frac{1}{n} \begin{pmatrix} 1 & 1 & \cdots & 1 \\ r_1 & r_2 & \cdots & r_n \\ r_1^2 & r_2^2 & \cdots & r_n^2 \\ \vdots & \vdots & \ddots & \vdots \\ r_1^{n-1} & r_2^{n-1} & \cdots & r_n^{n-1} \end{pmatrix} \begin{pmatrix} s_1 \\ s_2 \\ s_3 \\ \vdots \\ s_n \end{pmatrix} = \begin{pmatrix} m^*_{0,1} \\ m^*_{1,1} \\ m^*_{2,1} \\ \vdots \\ m^*_{n-1,1} \end{pmatrix}$$

where the above matrix is a Vandermonde matrix, which is nonsingular since the r_i's are distinct. Note that the unique solution of this linear system will yield $s_1, \ldots, s_n$, and associate each of them appropriately with the corresponding r_i, which will enable one to correctly pair the known x_i's and y_j's. □

Remark 2: Given the definition of $m^*_{j,k}$, it is apparent that the order of the n bivariate sample points does not matter, since the pair of ranks corresponding to a given bivariate observation will remain unchanged. Thus, in applications, it suffices to set a certain order for the r_i's, and the s_j's to be associated with these r_i's will be properly ordered in the solution vector of the linear system.

Example 9: Let the sample be $\{(1, 7), (2, 2), (5, 3)\}$. Given the marginal moments on each variable, one can retrieve the observations on the first variables, namely, 1, 2 and 5, as well as the observations on the second variables which are 2, 3 and 7. It now remains to pair them using Proposition 3. We have to determine the second component of the following paired ranks: $[r_1, s_1] = [1, 3], [r_2, s_2] = [2, 1]$ and $[r_3, s_3] = [3, 2]$, that is, $[s_1, s_2, s_3] = [3, 1, 2]$. The joint moments of the ranks of orders (0,1), (1,1) and (2,1) are $m^*_{0,1} = 2$, $m^*_{1,1} = 11/3$ and $m^*_{2,1} = 25/3$, respectively. Solving the following system will yield the ranks of the second component, that is, [3,1,2], and enable one to correctly pair the data points:

$$\frac{1}{3}\begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \\ 1^2 & 2^2 & 3^2 \end{pmatrix}\begin{pmatrix} s_1 \\ s_2 \\ s_3 \end{pmatrix} = \begin{pmatrix} 2 \\ 11/3 \\ 25/3 \end{pmatrix}.$$

This result can be generalized to the multivariate case by proceeding as in the generalization of the Proposition 1, except that in this case, the joint moments of the ranks are utilized in addition to the marginal sample moments of each variable. As well, joint moments of the ranks other than those specified in Proposition 3 can be utilized as was done in Example 6 in conjunction with certain joint moments of the observations.

Example 10: Let the sample be $\{(2, 4, 6), (7, 3, 1), (5, 6, 3)\}$. Given the marginal moments on each variable, one can retrieve the observations on the first, second and third variables, that is, {2, 5, 7}, {3, 4, 6}, and {1, 3, 6}, respectively. We then have to determine the ranks of the entries in second and third components, namely, $[s_1, s_2, s_3] = [2, 1, 3]$ and $[t_1, t_2, t_2] = [3, 1, 2]$ and end up with the following set of ranks: $[r_1, s_1, t_1] = [1, 2, 3]$, $[r_2, s_2, t_2] = [3, 1, 1]$, and $[r_3, s_3, t_3] = [2, 3, 2]$, which enables us to retrieve the original data set.

The joint moments of the ranks of orders (0,1,0), (1,1,0) and (2,1,0) are $m^*_{0,1,0} = 2$, $m^*_{1,1,0} = 11/3$ and $m^*_{2,1,0} = 23/3$, respectively. Let the given joint moments of the ranks of orders (0,0,1), (0,1,1) and (0,2,1) be $m^*_{0,0,1} = 2$, $m^*_{0,1,1} = 13/3$ and $m^*_{0,2,1} = 31/3$, respectively.

We started off with $r_1 = 1$, $r_2 = 3$ and $r_3 = 2$; however, as per Remark 2, any permutation thereof will lead to the data set with its trivariate observations appearing in a different order. Thus, we first solve the following linear system, which will yield the ranks of the second component entries, that is, [2,1,3]:

$$\frac{1}{3}\begin{pmatrix} 1 & 1 & 1 \\ 1 & 3 & 2 \\ 1^2 & 3^2 & 2^2 \end{pmatrix}\begin{pmatrix} s_1 \\ s_2 \\ s_3 \end{pmatrix} = \begin{pmatrix} 2 \\ 11/3 \\ 23/3 \end{pmatrix}.$$

The solution of the linear system that follows will then yield the ranks of the third component entries, which are [3,1,2]:

$$\frac{1}{3}\begin{pmatrix} 1 & 1 & 1 \\ 2 & 1 & 3 \\ 2^2 & 1^2 & 3^2 \end{pmatrix}\begin{pmatrix} t_1 \\ t_2 \\ t_3 \end{pmatrix} = \begin{pmatrix} 2 \\ 13/3 \\ 31/3 \end{pmatrix}.$$

Proposition 4: A bivariate sample of size n can be retrieved on the basis of the first n marginal sample moments of each variable in conjunction with any *single* additional joint sample moment that does not involve moments of order zero.

Proof: On applying Theorem 1, the set of observations on each variable can be determined from the marginal moments. Then, given the ordered observations on the first variable, there will be a unique permutation of the observations on the second variable that will yield the given joint moment. □

This assumes that the observations have been recorded with sufficient precision.

Example 11: Consider the sample $\mathcal{S} = \{(1,7),(2,2),(5,3)\}$. Given the marginal moments of each variables, it can be determined that the first and second component values will respectively be {1,2,5} and {2,3,7}. Assuming for instance that, it is known that the joint moment $m_{1,1} = 26/3$, and, for instance, setting the the observations on the first component in increasing order, that is, 1,2,5, we are seeking the permutation of {2,3,7} among the 6 possible ones that will yield the same joint moment of orders 1 and 1. This process will lead to the identification of the correct bivariate data points that constitute the sample $\mathcal{S}$. The 6 possible pairs of observations and their joint moment of order (1,1) are:

$\{(1,7),(2,3),(5,2)\} \Rightarrow\ m_{1,1} = 23/3$
$\{(\mathbf{1},\mathbf{7}),(\mathbf{2},\mathbf{2}),(\mathbf{5},\mathbf{3})\} \ \Rightarrow\ m_{1,1} = \mathbf{26/3}$
$\{(1,2),(2,7),(5,3)\} \ \Rightarrow\ m_{1,1} = 31/3$
$\{(1,2),(2,3),(5,7)\} \ \Rightarrow\ m_{1,1} = 43/3$
$\{(1,3),(2,2),(5,7)\} \ \Rightarrow\ m_{1,1} = 42/3$
$\{(1,3),(2,7),(5,2)\} \ \Rightarrow\ m_{1,1} = 27/3.$

Accordingly, we select the bold-faced set as the original sample since its joint moment of order (1,1) coincides with that of $\mathcal{S}$.

Proposition 4 which, incidentally, is implementable in the case of moderately sized samples, can readily be extended to sets of multivariate observations.

Example 12: Consider the sample $\mathcal{S} = \{(2,4,6),(7,3,1),(5,6,3)\}$. Given the first three marginal moments of each variable, it can be determined that the observations on the first, second and third components are {2,5,7}, {3,4,6}, and {1,3,6}, respectively. Assuming for instance that, it is known that the joint moment $m_{1,1,1} = 53$, and setting the observations on the first component in increasing order, that is, $\{2,5,7\}$, we are seeking the permutation of $\{3,4,6\}$ and that of $\{1,3,6\}$ that will yield the same joint moment. This will enable us to identify the correct triplet of trivariate observations comprising $\mathcal{S}$. The 36 possible sets of observations and their joint moments of order (1,1,1) are:

$\{(2,3,1),\ (5,4,3),\ (7,6,6)\} \Rightarrow m_{1,1,1} = 106,$
$\{(2,3,1),\ (5,4,6),\ (7,6,3)\} \Rightarrow m_{1,1,1} = 84,$
$\{(2,3,3),\ (5,4,1),\ (7,6,6)\} \Rightarrow m_{1,1,1} = \frac{290}{3},$
$\{(2,3,3),\ (5,4,6),\ (7,6,1)\} \Rightarrow m_{1,1,1} = 60,$
$\{(2,3,6),\ (5,4,1),\ (7,6,3)\} \Rightarrow m_{1,1,1} = \frac{182}{3},$
$\{(2,3,6),\ (5,4,3),\ (7,6,1)\} \Rightarrow m_{1,1,1} = 46,$
$\{(2,3,1),\ (5,6,3),\ (7,4,6)\} \Rightarrow m_{1,1,1} = 88,$
$\{(2,3,1),\ (5,6,6),\ (7,4,3)\} \Rightarrow m_{1,1,1} = 90,$
$\{(2,3,3),\ (5,6,1),\ (7,4,6)\} \Rightarrow m_{1,1,1} = 72,$

$\{(2,3,3),\,(5,6,6),\,(7,4,1)\} \Rightarrow m_{1,1,1} = \frac{226}{3}$,
$\{(2,3,6),\,(5,6,1),\,(7,4,3)\} \Rightarrow m_{1,1,1} = 50$,
$\{(2,3,6),\,(5,6,3),\,(7,4,1)\} \Rightarrow m_{1,1,1} = \frac{154}{3}$,
$\{(2,4,1),\,(5,3,3),\,(7,6,6)\} \Rightarrow m_{1,1,1} = \frac{305}{3}$,
$\{(2,4,1),\,(5,3,6),\,(7,6,3)\} \Rightarrow m_{1,1,1} = \frac{224}{3}$,
$\{(2,4,3),\,(5,3,1),\,(7,6,6)\} \Rightarrow m_{1,1,1} = 97$,
$\{(2,4,3),\,(5,3,6),\,(7,6,1)\} \Rightarrow m_{1,1,1} = 52$,
$\{(2,4,6),\,(5,3,1),\,(7,6,3)\} \Rightarrow m_{1,1,1} = 63$,
$\{(2,4,6),\,(5,3,3),\,(7,6,1)\} \Rightarrow m_{1,1,1} = 45$,
$\{(2,4,1),\,(5,6,3),\,(7,3,6)\} \Rightarrow m_{1,1,1} = \frac{224}{3}$,
$\{(2,4,1),\,(5,6,6),\,(7,3,3)\} \Rightarrow m_{1,1,1} = \frac{251}{3}$,
$\{(2,4,3),\,(5,6,1),\,(7,3,6)\} \Rightarrow m_{1,1,1} = 60$,
$\{(2,4,3),\,(5,6,6),\,(7,3,1)\} \Rightarrow m_{1,1,1} = 75$,
$\{(2,4,6),\,(5,6,1),\,(7,3,3)\} \Rightarrow m_{1,1,1} = 47$,
$\{(\mathbf{2},\mathbf{4},\mathbf{6}),\,(\mathbf{5},\mathbf{6},\mathbf{3}),\,(\mathbf{7},\mathbf{3},\mathbf{1})\} \Rightarrow \mathbf{m_{1,1,1}} = \mathbf{53}$,
$\{(2,6,1),\,(5,3,3),\,(7,4,6)\} \Rightarrow m_{1,1,1} = 75$,
$\{(2,6,1),\,(5,3,6),\,(7,4,3)\} \Rightarrow m_{1,1,1} = 62$,
$\{(2,6,3),\,(5,3,1),\,(7,4,6)\} \Rightarrow m_{1,1,1} = 73$,
$\{(2,6,3),\,(5,3,6),\,(7,4,1)\} \Rightarrow m_{1,1,1} = \frac{154}{3}$,
$\{(2,6,6),\,(5,3,1),\,(7,4,3)\} \Rightarrow m_{1,1,1} = 57$,
$\{(2,6,6),\,(5,3,3),\,(7,4,1)\} \Rightarrow m_{1,1,1} = \frac{145}{3}$,
$\{(2,6,1),\,(5,4,3),\,(7,3,6)\} \Rightarrow m_{1,1,1} = 66$,
$\{(2,6,1),\,(5,4,6),\,(7,3,3)\} \Rightarrow m_{1,1,1} = 65$,
$\{(2,6,3),\,(5,4,1),\,(7,3,6)\} \Rightarrow m_{1,1,1} = \frac{182}{3}$,
$\{(2,6,3),\,(5,4,6),\,(7,3,1)\} \Rightarrow m_{1,1,1} = 59$,
$\{(2,6,6),\,(5,4,1),\,(7,3,3)\} \Rightarrow m_{1,1,1} = \frac{155}{3}$,
$\{(2,6,6),\,(5,4,3),\,(7,3,1)\} \Rightarrow m_{1,1,1} = 51$

Accordingly, we select the bold-faced set as the original sample since its joint moment of order (1,1,1) coincides with that of $\mathcal{S}$.

Proposition 4 can as well be extended to complex-valued samples.

Example 13: Consider the sample $\mathcal{S}$=$\{(5.4 + 6.1\,i, 9 + 3.4\,i), (6.7, 3.3\,i), (8\,i, 1.9)\}$. Given the first three marginal moments of the first and second components, which are respectively $\{121/30+(47\,i)/10, -679/75+(549\,i)/25, -5783/120-(68451\,i)/1000\}$ and $\{109/30+(67\,i)/30, 518/25+(102\,i)/5, 423739/3000+(750959\,i)/3000\}$, one can determine the three entries in each of the two components as was done in Example 2 for the univariate case. Now, assume that, additionally, $m_{1,1}$=1393/150 + (11057 i)/300, is provided. On keeping the observations on first component in a given order and permuting those of the second component, only one of the six joint moments of orders 1 and 1 so obtained will equal $m_{1,1}$, the corresponding set of paired observations being those included in $\mathcal{S}$.

4. Concluding remarks

Four methodologies were introduced for the purpose of recovering a multivariate data set from certain of its associated marginal and joint moments as evaluated from the observations or their component-wise ranks. In fact, two of them also hold in the complex domain. For a given multivariate sample, the evaluation of the marginal and joint moments is straightforward and constitutes a direct problem. As explained in the Introduction, the results introduced in this paper actually solve the inverse problem consisting of recovering the original observations on the basis of certain marginal and joint moments.

Interestingly, a parallel can be established between Proposition 2 which makes use of a number of marginal moments and all the component-wise ranks of the observations—or, equivalently, the pseudo-observations—to recover the entire sample, and Sklar's theorem as introduced by Sklar (1959), which states that a joint distribution can be expressed in terms of the marginal distributions and a function that depends only on the pseudo-observations, which is referred to as a copula. In fact, copulas completely account for the dependence between the variables. Several nonparametric copula density estimation techniques were recently proposed in Provost and Zang (2024). For an introduction to copulas and related results, the reader is referred to Nelsen (2006). All the calculations were carried out with the symbolic computing package *Mathematica*, the code being available from the first author upon request.

Acknowledgements

The financial support of the Natural Sciences and Engineering Research Council of Canada is gratefully acknowledged.

References

Blanken, L. and Oomen, T. (2020). Kernel-based identification of non-causal systems with application to inverse model control. *Automatica*, **114**, 1–7.

Deheuvels, P. (1979). La fonction de dépendance empirique et ses propriétés: un test non paramétrique d'indépendance. *Académie Royale de Belgique–Bulletin de la Classe des Sciences*, **65**, 274–292.

Escárate, P., Curé, M., Araya, I., Coronel, M., Cedeño, A. L., Celedon, L., Cavieres, J., Aguero, J. C., Arcos, C., Cidale, L. S., and Levenhagen, R. S. (2023). A method to deconvolve stellar profiles – The non-rotating line utilizing Gaussian sum approximation. *Astronomy & Astrophysics*, **676**, A44. doi.org/10.1051/0004-6361/202346587

Giovannelli, J. F. and Idier, J. (2015). *Regularization and Bayesian Methods for Inverse Problems in Signal and Image Processing*. New York: John Wiley & Sons, Inc.

Jin T., Provost S. B, and Ren J. (2016). A moment-based methodology for approximating the distribution of aggregate losses. *Scandinavian Actuarial Journal*, **16**, 216–245.

Kang, J. S. J., Provost, S. B., and Ren, J. (2019). Moment-based density approximation techniques as applied to heavy-tailed distributions. *International Journal of Statistics and Probability*, **8**, 1–23.

Klyuchinskiy, D., Novikov, N., and Shishlenin, M. (2020). A modification of gradient descent method for solving coefficient inverse problem for acoustics equations. *Computation*, **8**, 1–14. doi:10.3390/computation8030073

Kofler, A., Altekruger, F., Antarou Ba, F., Kolbitsch, C., Papoutsellis, E., Schote, D., and Papafitsoros, K. (2023). Learning regularization parameter-maps for variational image reconstruction using deep neural networks and algorithm unrolling. *SIAM Journal on Imaging Sciences*, **16**, 2202–2246.

Mohamad-Djafari, A. (2013). *Inverse Problems in Vision and 3D Tomography*. New York: John Wiley & Sons. ISBN 978-1-118-60046-7.

Nakanishi, T. (2024). *Detection of Latent Gender Biases in Data and Models Using the Approximate Generalized Inverse Method*. In 2024 IEEE 18th International Conference on Semantic Computing (ICSC), 191–196.

Nelsen, R .B. (2006) *An Introduction to Copulas*, 2nd ed. New York: Springer.

Provost, S. B. and Ha, H.-T. (2016). Distribution approximation and modelling via orthogonal polynomial sequences. *Statistics*, **50**, 454–470.

Provost, S. B. and Zang, Y. (2024). Nonparametric Copula Density Estimation Methodologies. *Mathematics*, **12**, 398. https://doi.org/10.3390/math12030398

Provost, S. B., Zareamoghaddam, H., Ahmed, S. E., and Ha, H.-T. (2020). The generalized Pearson family of distributions and explicit representation of the associated density functions. *Communications in Statistics - Theory and Methods*, **51**, 5590–5606.

Provost, S. B. and Zheng, S. Z. (2015). Polynomially adjusted saddlepoint density approximations. *International Journal of Statistics and Probability*, **4**, 1–11.

Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publication de l'Institut Statistique de l'Université de Paris*, **8**, 22–231.

Zareamoghaddam, H., Provost, S. B., and Ahmed, E. (2017). A moment-based bivariate density estimation methodology applicable to big data modeling. *Journal of Probability and Statistical Science*, **30**, 155–162.

Zhdanov, M. S. (2015). *Inverse Theory and Applications in Geophysics*. Amsterdam: Elsevier.

Zou, Z., Shi, T., Shi, Z., and Ye, J. (2021). Adversarial training for solving inverse problems in image processing. *IEEE Transactions on Image Processing*, **30**, 2513–2525.

APPENDIX

Proof of Theorem 1

Let $\mathcal{S} = \{x_1, x_2, \ldots, x_n\}$ be a sample of size n and $\mathcal{M} = \{m_1, m_2, \ldots, m_n\}$ where $m_h = \sum_{i=1}^{n} x_i^h / n$. According to the fundamental theorem of algebra, $p(z) = a_0 + a_1 z + \cdots + a_{n-1} z^{n-1} + z^n$ is uniquely defined by its coefficients a_i's and it is also uniquely specified by its n roots x_i's for $i = 1, \ldots, n$. Moreover, given $\mathcal{S}$, the coefficients of $p(x)$ can be expressed in terms of the sequence of moments $\mathcal{M}$ via the Newton-Girard identity. Accordingly, a given polynomial of degree n, say $p(x)$, can be represented as follows:

$$\prod_{i=1}^{n} (x - x_i) = \sum_{k=0}^{n} (-1)^{n-k} \, e_{n-k} \, x^k, \tag{1}$$

where $e_0 = 1$ and

$$e_\ell = \frac{n}{\ell} \sum_{h=1}^{\ell} (-1)^{h-1} \, e_{\ell-h} \, m_h, \quad \ell = 1, \ldots, n. \tag{2}$$

Thus, given the first n sample moments associated with $\mathcal{S}$, a sample of size n, one can express the right-hand side of (1) as a polynomial whose roots are precisely $\{x_1, x_2, \ldots, x_n\}$. This establishes that $\mathcal{S}$ is uniquely specified by $\mathcal{M}$.