STAT 330 Project: Multivariate Skewness-based Projection Pursuit

Samuel Horvath
VCC, KAUST
Thuwal, Saudi Arabia
sameul.horvath@kaust.edu.sa

Abstract

Projection pursuit is a multivariate statistical technique aimed at finding interesting low-dimensional data projections by maximizing a measure of interestingness commonly known as projection index. Building on the original work of Loperfido [4], which focus on the computational difficulties inherent to the maximization of the projection index in one dimension, we propose the extension to the multivariate approach. Our problem is addressed within the framework of skewness-based projection pursuit, focused on data projections with the highest third standardized cumulants. First, we use the right dominant singular vector of the third multivariate, standardized moment to start the maximization procedure. Second, we maximize the proposed an iterative coordinate-wise algorithm for skewness maximization with respect to generally accepted Mardia's definition of skewness [6], which we show is the right measure to consider and which is equivalent to the maximization of a sixth-order polynomial with normalized entries. We propose an efficient implementation of our method and illustrate its usage on several real-world datasets.

1 Introduction

We start with the definition of univariate skewness, then we move to the extensions to multivariate distributions such as projection pursuit or Mardia's definition of multivariate skewness.

1.1 Univariate skewness

In probability theory and statistics, skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. The skewness value can be positive, zero, negative, or undefined.

For a unimodal distribution, negative skew commonly indicates that the tail is on the left side of the distribution, and positive skew indicates that the tail is on the right. In cases where one tail is long but the other tail is fat, skewness does not obey a simple rule. For example, a zero value means that the tails on both sides of the mean balance out overall; this is the case for symmetric distribution, but can also be true for an asymmetric distribution where one tail is long and thin, and the other is short but fat.

Formally, skewness of univariate random variable X can be defined as the third standardized moment

$$\widetilde{\mu}_3 = \mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right], \tag{1}$$

where μ is the mean, σ is the standard deviation and $\mathrm{E}[\cdot]$ is the expectation operator. Sometimes this quantity is also referred to as Pearson's moment coefficient of skewness or simply the moment coefficient of skewness.

1.2 Multivariate Skewness

Here, we mention two extensions, which were proposed in the literature. The first one, introduced by Mardia in 1970 [6], focuses on the observation that the standard univariate skewness, which we defined in the previous section, can be approximately expressed as the correlation of the mean and variance of the random sample from a population. We denote these quantities by $\overline{X}$ and S^2, respectively. Formally,

$$corr^2(\overline{X}, S^2) \approx \frac{1}{2}\widetilde{\mu}_3^2 = \frac{1}{2}\beta_{1,1} \tag{2}$$

where the equation holds up to factor $\mathcal{O}(n^{-1})$ under assumption that 4-th cumulant κ_4 is negligible. Mardia [6] proposed the following extension of β_1 to the multivariate case. Similarly to the 1D case, he considered $\overline{\mathbf{X}}$ and $S_v^2 = vec(S^2)$, which needs to be vectorized as this is originally in matrix form. Since correlation is only defined for univariate variables, the canonical correlation, extension of the classical definition of correlation to the multivariate case, is used. One extra step is to only take trace (sum of the eigenvalues) of the matrix $\Sigma_{\overline{\mathbf{X}}\overline{\mathbf{X}}}^{-1/2}\Sigma_{\overline{X}S_v^2}\Sigma_{S_v^2S_v^2}^{-1}\Sigma_{S_v^2\overline{\mathbf{X}}}\Sigma_{\overline{\mathbf{X}}\overline{\mathbf{X}}}^{-1/2}$ to avoid expensive maximization. This procedure leads to the well-known formula of Mardia's definition of multivariate skewness

$$\beta_{1,p} = \mathrm{E}\left[\left(\left(\mathbf{X}_1 - \mu\right)^\top \Sigma^{-1}\left(\mathbf{X}_2 - \mu\right)\right)^3\right], \tag{3}$$

where $\mathbf{X}_1$ and $\mathbf{X}_2$ are independent identically distributed copies of a random p-dimensional vector $\mathbf{X}$.

The second measure is to Malkovich and Afifi [5], originally referred to as a generalized measure of skewness. Here, the authors focus on projection pursuit, where they look for a vector $c \in \mathbb{R}^p$ such that 1D skewness β_1 of $c^\top\mathbf{X}$ is maximized. Formally,

$$\beta_1^\star = \max_{c\in\mathbb{R}^p}\beta_1(c) = \max_{c\in\mathbb{R}^p}\mathrm{E}\left[\left(\frac{c^\top\mathbf{X} - c^\top\mu}{\sqrt{c^\top\Sigma c}}\right)^3\right]. \tag{4}$$

Note that we can limit vector c to lie on the unit ball in $\mathbb{R}^p$ denoted by $\mathbb{S}^{p-1}$ and this would lead to the equivalent result.

2 Problem formulation

In this section, we define the framework in which we operate. Let $\mathbf{X} = (X_1, X_2, \ldots, X_p)^\top$ be a p-dimensional random vector with bounded finite third moment $\mathrm{E}\left[|X_i|^3\right] < \infty$ for $i = 1, 2, \cdots, p$. We use standard definition of the third moment of $\mathbf{X}$ is $p^2 \times p$ matrix $\mathbf{M}_{3,\mathbf{X}} = \mathrm{E}\left[\mathbf{X}\otimes\mathbf{X}^\top\otimes\mathbf{X}\right]$. For instance, for $p = 3$

$$\mathbf{M}_{3,\mathbf{X}} = \mathrm{E}\left[(X_1X_2X_3)\otimes\begin{pmatrix}X_1\\X_2\\X_3\end{pmatrix}\otimes(X_1X_2X_3)\right] = \begin{pmatrix}m_{111} & m_{112} & m_{113}\\ m_{112} & m_{122} & m_{123}\\ m_{113} & m_{123} & m_{133}\\ m_{112} & m_{122} & m_{123}\\ m_{122} & m_{222} & m_{223}\\ m_{123} & m_{233} & m_{233}\\ m_{113} & m_{123} & m_{133}\\ m_{123} & m_{223} & m_{233}\\ m_{133} & m_{233} & m_{333}\end{pmatrix}. \tag{5}$$

This definition let us to equivalently reformulate (4) to

$$\beta_1^\star = \max_{c\in\mathbb{R}^p}\frac{(c^\top\otimes c^\top)\mathbf{M}_{3,X-\mu}c}{\left(c^\top\Sigma c\right)^{3/2}}. \tag{6}$$

Further normalization of distribution X to $\mathbf{Z} = \Sigma^{-1/2}(\mathbf{X} - \mu)$ leads to

$$\beta_1^\star = \max_{c \in \mathbb{S}^{p-1}} (c^\top \otimes c^\top) \mathbf{M}_{3,\mathbf{Z}} c. \tag{7}$$

For the Mardia's definition of multivariate skewness this would lead to

$$\beta_{1,p} = \mathrm{E}\left[(\mathbf{Z}_1^\top \mathbf{Z}_2)^3\right], \tag{8}$$

where $\mathbf{Z}_1$ and $\mathbf{Z}_2$ are independent random variables with the same distribution as Z.

Equipped with all of these equations, we are ready to proceed with the problem formulation. We are interested to find matrix C of the following structure

$$C = \begin{bmatrix} — \ c_1^\top \ — \\ — \ c_2^\top \ — \\ \vdots \\ — \ c_k^\top \ — \end{bmatrix}, \tag{9}$$

where $k < p$ and

$$\langle c_i, c_j \rangle = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases},$$

such that the multivariate skewness of $C\mathbf{Z}$ is maximized. Note that this is equivalent to the problem of finding such k dimensional subspace in which the skewness of $\mathbf{X}$ is maximized. This can be as an optimization problem of the form

$$C, \lambda = \arg\min_{C,\lambda} \left\| \mathbf{M}_{3,\mathbf{Z}} - \sum_{i=1}^{k} \lambda_i c \otimes c^\top \otimes c \right\|^2 \tag{10}$$

Loperfido [4] proposed to use sequential approach, wherein each step such c is found, which is an optimal solution of (7) and then $\mathbf{M}_{3,Z}$ is updated such that effect of vector c is removed from data using linear regression and $\mathbf{M}_{3,Z}$ is recomputed. This approach is very similar to Principal Component Analysis (PCA). Unfortunately, unlike PCA, this approach does not lead to the best result as for high-order tensor (greater than 2), sequential optimization potentially leads to different, not necessarily optimal, solution [3]. Since we deal with high-order tensors, there is no hope to construct such a sequential algorithm, which can provably converge to the optimal solution. On the other hand, it was shown that best lower rank approximation is plagued by computational difficulties [1], thus we would like to still have a sequential algorithm and the question is whether the method proposed in [4] is an optimal sequential algorithm. The answer is that it is not for the case $k > 1$ for a simple reason, which is that it does not optimize the true objective as it assumes independence. Thankfully, this can be fixed using Mardia's definition of multivariate skewness. The most important observation is the following lemma

Lemma 1. *Optimizing the skewness in form of* (10) *is equivalent to maximization of Mardia's definition of skewness for* $\widetilde{\mathbf{Z}} = C\mathbf{Z}$.

Proof. First, note, that optimization of (10) is equivalent to finding such low-dimensional projection of normalized random variable $\mathbf{Z}$ for which the Frobenius norm of the third moment of the projection is maximized. Formally, this is equivalent to

$$\max_C \left\| \mathbf{M}_{3,C\mathbf{Z}} \right\|^2. \tag{11}$$

Since

$$E(\widetilde{\mathbf{Z}}) = C\mathrm{E}\left[\mathbf{Z}\right] = 0 \text{ and } \Sigma_{\widetilde{\mathbf{Z}},\widetilde{\mathbf{Z}}} = CI_pC^\top = I_k,$$

(11) can be rewritten to

$$\max_C \sum_{i,j,k} \mathrm{E}\left[\widetilde{Z}_i \widetilde{Z}_j \widetilde{Z}_k\right]^2,$$

which exactly matches Equation (2.16)– definition of $\beta_{1,k}$ for $\widetilde{\mathbf{Z}}$ in [6]. This concludes the proof. □

This Lemma motivates us to propose a new optimal sequential algorithm, which in each step consider as an objective Mardia's multivariate skewness. We discuss the implementation details in the next section. Note that this is not the optimal algorithm in terms of a global solution for the reasons we discussed at the beginning of this section.

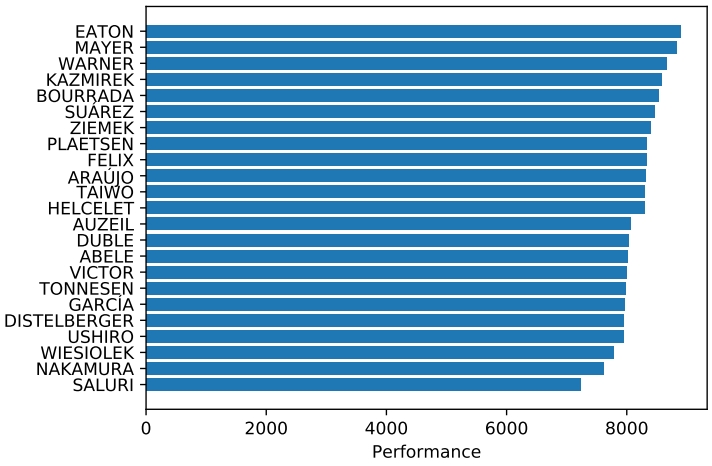

Figure 1: Bar chart of the total points scored by each athlete.

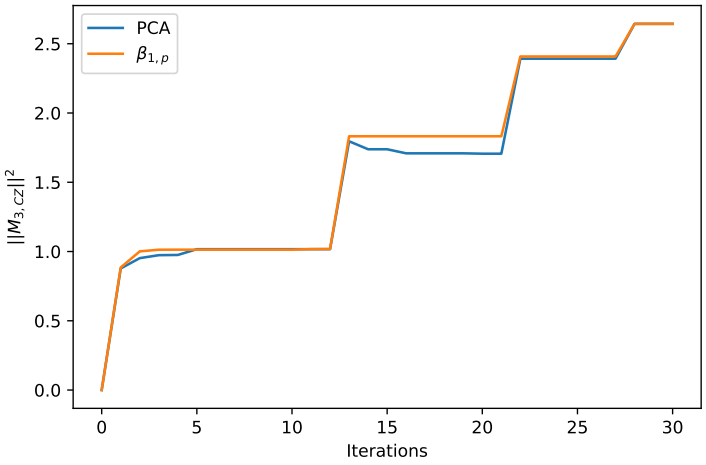

Figure 2: Comparison of PCA like method of Loperfido and our newly proposed method, where the objective is multivariete Mardia's definition of skewness.

3 New Algorithm

Up to this point, we focused on population quantities. From this point onwards, we shift our focus to samples. We assume to have access to n i.i.d. samples $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \in \mathbb{R}^p$, which are rows of data matrix $\mathbf{X}$. As skewness is invariant to linear transformations, we work with $\mathbf{Z}$ which is a normalized version of $\mathbf{X}$ with each row

$$\mathbf{z}_i = S^{-1/2}(\mathbf{x}_i - \overline{x}) \tag{12}$$

where $\overline{\mathbf{x}}$ is the sample mean

$$\overline{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i$$

and S is the sample covariance matrix

$$S = \frac{1}{n}\sum_{i=1}^{n} (\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^\top.$$

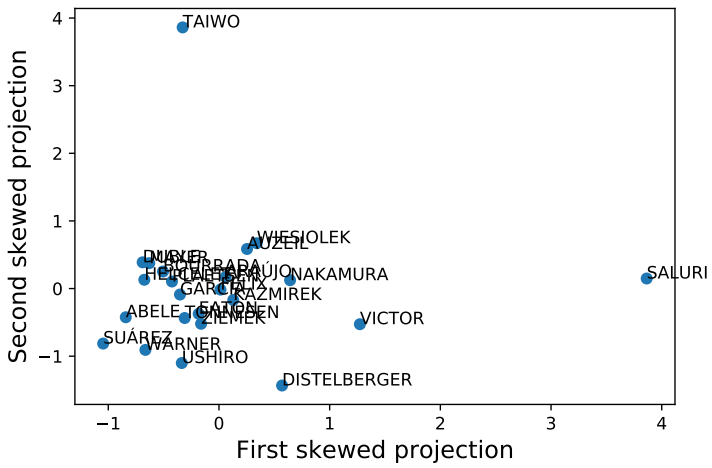

Figure 3: Scatterplot of the most skewed projection (horizontal axis) and the most skewed projection among those orthogonal to it (vertical axis).

On the sample level, our objective function has the following form

$$\max_{\widetilde{\mathbf{Z}}=\mathbf{Z}C^\top} \frac{1}{n^2}\sum_{i,j}^{n}\left[\widetilde{\mathbf{z}}_i^\top \widetilde{\mathbf{z}}_j\right]^3 = \max_{\widetilde{\mathbf{Z}}=\mathbf{Z}C^\top} \frac{1}{n^2}\sum_{i,j}^{n}\left[\sum_{l=1}^{k}(c_l^\top \mathbf{z}_i)(c_l^\top \mathbf{z}_j)\right]^3. \tag{13}$$

We denote $c_0 = \vec{0}$

$$(\mathbf{M}^k_{shift})_{ij} = \sum_{l=0}^{k}(c_l^\top \mathbf{z}_i)(c_l^\top \mathbf{z}_j).$$

Its matrix form is following

$$\mathbf{M}^k_{shift} = \sum_{l=0}^{k}(\mathbf{Z}c_l)(\mathbf{Z}c_l)^\top. \tag{14}$$

Using this notation, we can rewrite our objective to

$$\max_{\widetilde{\mathbf{Z}}=\mathbf{Z}C^\top} \frac{1}{n^2}\sum_{i,j}^{n}(\mathbf{M}^k_{shift})^3_{ij}$$

Note that this matrix interpretation allows us to exploit vectorization in our implementation, which can significantly speed-up our computations.

We denote set $\mathbb{S}_l$ to be intersection of unit ball $\mathbb{S}^{p-1}$ and complement of range space of vector $c_0, \ldots, c_{l-1}$. Formally, $\mathbb{S}_l = \mathbb{S}^{p-1} \cap \text{range}(c_0, \ldots, c_{l-1})^\perp$. Our algorithm follows

Algorithm 1

1: **for** $l = 1, \ldots k$ **do**
2: $\quad c_l = \arg\max_{c\in\mathbb{S}_l} \frac{1}{n^2}\sum_{i,j}^{n}(\mathbf{M}^{k-1}_{shift} + (\mathbf{Z}c)(\mathbf{Z}c)^\top)^3_{ij}$
3: **end for**

If we replace $\mathbf{M}^{k-1}_{shift}$ by zero matrices at line 2 at each step, it would lead to the algorithm proposed in [4]. The maximization procedure does not have a closed-form solution and needs to be solved numerically. For this, we use a coordinate-ascend algorithm, wherein each epoch (p iterates), we randomly shuffle the order of coordinates. For each coordinate, we optimize our objective using L-BFGS. Unlike [4], we can't find an analytical solution as our objective is not third but sixth-order polynomial. On the other hand, this does not bring much overhead as the convergence is usually

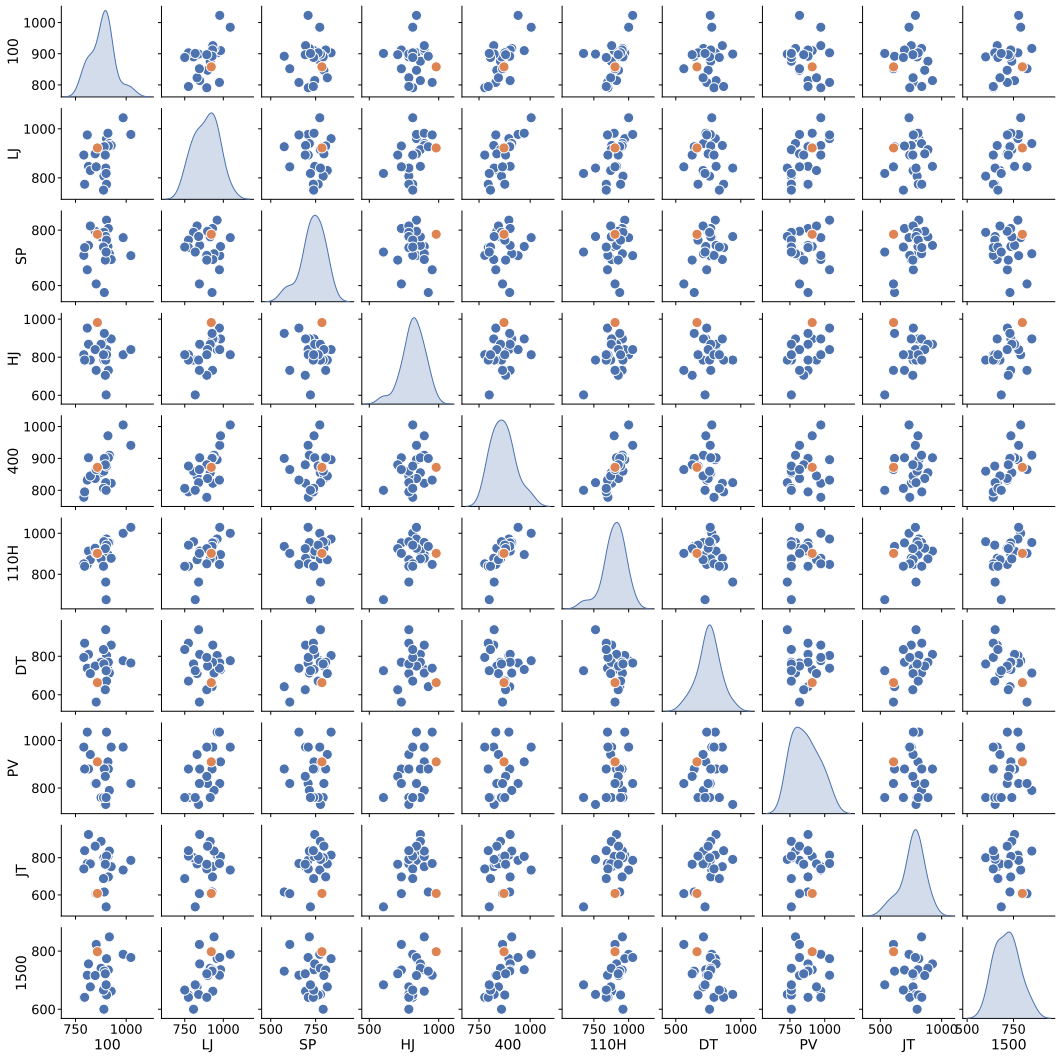

Figure 4: Multiple scatterplot of the points scored by the decathletes in the ten events. Taiwo is depicted in orange.

reached within a few iterations. Moreover, our method still guarantees provable progress in each iteration in contrast to high-order power methods. As proposed by previous works [2, 4], our initial guess for c_l is the is dominant right singular vector of the third cumulant of $\mathbf{Z}^l$ ($\mathbf{M}_{3,\mathbf{Z}^l}$), where $\mathbf{Z}^l$ is projection of $\mathbf{Z}$ onto range$(c_0, \ldots, c_{l-1})^\perp$.

4 Experiments

We implemented our method in Python using fast C based *numpy* package for handling vector operations and *scipy* package for Singular Value Decomposition (SVD), computing fractional power ($S^{-1/2}$) and L-BFGS for maximization. Our code is publicly available on `github.link.com`.

4.1 Guaranteed Improvement

As the first experiment, we demonstrate also empirically that the method proposed by Loperfido [4] does not optimize true objective and that actual skewness might decrease and it is no guarantee to increase in every single step. We demonstrate this behaviour on the well-known iris dataset. We construct all the skewed projections ($p = 4$), where for each projection, we run three epochs.

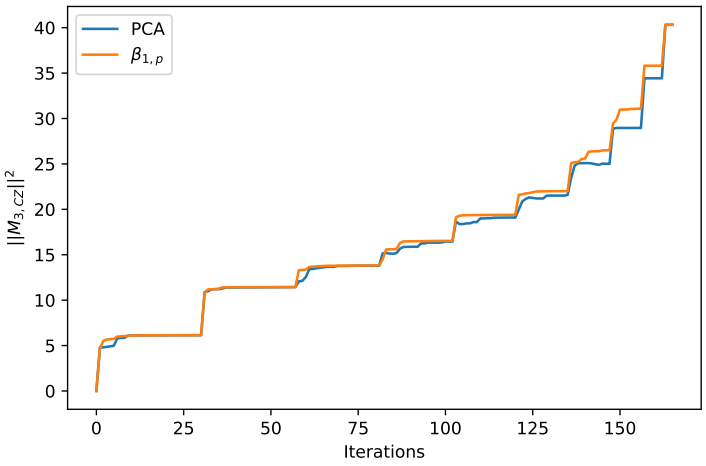

Figure 5: Comparison of PCA like method of Loperfido and our newly proposed method, where the objective is multivariete Mardia's definition of skewness.

Figure 2 shows that along the optimization process the overall skewness for independent PCA like approach [4] does not need to increase with every iteration, while our newly proposed method guarantees improvement in every single iteration. Note, that this is exactly predicted by our theory as our new formulation maximizes the true objective.

4.2 Decathlon data

The second experiment illustrates skewness-based projection pursuit using the results of the athletes competing in the decathlon at the Games of the XXXI Olympiad (Rio de Janeiro, Brazil, the year 2016). The dataset contains ten variables (the points scored in each event) and 23 cases (the decathletes who scored points in each event). It is freely available at `www.iaaf.org`, the official website of the IAAF (International Association of Athletics Federations). Confirmatory data analysis is very difficult due to the limited number of cases (especially when compared to the number of variables) and their nonrandom nature (each athlete is extraordinarily gifted for the decathlon). We shall, therefore, follow an exploratory approach, mainly based on graphical tools. Decathlon is a combined event in athletics consisting of ten events: one hundred meters (100M), long jump (LONGJ), shot put (SHOTP), high jump (HIGHJ), four hundred meters (400M), one hundred and ten meters hurdles (110H), discus throw (DISCT), pole vault (POLEV), javelin throw (JAVET), one thousand and five hundred meters (1500M). Each athlete had his performances recorded in seconds for a track event, meters for throwing events and centimetres for jumping events. All performances are converted into decathlon points (simply points, henceforth) according to the IAAF scoring tables. We chose to use points rather than performances because points are what ultimately matter for ranking decathletes.

On cannot spot any striking particular feature just by looking into Figure 3, which depicts the bar chart of the total points scored by each athlete. Ashton Eaton, the winner of the decathlon, does not stand far from his competitors, despite having achieved the second-best performance of all times (and is regarded as one of the best decathletes ever).

Similarly to Loperfido [4], we look into the first and second most skewed projections, which we can visualize and they could help us to indicate potential outliers. We depict this graph in Figure **??**. This graph indicates that possible candidates are Saluri (first skewed projection) and Taiwo (second skewed projection). Saluri might be considered as a natural candidate as he ended up last and scored worst, or nearly so, in several events, while Taiwo is more surprising one since he ranked eleventh and obtained about average scores in nearly all events. The possible explanation might be his extremely good or very bad performance as displayed in pair plot, Figure 4 (Taiwo's performance is orange dot).

As for the iris dataset, we also include the construction of all the projections to showcase the superiority of our new approach. For each projection, we run 3 epochs. The optimization procedure is displayed in Figure 5.

5 Conclusion

In this work, we propose the new sequentially optimal method for the multivariete skewness-based projection pursuit. We achieve the optimality by changing the objective, which we modify to be Mardia's definition of multivariete skewness. We show that this quantity is right objective to consider via proving its equivalence to high-order tensor factorization. Our method enjoys several desired features such as guarantee of improvement in every single step or possibility of efficient vectorized implementation.

References

[1] Pierre Comon, Gene Golub, Lek-Heng Lim, and Bernard Mourrain. Symmetric tensors and symmetric tensor rank. *SIAM Journal on Matrix Analysis and Applications*, 30(3):1254–1279, 2008.

[2] Lieven De Lathauwer, Pierre Comon, Bart De Moor, and Joos Vandewalle. Higher-order power method. *Nonlinear Theory and its Applications, NOLTA'95*, 1:91–96, 1995.

[3] Eleftherios Kofidis and Phillip A Regalia. On the best rank-1 approximation of higher-order supersymmetric tensors. *SIAM Journal on Matrix Analysis and Applications*, 23(3):863–884, 2002.

[4] Nicola Loperfido. Skewness-based projection pursuit: A computational approach. *Computational Statistics & Data Analysis*, 120:42–57, 2018.

[5] James Francis Malkovich and AA Afifi. On tests for multivariate normality. *Journal of the american statistical association*, 68(341):176–179, 1973.

[6] Kanti V Mardia. Measures of multivariate skewness and kurtosis with applications. *Biometrika*, 57(3):519–530, 1970.