Journal of Quantitative Analysis in Sports

Volume 7, *Issue* 1 | 2011 | *Article* 4

Model Averaging in Factor Analysis: An Analysis of Olympic Decathlon Data

Michael Schomaker, *Ludwig-Maximilians-Universität*
Christian Heumann, *Ludwig-Maximilians-Universität*

Recommended Citation:
Schomaker, Michael and Heumann, Christian (2011) "Model Averaging in Factor Analysis: An Analysis of Olympic Decathlon Data," *Journal of Quantitative Analysis in Sports*: Vol. 7: Iss. 1, Article 4.
Available at: http://www.bepress.com/jqas/vol7/iss1/4
DOI: 10.2202/1559-0410.1249

Model Averaging in Factor Analysis: An Analysis of Olympic Decathlon Data

Michael Schomaker and Christian Heumann

Abstract

This article presents a multivariate analysis of Olympic decathlon data based on maximum likelihood factor analysis. All results explicitly account for model selection uncertainty, which is inherent in any data-based selection process but mostly ignored in reports related to multivariate sports data. For this purpose, some well-established frequentist procedures that have so far been applied almost exclusively to regression analysis are adopted and transferred to the factor analytical context. The findings support the claim that decathlon contests consist of three dimensions. These dimensions seem to be similar to, but not exactly the same, as those found by Cox and Dunn (2002) via hierarchical cluster analysis.

KEYWORDS: dimension, frequentist procedures, multivariate analysis, averaging techniques, small samples

1 Introduction

The analysis of athletic contests has always played an important role for statistics in sports. For example, Brown (1946) studied paradoxes in traditional scoring systems, Chang et al. (2003) suggested a new score awarding method for decathlon events and Dawkins et al. (1994) analyzed olympic heptathlon data based on cluster and correspondence analysis.

In this article, we are concerned with results of the olympic decathlon contest in Athens, August 2004. The decathlon competition consists of ten track-and-field events run over two consecutive days; these are the 100 m race, long jump (LJ), shot-put (SP), high jump (HJ), 400 m race, 110 m hurdles (110 mh), discus (Dis), pole-vault (PV), javeline (Jav) and 1500 m race. The data are taken from the IAAF (International Amateur Athletic Federation) website archive and used to perform multivariate methods for dimension reduction, namely maximum likelihood factor analysis that explicitly accounts for model selection uncertainty. We are particulary interested in grouping the different disciplines to find latent factors which reflect the nature of decathlon competitions.

Decathlon data sets are typically hard to analyze as they are small, multidimensional and therefore weak in any structure supported from statistical methodology. This can be exemplarily seen in the detailed and carefully conducted multivariate description of decathlon data from Cox and Dunn (2002), where cluster analysis is applied upon five different data sets of decathlon championships held from 1991 to 1999: here, each data set supports other combinations of clusters for the ten events and this raises the question how stable recent findings really are. The problem of an adequate, robust and good choice for the description of such data relates to a typical model selection problem, whereas model selection uncertainty is obviously apparent and has to be incorporated fully into any statistical inference subsequent to a data-based selection step. It is nowadays mostly accepted that this can be primarily realized by applying model averaging schemes, which is the combination of estimators from many potential models, rather than relying on traditional model selection estimators. Several methods have been developed both from a Bayesian (see, e.g., Hoeting et al. (1999)) and frequentist (see, e.g., Wang et al. (2009)) point of view. Unfortunately these methods were applied and studied mainly in the context of regression analysis, where due to interaction effects and a large amount of possible transformations the number of potentially good models is basically larger than in other statistical fields such as autoregressive models, signal detection and factor analysis.

The aim of this paper is to explore the dimension of decathlon competitions and devise a good and robust model to describe this dimension appropriately. This relates to a process of reducing the number of variables under consideration to a suitable amount of latent factors having impact on the ten disciplines of decathlon. For this purpose we use maximum likelihood factor analysis and account for model selection uncertainty by applying well-developed, frequentist model averaging schemes.

The balance of this paper begins with the description of frequentist model averaging techniques and their application in factor analysis in section 2, followed by a detailed analysis of the 2004 olympic decathlon data in section 3 and a discussion about methodology, results and alternatives in section 4.

2 Methods

2.1 General Framework

Consider the $n \times p$ data matrix $X = (X_1, \ldots, X_p)$, where $X_j = (x_{1j}, \ldots, x_{nj})'$ is an $n \times 1$ vector of values of the j^{th} variable, $j = 1, \ldots, p$, and $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$, an $1 \times p$ vector containing the i^{th} observation of each of the p variables, $i = 1, \ldots, n$. Let X be multivariate distributed with density $f(X;\theta) \in \mathcal{F}$ where θ is an unknown parameter vector, $\mathcal{F} = \{f(X;\theta),\, \theta \in \Theta\}$ is a parameterized family of probability distributions and Θ is the corresponding parameter space.

Now, let $\mathcal{M} = \{M_1, \ldots, M_{k^*}\} \subset \mathcal{F}$ be a set of candidate models to describe the structure of X appropriately. A *model selection* procedure is one that singles out a "winning" model from the set $\mathcal{M}$ on the basis of a data-based model selection criterion Υ, for instance by choosing the model that minimizes the information criterion of Akaike (1973), AIC. Typically, all subsequent inference is then conducted within this single chosen model as if it was given *a priori*. This approach, though popular in the statisticians' daily routine, neglects uncertainty about the model choice. The consequences may be serious: estimates can be biased and the corresponding standard errors tend to be underestimated; see Hjort and Claeskens (2003) for a variety of examples. One possibility to overcome this problem exists in the application of *model averaging*, that compromises across some or all models of interest. Consequently, this results in the compromise estimator

$$\hat{\bar{\theta}} = \sum_{\kappa=1}^{k^*} w_\kappa \hat{\theta}_\kappa \qquad (1)$$

based on the estimators $\hat{\theta}_\kappa$ for each of the candidate models M_κ belonging to $\mathcal{M}$. The compromise estimator in (1) may be called a FMA estimator if θ in each model is estimated by a frequentist principle. It can be seen that any frequentist model selection (FMS) estimator is a special case of $\hat{\bar{\theta}}$ by assigning a value of 1 to a particular w_κ and 0 to all other w_κ's. In regard to the weight choice in (1), Buckland et al. (1997) proposed the following exponential AIC weights:

$$w_\kappa \;=\; \frac{\exp(-\frac{1}{2}\mathrm{AIC}_\kappa)}{\sum_{\kappa=1}^{k^*}\exp(-\frac{1}{2}\mathrm{AIC}_\kappa)}\,, \tag{2}$$

where AIC_κ is the AIC value of model $M_\kappa \in \mathcal{M}$. One may also construct weights based on values of a cross-validation criterion (Stone (1974)) or the Focused Information Criterion (Claeskens and Hjort (2003)) scores or by minimizing a Mallows criterion as suggested in recent studies by Hansen (2007). An important limitation of Hansen's approach, however, is that the optimality properties regarding the Mallows criterion apply only in the context of linear regression and not elsewhere.

Approximate standard errors for (1) may be obtained either via bootstrapping or the formula of Buckland et al. (1997),

$$\mathrm{s.e.}(\hat{\bar{\theta}}) \;=\; \sum_{\kappa=1}^{k^*} w_\kappa \sqrt{\widehat{\mathrm{Var}}(\hat{\theta}_\kappa) - (\hat{\theta}_\kappa - \hat{\bar{\theta}})^2}\,, \tag{3}$$

where the first term under the square root reflects sampling uncertainty and the second one uncertainty due to model selection; see also Burnham and Anderson (2002) for an insightful discussion on assumptions and correct use of approximate standard errors in model averaging.

2.2 Model Averaging in Factor Analysis

In the forthcoming analysis we apply maximum likelihood factor analysis and consider the model

$$X' \;=\; \Gamma^{(k)}F^{(k)} + U\,, \tag{4}$$

where X is a $n \times p$ matrix of data, $\Gamma^{(k)}$ is a $p \times k$ matrix of loadings, F is a $k \times n$ matrix consisting of k factors and U is the $p \times n$ matrix of stochastic errors, $k < p$. We assume the matrices F, U and X to be multivariate normal, with expectation 0 and corresponding covariance matrices I, $\Psi = \mathrm{diag}(\Psi_1^2, \ldots, \Psi_p^2)'$ and $\Sigma = \Gamma^{(k)}\Gamma^{(k)\prime} + \Psi$. Typically, the Ψ_i^2, $i = 1, \ldots, p$, are termed as 'uniquenesses'.

In the context of factor analysis, the model selection problem relates to the appropriate choice for the number of latent factors, whereas the frequentist model averaging scheme compromises between different models that contain different numbers of factors. We consider the AIC criterion (Akaike (1973), Akaike (1987))

$$AIC = -2\mathcal{L}(\hat{\Gamma}^{(k)}, \hat{\Psi}) + 2K \tag{5}$$

where $\mathcal{L}(\cdot)$ is the likelihood function, K is the number of parameters and $\hat{\Gamma}^{(k)}$ and $\hat{\Psi}$ are the corresponding ML-estimates. The model M_κ that minimizes the AIC over $\mathcal{M}$ yields the traditional maximum likelihood FMS estimators $\hat{\Gamma}^{(k)}_\kappa$ and $\hat{\Psi}_\kappa$. To incorporate model selection uncertainty one may use the compromise estimator (1) of Γ and Ψ, that is

$$\hat{\bar{\Gamma}} = \sum_{\kappa=1}^{k^*} w_\kappa \, \hat{\Gamma}^{(k)}_\kappa \qquad \text{and} \qquad \hat{\bar{\Psi}} = \sum_{\kappa=1}^{k^*} w_\kappa \, \hat{\Psi}^{(k)}_\kappa \,, \tag{6}$$

where k^* is the number of potential models and w_κ is given by (2). The application of (6) guarantees the incorporation of model selection uncertainty, but only makes sense if we use factor loadings based on the same rotation principle, for instance a varimax rotation.

It is worthwile to note, that a current working paper of Dunson (2009) highlights the chances of model averaging in factor analysis from the Bayesian point of view.

3 Analysis

The complete data with the results of the olympic decathlon is presented in Table 1 and consists of all 30 athletes who finished the competition. Two athletes, Eugene Martineau from the Netherlands and Victor Covalenco from Moldavia, have error trials in the shot-put and pole-vault event, respectively. These two values may be treated as missing and since it is well-known that the ignorance of missing observations can lead to biased estimates and may even lead to the choice of inappropriate models (see, for instance, Schomaker et al. (2010) and Little and Rubin (2002)) we conduct an imputation for these two values based on a k-nearest-neighbor methodology: The idea of this procedure is simple: based on the Euclidian distance one chooses k rows that are nearest to the row that contains missing values; these k rows must not contain any missing observation. The missing values in the row under consideration are then replaced by the average of the observations in these neighboring k rows.

DOI: 10.2202/1559-0410.1249

		100m (sec)	LJ (m)	SP (m)	HJ (m)	400m (sec)	110mh (sec)	Dis (m)	PV (m)	Jav (m)	1500m (sec)
Roman Sebrle	(CZE)	10.85	7.84	16.36	2.12	48.36	14.05	48.72	5.00	70.52	280.01
Bryan Clay	(USA)	10.44	7.96	15.23	2.06	49.19	14.13	50.11	4.90	69.71	282.00
Dmitriy Karpov	(KAZ)	10.50	7.81	15.93	2.09	46.81	13.97	51.65	4.60	55.54	278.11
Dean Macey	(GBR)	10.89	7.47	15.73	2.15	48.97	14.56	48.34	4.40	58.46	265.42
Chiel Warners	(NED)	10.62	7.74	14.48	1.97	47.97	14.01	43.73	4.90	55.39	278.05
Attila Zsivoczky	(HUN)	10.91	7.14	15.31	2.12	49.40	14.95	45.62	4.70	63.45	269.54
Laurent Hernu	(FRA)	10.97	7.19	14.65	2.03	48.73	14.25	44.72	4.80	57.76	264.35
Erki Nool	(EST)	10.80	7.53	14.26	1.88	48.81	14.80	42.05	5.40	61.33	276.33
Claston Bernard	(JAM)	10.69	7.48	14.80	2.12	49.13	14.17	44.75	4.40	55.27	276.31
Roland Schwarzl	(AUT)	10.98	7.49	14.01	1.94	49.76	14.25	42.43	5.10	56.32	273.56
Aleksandr Pogorelov	(RUS)	10.95	7.31	15.10	2.06	50.79	14.21	44.60	5.00	53.45	287.63
Florian Schönbeck	(GER)	10.90	7.30	14.77	1.88	50.30	14.34	44.41	5.00	60.89	278.82
Romain Barras	(FRA)	11.14	6.99	14.91	1.94	49.41	14.37	44.83	4.60	64.55	267.09
Marice Smith	(JAM)	10.85	6.81	15.24	1.91	49.27	14.01	49.02	4.20	61.52	272.74
Nikolay Averyanov	(RUS)	10.55	7.34	14.44	1.94	49.72	14.39	39.88	4.80	54.51	271.02
⋮		⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table 1: Results of the Olympic Decathlon in Athens, 23.8./24.8.2004 (Part I)

		100m (sec)	LJ (m)	SP (m)	HJ (m)	400m (sec)	110mh (sec)	Dis (m)	PV (m)	Jav (m)	1500m (sec)
⋮		⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Jaako Ojaniemi	(FIN)	10.68	7.50	14.97	1.94	49.12	15.01	40.35	4.60	59.26	275.71
Vitaliy Smirnov	(UZB)	10.89	7.07	13.88	1.94	49.11	14.77	42.47	4.70	60.88	263.31
Haifeng Qi	(CHN)	11.06	7.34	13.55	1.97	49.65	14.78	45.13	4.50	60.79	272.63
Stefan Drews	(GER)	10.87	7.38	13.07	1.88	48.51	14.01	40.11	5.00	51.53	274.21
Aleksandr Parkhomenko	(BLR)	11.14	6.61	15.69	2.03	51.04	14.88	41.90	4.80	65.82	277.94
Paul Terek	(USA)	10.92	6.94	15.15	1.94	49.56	15.12	45.62	5.30	50.62	290.36
David Gomez	(ESP)	11.08	7.26	14.57	1.85	48.61	14.41	40.95	4.40	60.71	269.70
Indrek Turi	(EST)	11.08	6.91	13.62	2.03	51.67	14.26	39.83	4.80	59.34	290.01
Santiago Lorenzo	(ARG)	11.10	7.03	13.22	1.85	49.34	15.38	40.22	4.50	58.36	263.08
Janis Karlivans	(LAT)	11.33	7.26	13.30	1.97	50.54	14.98	43.34	4.50	52.92	278.67
Prodromos Korkizoglou	(GRE)	10.86	7.07	14.81	1.94	51.16	14.96	46.07	4.70	53.05	317.00
Hans Olav Uldal	(NOR)	11.23	6.99	13.53	1.85	50.95	15.09	43.01	4.50	60.00	281.70
Paolo Casarsa	(ITA)	11.36	6.68	14.92	1.94	53.20	15.39	48.66	4.40	58.62	296.12
Eugene Martineau	(NED)	10.99	6.84	–	2.00	49.10	15.02	40.00	4.80	63.62	271.79
Victor Covalenco	(MDA)	11.28	7.20	13.04	1.85	51.82	15.80	38.19	–	53.46	263.81

Table 1: Results of the Olympic Decathlon in Athens, 23.8./24.8.2004 (Part II)

DOI: 10.2202/1559-0410.1249

Due to the small sample size we find this procedure preferable to parametric imputation approaches, such as an EM-algorithm based imputation. Using a rule of thumb for small data, we take $k = 2$ and obtain imputed values of 14.41 meter for the shot-put result of Eugene Martineau and 4.65 meter for the pole-vault result of Victor Covalenco. The new, updated data set that contains the imputed values is used for the forthcoming analysis.

Based on the multivariate description of decathlon data from Cox and Dunn (2002), individual considerations, and an explorative look at the data we consider five competing models to capture the structure of the data: these are all factor analytical models $X' = \Gamma^{(k)} F^{(k)} + U$, $k = 1, \ldots, 5$, where X is the 30×10 matrix of the decathlon results, $\Gamma^{(k)}$ is the $10 \times k$ matrix of loadings, F is a $k \times 30$ matrix consisting of k factors and U is the 10×30 matrix of stochastic errors. We use the statistical software package R (R Development Core Team (2008, Version 2.8.1)) to perform the analysis[1] and calculate the AIC[2] and the corresponding AIC weights for each of the five models[3]. The results are presented in Table 2.

	M_1	M_2	M_3	M_4	M_5
AIC	-12.85	-24.90	-25.77	-18.32	-8.50
Weights	0.00	0.39	0.60	0.01	0.00

Table 2: AIC and AIC weights for the five competing models

It is clear that the 3-factor model M_3 has the minimum AIC value (-25.77) and its maximum likelihood parameter estimates $\hat{\Gamma}_3^{(3)}$ and $\hat{\Psi}_3$ would be chosen from the corresponding frequentist model selection procedure. However, the 2-factor model has only a slightly larger AIC value (-24.90) and a look at the corresponding AIC weights, 0.39 for M_2 and 0.60 for M_3, confirms that both the 2- and the 3-factor model have a considerable appeal to describe the

[1]The exact results of any maximum likelihood factor analysis depend on the utilized optimization techniques and therefore may vary slightly depending on the software used by the statistical analyst. Especially when confronted with Heywood cases (where one or more uniquenesses – the diagonal elements of Ψ - are essentially zero) this can be observed regularly

[2]There exist small sample corrections for the AIC (Sugiura (1978), Hurvich and Tsai (1989)). Since they are not valid in the context of factor analysis we use the traditional non-corrected AIC

[3]Most statistical software packages do not provide Akaikes Information Criterion (AIC) in the context of factor analysis. However, via the relationship of the AIC to the χ^2-statistic (see Akaike (1987, p. 321)) it can be calculated easily using standard statistical software packages like R, S-Plus and others'

structure of the data appropriately. The factor analytical models that contain one, four or even five factors seem to be of no practical importance as can be clearly seen from Table 2.

The matrix of the estimated loadings (after a varimax rotation) and the corresponding estimated uniquenesses of the 2-, 3- and 4-factor model are presented in Tables 3, 4 and 5; all elements $\gamma_{ii} \in \Gamma$ that are greater 0.5 are underlined.

	Loadings Γ					Ψ_i^2
100 m	0.79	-0.31	0.00	0.00	0.00	0.28
Long Jump	-0.79	0.10	0.00	0.00	0.00	0.37
Shot Put	-0.16	0.94	0.00	0.00	0.00	0.09
High Jump	-0.24	0.64	0.00	0.00	0.00	0.53
400 m	0.79	-0.15	0.00	0.00	0.00	0.36
110 m Hurdles	0.63	-0.29	0.00	0.00	0.00	0.51
Discus	-0.16	0.72	0.00	0.00	0.00	0.46
Pole-Vault	-0.27	-0.03	0.00	0.00	0.00	0.93
Javeline	-0.01	0.41	0.00	0.00	0.00	0.83
1500 m	0.25	0.26	0.00	0.00	0.00	0.87

Table 3: Matrix of loadings and uniquenesses, 2-factor model

	Loadings Γ					Ψ_i^2
100 m	0.83	-0.27	-0.07	0.00	0.00	0.23
Long Jump	-0.78	0.10	-0.03	0.00	0.00	0.39
Shot Put	-0.19	0.91	0.08	0.00	0.00	0.14
High Jump	-0.23	0.66	-0.03	0.00	0.00	0.52
400 m	0.77	-0.17	0.38	0.00	0.00	0.23
110 m Hurdles	0.63	-0.29	0.00	0.00	0.00	0.52
Discus	-0.18	0.72	0.20	0.00	0.00	0.41
Pole-Vault	-0.32	-0.10	0.24	0.00	0.00	0.83
Javeline	0.04	0.46	-0.29	0.00	0.00	0.70
1500 m	0.12	0.16	0.98	0.00	0.00	0.00

Table 4: Matrix of loadings and uniquenesses, 3-factor model

A close look at the results of the 2-factor model M_2 yields the following evidence: All short distance races as well as long jump are loading high on the first factor. This may be interpreted as a speed, athletic and liveliness component of the decathlon contest. On the contrary, it is mainly shot-put, high jump and discus that are loading high on the second factor, which could be taken as a summary of strength-and-technique events. Of course, the javeline

	Loadings Γ					Ψ_i^2
100 m	0.82	-0.25	-0.08	0.24	0.00	0.20
Long Jump	-0.81	0.03	-0.02	-0.01	0.00	0.35
Shot Put	-0.21	0.96	0.18	-0.02	0.00	0.00
High Jump	-0.29	0.59	0.07	0.09	0.00	0.55
400 m	0.77	-0.19	0.37	0.04	0.00	0.24
110 m Hurdles	0.65	-0.22	-0.03	0.01	0.00	0.52
Discus	-0.34	0.59	0.37	0.62	0.00	0.00
Pole-Vault	-0.24	-0.05	0.17	-0.44	0.00	0.72
Javeline	0.00	0.47	-0.22	0.09	0.00	0.72
1500 m	0.12	0.05	0.98	-0.10	0.00	0.00

Table 5: Matrix of loadings and uniquenesses, 4-factor model

event also has a certain contribution to the second factor. But as we look at the uniquenesses, we find that most of its variance cannot explained by the two factors. Therefore such a kind of interpretation has to be handled with care. Both pole-vault and the 1500 m race are not loading high on any of the two factors which is reflected in the corresponding large uniquenesses.

The 3-factor model, which is slightly favored from the AIC, relates to similar statements. The first and the second factor may be interpreted as above. However, in regard to the third factor, one can see that it is only the 1500 m race that is loading high on it. A possible interpretation might be that the third factor reflects endurance, which is supported from the loading of the 400 m race (0.38). However, it is well known that results from the 1500 m race are different from the other nine events since it is the last one and each athlete knows about his possible success and is clearly affected by that[4]. Therefore, the third factor might also be seen as a 'last-event-effect'.

The 4-factor model is not well-supported by the Akaike criterion. A look at the estimated loadings in Table 5 confirms that with regards to content. The first three factors are basically the same as in the 3-factor model. Moreover, it is mainly the discus event that is loading high on the fourth factor, which is – no doubt – dispensable and unnecessary.

However, based on the thoughts of section 2 we are interested in incorporating model selection uncertainty in our final estimations of Γ and Ψ. We therefore calculate the compromise estimators (6) as illustrated in Table 6.

A close look at the table still supports the assumption of three strong factors in decathlon competitions - despite the remarkable evidence of a 2-factor

[4]A famous example is Bryan Clay from the United States, who dominated the 2008 olympic decathlon in Beijing in the first nine events. Being sure to win, he came in last in the 1500 m race

	Loadings $\bar{\Gamma}$					$\bar{\Psi}_i^2$
100 m	0.81	-0.29	-0.05	0.00	0.00	0.25
Long Jump	-0.78	0.10	-0.02	0.00	0.00	0.38
Shot Put	-0.18	0.92	0.05	0.00	0.00	0.12
High Jump	-0.23	0.65	-0.01	0.00	0.00	0.52
400 m	0.78	-0.16	0.24	0.00	0.00	0.28
110 m Hurdles	0.63	-0.29	0.00	0.00	0.00	0.52
Discus	-0.17	0.72	0.13	0.01	0.00	0.42
Pole-Vault	-0.30	-0.07	0.15	-0.01	0.00	0.87
Javeline	0.02	0.44	-0.18	0.00	0.00	0.75
1500 m	0.17	0.19	0.60	0.00	0.00	0.34

Table 6: Weighted matrix of loadings and uniquenesses that incorporate model selection uncertainty

model. Hence, the final results are similar to the model selection estimator from the three factor model:

- The first factor contains essentially 100 m, 400 m, 110 m hurdles and long jump. This may reflect the speed-and-athletic component of the decathlon contest.
- The second factor contains essentially shot-put, high jump and discus. This may reflect the strength-and-technique component of the decathlon contest.
- The third factor contains essentially the 1500 m race. This may reflect endurance as well as the special status of the last event.
- Both javeline and pole-vault do not load high on any of these three factors and are high in their uniquenesses. One possible explanation relates to the claim that these events are the most sophisticated and demanding within a decathlon contest.

Effectively, the support of three distinctive factors lies in correspondence with the final results of the cluster analysis of Cox and Dunn (2002)[5]; see also Table 7. However, although the first factor of our analysis complies with the first cluster of Cox and Dunn (2002), the second and third do not. The main difference is the pattern of the third cluster, which Cox and Dunn (2002) claim to contain both the 1500 m race and the high jump. Though somewhat supported by the data, it is questionable whether such a cluster makes sense.

[5]For further insightful thoughts regarding the results from Cox and Dunn (2002) see Woolf et al. (2007)

DOI: 10.2202/1559-0410.1249

Of course, cluster analysis is a different methodology in multivariate statistics and obviously has to account for all ten events. Therefore both the javeline and pole-vault event are contained in cluster 2, which makes more sense than to be contained in any of the other 2 clusters and this is not in direct contradiction to our results.

Cluster 1	110 m	400 m	long jump	110 m hurdles
Cluster 2	shot-put	discus	javeline	pole-vault
Cluster 3	high jump	1500 m		

Table 7: Final results of the cluster analysis from Cox and Dunn (2002, p. 181)

4 Discussion

The aim of this article, as mentioned above, is to explore the dimension of decathlon contests by example of the olympic data from Athens 2004. Our approach is based on maximum likelihood factor analysis and explicitly accounts for the uncertainty associated with the determination of the number of latent factors by means of modern frequentist model averaging techniques.

It is found that the data supports three distinctive factors. One possible interpretation relates to a 'speed-and-athletic' factor, a 'strength-and-technique' factor and a 'last-event' factor. Both the javeline and the pole-vault event do not contribute much to these factors. Of course, all kind of metaphorical interpretations in factor analysis have to be handled with care.

It would be interesting to think of other possible extensions for grouping decathlon disciplines. It is well-known that a restriction of conventional factor analytical approaches is that no covariates, such as age and origin of athletes, can be included into the model and the estimation process. Hence, latent variable models, as for example in Fahrmeir and Raach (2007), offer an interesting opportunity to broaden our knowledge about the nature of decathlon competitions and revise existing results.

Despite these future challenges, our study has offered some interesting insights into some very practical questions on the dimension of decathlon contests and the treatment of modeling uncertainty that certainly warrant further studies.

Bibliography

Akaike, H. (1973) *Information theory and an extension of the maximum likelihood principle*. Proceeding of the Second International Symposiumon Information Theory Budapest 267–281

Akaike, H. (1987) *Factor analysis and AIC*. Psychometrika 52:317–332

Brown, B.H. (1946) *The scoring of athletic contests*. The Scientific Monthly 62:233–237

Buckland, S.T., Burnham, K.P., Augustin, N.H. (1997) *Model selection: an integral part of inference*. Biometrics 53:603–618

Burnham, K., Anderson, D. (2002) *Model selection and multimodel inference. A practical information-theoretic approach*. Springer, New York

Chang, C-L., Tsai, C-H., Chen, L. (2003) *Applying grey relational analysis to the decathlon evaluation model*. International Journal of the Computer, the Internet and Management 11:54–62

Claeskens, G., Hjort, N.L. (2003) *The focused information criterion [with discussion]*. Journal of the American Statistical Association 98:900–916

Cox, T.F., Dunn, R.T. (2002) *An analysis of decathlon data*. The Statistician 51:179–187

Dawkins, B.P., Andreae, P.M., O'Connor, P.M. (1994) *Analysis of olympic hepathlon data*. Journal of the American Statistical Association 89:1100–1106

Dunson, D.B. (2009) *Efficient bayesian model averaging in factor analysis*. Working Paper, `http://ftp.stat.duke.edu/WorkingPapers/06-03.pdf`

Fahrmeir, L., Raach, A. (2007) *A Bayesian semiparametric latent variable model for mixed responses*. Psychometrika 72:327–346

Hansen, B.E. (2007) *Least squares model averaging*. Econometrica 75:1175–1189.

Hoeting, J., Madigan, D., Raftery, E., Volinsky, C. (1999) *Bayesian model averaging: a tutorial*. Statistical Science 14:382–417

Hjort, L., Claeskens, G. (2003) *Frequentist model average estimators*. Journal of the American Statistical Association 98:879–945

DOI: 10.2202/1559-0410.1249

Hurvich, C.M., Tsai, C.L. (1989) *Regression and time series model selection in small samples.* Biometrika 76:297–307

Little R., Rubin D. (2002) *Statistical analysis with missing data.* Wiley, New York

R Development Core Team (2008) *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org

Schomaker, M., Wan, A.T.K., Heumann, C. (2010) *Frequentist model averaging with missing observations.* Computational Statistics & Data Analysis 54:3336–3347

Stone, M. (1974) *Cross-validatory choice and assessment of statistical predictions.* Journal of the Royal Statistical Society B 36:111–147

Sugiura, N. (1978) *Further analysis of the data by Akaike's Information Criterion and the finite corrections.* Communications in Statistics, Theory and Methods A7:13–26

Wang, H., Zhang, X., Zou, G. (2009) *Frequentist model averaging: a review.* Journal of Systems Science and Complexity 22:732–748

Woolf, A., Ansley, L., Bidgood, P. (2007) *Grouping of decathlon disciplines.* Journal of Quantitative Analysis of Sports Vol. 3 : Iss. 4, Article 5.