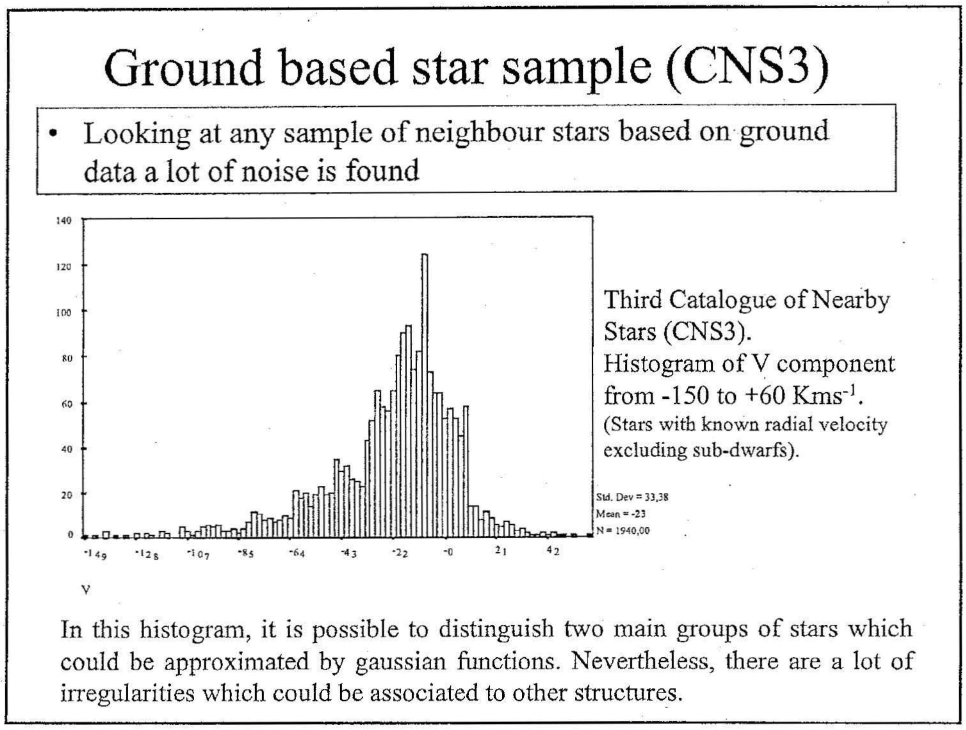
Non-informative filtering of the local stellar velocity distribution in order to obtain two nearly-pure statistical populations

R. Cubarsi, S. Alcobé, and M. A. Català-Poch
Dept. Matematica Aplicada i Telematica
Universitat Politecnica de Catalunya
Jordi Girona 1-3, E08034 Barcelona, Spain

Abstract

Determination of stellar population constants from the statistics of a global sample is extremely sensitive to the noise introduced by few non-typical stars. The local velocity distibution can be better determined from a neighbour stars calatogue including Hipparcos data and by applying a non-informative filtering to the global sample. In order to produce an optimal segregation into two population components an auxiliar parameter P, depending on the distribution variables, such as the absolute velocity referred to the slower subcentroid, is introduced. The parameter must induce an ordered incorporation of stars to the population components, in the sense that the greater the P value, the greater the number of stars in each component. Then a subsample S(P) is drawn from the global one. Depending on this parameter the statistical entropy H(P) of the mixture probability is computed, and the optimal subsample S(P) is selected in order to maximize H(P). The method is applied to segregate the local velocity distribution into two main trivariate normal components (Cubarsi et. al) where a total accordance between the best fit from chi squared test and the maximum population entropy H(P) is produced. Furthermore the method can be used recursively in order to segregate a global sample in more than two populations.

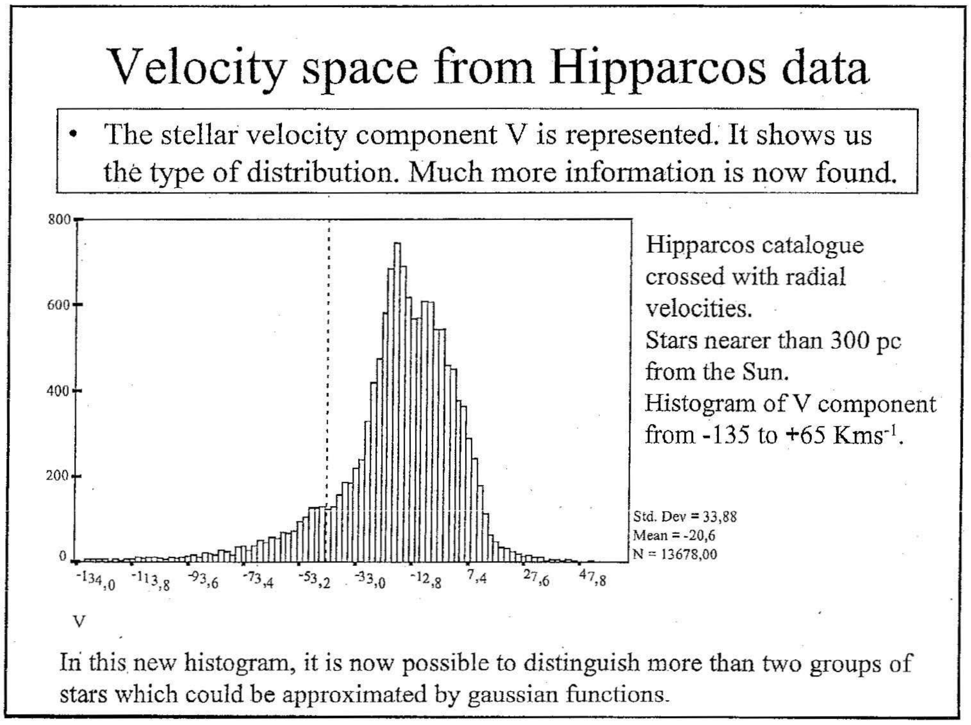


Ground based star sample (CNS3)

- Looking at any sample of neighbour stars based on ground data a lot of noise is found

Third Catalogue of Nearby Stars (CNS3).
Histogram of V component from -150 to +60 Kms^{-1}.
(Stars with known radial velocity excluding sub-dwarfs).

In this histogram, it is possible to distinguish two main groups of stars which could be approximated by gaussian functions. Nevertheless, there are a lot of irregularities which could be associated to other structures.



Velocity space from Hipparcos data

- The stellar velocity component V is represented. It shows us the type of distribution. Much more information is now found.

Hipparcos catalogue crossed with radial velocities.
Stars nearer than 300 pc from the Sun.
Histogram of V component from -135 to +65 Kms^{-1}.

In this new histogram, it is now possible to distinguish more than two groups of stars which could be approximated by gaussian functions.

Numerical method. Generalities.

- A subsample S(P) is drawn from the overall stellar velocity sample, depending on a parameter P, which describes a specific property.
- The number of stars of the subsample is noted #S(P) and we assume that a finite number of c populations are present in this subsample.
- The partition $A=\{A_1, A_2,..., A_c\}$ represents the c population components
- The mixture density function for the stellar velocity **v** is given by: $f(v)=\sum_{i=1}^{c} p(A_i)g_i(v|o_i)$ (1) where $p(A_i)= n^{(i)}$ is the mixing proportion of the population A_i, so that $\sum_{i=1}^{c} p(A_i)=1$ and $g_i(v|o_i)$ is the partial density function for the i-th population, depending on a characteristic parameter vector $o_i=\{v^{(i)},M_2^{(i)}$ that is, means and 2nd order central moments.
- The partial densities are assumed to be normal trivariant functions for all the populations. Thus $g_1=g_2=.....=g_c\equiv g$
- The algorithm (Cubarsi 1992) used to compute the parameter vectors involved in Eq. (1) and the mixing proportion provides a density function in the form $f(v)=\sum_{i=1}^{c} n^{(i)}(\mathbf{a}_p)g(v|h_i(\mathbf{a}_p))$ (2) where $\mathbf{a}_p=\mathbf{a}_p(o_1, o_2, ...o_c)$ is a reduced vector of auxiliar parameters and the functions $h(\mathbf{a}_p)$ are different for each component. The sub-index p is written in order to emphasize that the new parameter vector $\mathbf{a}_p$ depends on the parameter P of the selected subsample S(P).

* The purpose of the present work is to obtain a segregation for the subsample S(P), so that the parameters describing each population are representative of the maximum number of stars. Thus, the overall sample is filtered in order to exclude stars with an information too specific and different from the characteristic population values. That is: data too much informative is considered noise.

* This purpose is accomplished with two steps:

1) Maximization of the entropy H(A) of the partition A, as a function of the parameter P. Thus, we shall write the entropy as $H_c(P)$ for c components.

2) To choose the property closest to ideal parameter P so that the maximization of $H_c(P)$ is possible.

Entropy

- The function $z(t)=\begin{cases}-t\ln t, & 0<t\leq 1\\ 0 & ,\ t=0\end{cases}$ is a non-negative, continous and strictly concave function, which is used in order to evaluate the entropy H(A) of the partition A. $\mathrm{H}(\mathrm{A})=\sum_{i=1}^{c} z(p(\mathrm{A}_i))$
- Using the previous notation, we can write $\mathrm{H}_c(P)=\sum_{i=1}^{c}-n^{(i)}(\mathbf{a}_p)\ln(n^{(i)}(\mathbf{a}_p))$ (3)
- This can be interpreted as the expected value of the incertainty $I(A_i)=-\ln(p(A_i))$ (Koch 1990) and entropy variations are interpreted as uncertainty variations for the mixture parameters.
- The greater the entropy, the less the information of the population parameters. Let us remember that the gaussian populations are also the less informative density functions depending on the parameters mean and second central moments.
- We want then to determine the values $\mathbf{a}_p$ and $n^{(i)}(\mathbf{a}_p)$ providing the maximum value of $H_c(P)$, depending on the parameter P. Since the model of superposition we use is a two-component model (c=2), we shall study this particular case, although it can be generalized.
- Hence, $H_2(P)=-n'(\mathbf{a}_p)\ln(n'(\mathbf{a}_p))-(1-n'(\mathbf{a}_p))\ln(1-n'(\mathbf{a}_p))$ (4)

  We asume that the first population A_1 is the prominent population $n'\geq\frac{1}{2}$
- The function $H_2(P)$, in Eq. (4), for $\frac{1}{2}<n'<1$ is a positive, decreasing and differentiable function satisfying: $0<\mathrm{H}_2(n')<1,\ \frac{d\mathrm{H}_2}{dn'}<0;\qquad \frac{1}{2}<n'<1$ (5)

Ideal selecting parameter

The parameter P used to select a two-component subsample S(P) must satisfy the two following conditions:

- The number of stars of S(P) increases with P, without loosing any star. That is $P_1<P_2\Rightarrow S(P_1)\subseteq S(P_2)\Rightarrow \#S(P_1)\leq \#S(P_2)$
- P induces an ordered incorporation of stars to the subsample S(P), so that stars of population A_1 are included first and, when such population is completed, stars of population A_2 are then included.

Hence, $P_1<P_2\Rightarrow n'(P_1)\geq n'(P_2)$ (6)

Moreover $n'(P_1)=n'(P_2)\Rightarrow n'=1$ (7a)

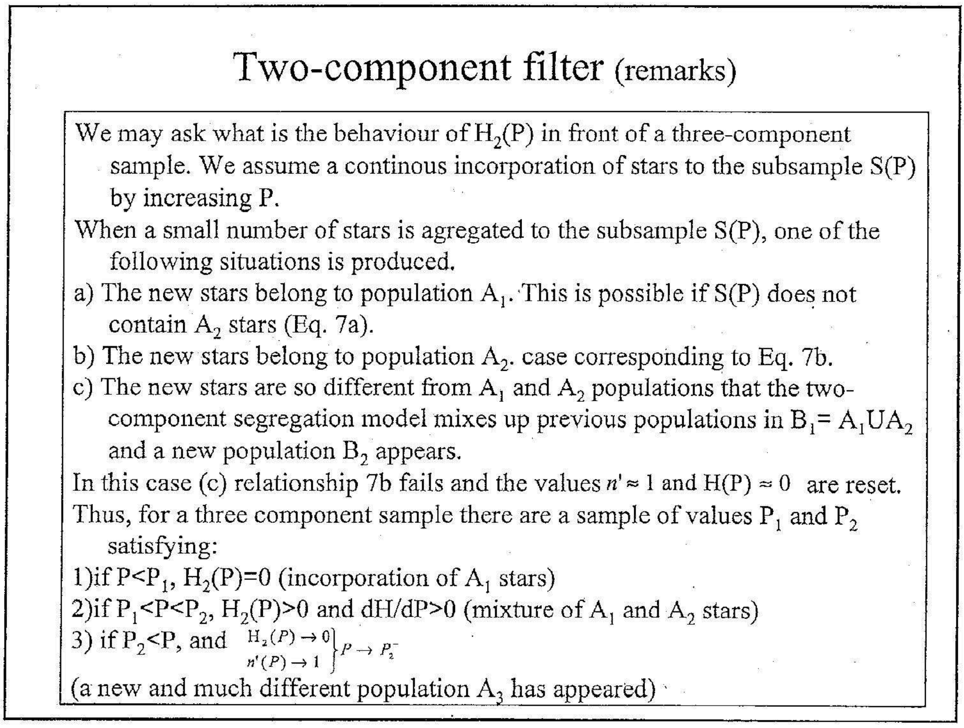
$n'(P_1)>n'(P_2)\Rightarrow n'<1$ (7b)

We asume that n'(P) is a non-increasing continous and differentiable function of P.

Then, the second condition is equivalent to $\frac{dn'}{dP}\leq 0$

Under previous hypothesis and Eq. (5), the entropy $H_2(P)$ is non-decreasing

$\frac{dH_2}{dP}=\frac{dH_2}{dn'}\frac{dn'}{dP}\geq 0,\quad \frac{1}{2}<n'<1;$ thus $P_1<P_2\Rightarrow \mathrm{H}_2(P_1)\leq \mathrm{H}_2(P_2)$



Two-component filter (remarks)

We may ask what is the behaviour of $H_2(P)$ in front of a three-component sample. We assume a continous incorporation of stars to the subsample S(P) by increasing P.

When a small number of stars is agregated to the subsample S(P), one of the following situations is produced.

a) The new stars belong to population A_1. This is possible if S(P) does not contain A_2 stars (Eq. 7a).

b) The new stars belong to population A_2. case corresponding to Eq. 7b.

c) The new stars are so different from A_1 and A_2 populations that the two-component segregation model mixes up previous populations in $B_1 = A_1UA_2$ and a new population B_2 appears.

In this case (c) relationship 7b fails and the values $n' \approx 1$ and $H(P) \approx 0$ are reset.

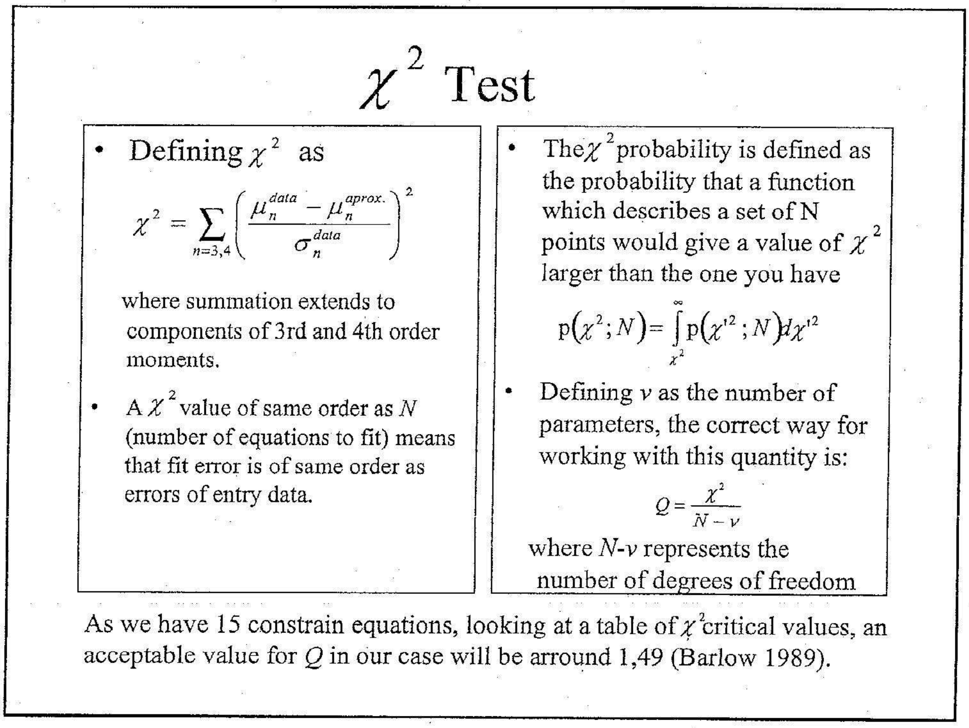
Thus, for a three component sample there are a sample of values P_1 and P_2 satisfying:

1)if $P<P_1$, $H_2(P)=0$ (incorporation of A_1 stars)

2)if $P_1<P<P_2$, $H_2(P)>0$ and $dH/dP>0$ (mixture of A_1 and A_2 stars)

3) if $P_2<P$, and $\left.\begin{array}{l} H_2(P) \to 0 \\ n'(P) \to 1 \end{array}\right\} P \to P_2^-$

(a new and much different population A_3 has appeared)



χ^2 Test

- Defining χ^2 as

$$\chi^2 = \sum_{n=3,4} \left(\frac{\mu_n^{data} - \mu_n^{aprox.}}{\sigma_n^{data}} \right)^2$$

where summation extends to components of 3rd and 4th order moments.

- A χ^2 value of same order as N (number of equations to fit) means that fit error is of same order as errors of entry data.

- The χ^2 probability is defined as the probability that a function which describes a set of N points would give a value of χ^2 larger than the one you have

$$p(\chi^2;N) = \int_{\chi^2}^{\infty} p(\chi'^2;N) d\chi'^2$$

- Defining ν as the number of parameters, the correct way for working with this quantity is:

$$Q = \frac{\chi^2}{N-\nu}$$

where N-ν represents the number of degrees of freedom

As we have 15 constrain equations, looking at a table of χ^2 critical values, an acceptable value for Q in our case will be arround 1,49 (Barlow 1989).

Hipparcos Catalogue. Central moments.

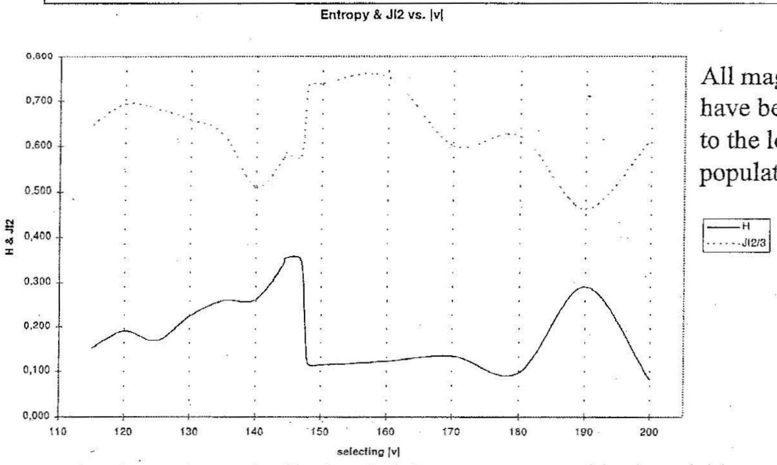
		Value	Error
100	(U_0)	-10.35	.27
010	(V_0)	-17.05	.18
001	(W_0)	-7.24	.13
ORDER2			
200		944.63	14.77
110		99.71	6.83
020		424.50	7.62
101		-13.43	5.63
011		10.83	4.03
002		240.36	5.28
ORDER3			
300		2497.22	983.54
210		-6279.45	386.69
120		-1150.26	298.98
030		-8662.00	448.13
201		-122.28	328.78
111		119.91	170.32
021		-299.52	194.91
102		50.37	226.55
012		-2265.19	185.45
003		30.14	325.81

ORDER4		
400	3796357.00	132175.24
310	183822.64	36654.71
220	631699.04	20993.54
130	118875.61	21987.01
040	953476.55	43783.72
301	-74239.09	36496.32
211	7438.00	11370.37
121	-11807.29	9077.28
031	23912.47	15755.64
202	422382.55	18225.00
112	18538.18	7942.35
022	216543.68	11228.97
103	-30101.79	15112.02
013	-3100.13	13918.00
004	428653.51	31433.98

Sample selected from HIPPARCOS crossed with radial velocities (13,678 stars). Selection is done according to |v| < 131 Kms^{-1} (13,315 stars remain)

Sample filtering (1)

Maximum entropy and best χ^2 fit (CNS3 sample)

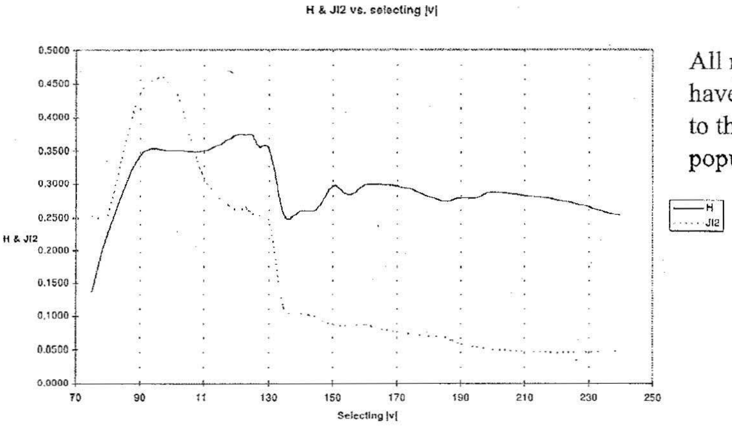


All magnitudes have been referred to the lower population mean.

After applying the numerical method, the maximum entropy for the mixture proportions is found for the selected sample. The correct selecting parameter is actually χ^2associated to the maximum entropy.

Sample filtering (2)

Maximum entropy and best χ^2 fit (HIPPARCOS sample)



All magnitudes have been referred to the lower population mean.

With the numerical method, the maximum entropy for the mixture proportions shows new stuctures which could be associated with stellar groups. (Not so clearly seen in the graphic due to scale effects).

Results for a local stellar sample (CNS3)

Comp. a	M11	M22	M33	M12	M13	M23	V1	V2	V3	%
value	1185	399	309	108	-73	6	-1.5	-1.6	-0.4	0.89
error	111	139	41	87	44	24	0.9	0.8	0.5	0.02

Comp. b	M11	M22	M33	M12	M13	M23	V1	V2	V3	%
value	3315	1579	1423	-28	432	-101	-23.3	-39.6	-0.2	0.11
error	751	1050	271	644	285	127	2.6	4.3	0.5	0.02

Component a (young disk): $\sigma_R : \sigma_\Phi : \sigma_z$ = 34±2 : 20±3 : 18±1

Component b (old disk): $\sigma_R : \sigma_\Phi : \sigma_z$ = 58±7 : 40±13 : 38±4

These are the results found after applying the superposition method which gives two main components for the local star sample. Looking at these kinematic parameters, we are able to identify the found components with young and old disk stars. Component b is compatible with an axisymmetric dynamic model.

Comp. a and b are the gaussian components which are superposed to get the global sample. Mij are the moments components. Vi are U,V,W velocities.

Results for HIPPARCOS sample

Comp. a	**M11**	**M22**	**M33**	**M12**	**M13**	**M23**	**V1**	**V2**	**V3**	**%**
value	660	240	159	53	-2.7	4.4	-0.4	-0.2	0.0	0.88
error	36	36	9	27	10	5	0.3	0.3	0.1	0.01

Comp. b	**M11**	**M22**	**M33**	**M12**	**M13**	**M23**	**V1**	**V2**	**V3**	**%**
value	2433	808	814	-239	-103	66	-23.8	-33.8	-0.5	0.12
error	230	244	53	181	58	26	1.0	1.3	0.1	0.01

Component a: $\sigma_R : \sigma_\Phi : \sigma_z = 25\pm1 : 16\pm1 : 13\pm0.3$

Component b: $\sigma_R : \sigma_\Phi : \sigma_z = 49\pm2 : 28\pm4 : 28\pm1$

The found results after applying the superposition method show much smaller errors. These kinematic parameters, allow us to identify the observed components with a subset of the CNS3 population **a** or (perhaps) with moving groups.

Now, component b shows not negligible vertex deviation. Then, an axisymmetric model is not enough to explain its dynamic behaviour.

Comp. a and b are the gaussian components which are superposed to get the global sample. Mij are the moments components. Vi are U,V,W velocities.

References

- Alcobé, S., Cubarsi R., Catalá-Poch, M. A., 1995, Structure and Evolution of Stellar Systems, Proc. of the Int. Conference, 189, St. Petersburg
- Barlow, R., 1989, Statistics, a Guide to the Use of Statistical Methods in the Physical Sciences. Chichester: Wiley. Cubarsi, R. 1990, AJ, 99, 1558
- Cubarsi, R. 1992, AJ, 103, 1608
- Gliese W., Jahreiß, H., 1991, Third Catalogue of Nearby Stars, Astronomisches Rechen-Institut, Heidelberg
- Koch, K.R., 1990, Bayesian Inference with Geodetic Applications, Springer-Verlag, Berlín
- Papoulis, A., 1989, Probability, Random Variables and Stochastics Processes, McGraw Hill Co., Singapore
- Press,W.H.; Teukolsky, S.A.; Vetterling, W.T.; Flannery B.P.(1992) In Numerical Recipes in FORTRAN: The Art of Scientific Computing. Cambridge University Press, Cambridge, UK