Corso di Studio SM13 – CHIMICA

Introduzione alla chemiometria ed al disegno sperimentale (018CM)

10/11/2022 (lezione 3)

Pierluigi Barbieri

UNIVERSITÀ DEGLI STUDI DI TRIESTE

Dipartimento di Scienze Chimiche e Farmaceutiche

Gruppo di ricerca in Scienze Analitiche applicate alle InterazionI Uomo-Ambiente (ASHEI)

Michele Forina

http://www.sisnir.org/sisnir/download/fondamenta-per-la-chimica-analitica/category/3-fondamenta-per-la-chimica-analitica

Visualizzazione di statistiche uni- e bi-variate.

Visualizzazione di set di dati reali con diverse tecniche di rappresentazione grafica, analisi dei risultati tramite esplorazione visiva, tecniche di individuazione di dati anomali. Illustrazione del concetto di carta di controllo dei dati.

Analisi delle componenti principali e metodi fattoriali: aspetti teorici ed applicativi dell'analisi delle componenti principali e dei metodi fattoriali per la compressione ed interpretazione dell'informazione contenuta nei dati.

Visualizzazione di set di dati reali chimici, fisici e biologici

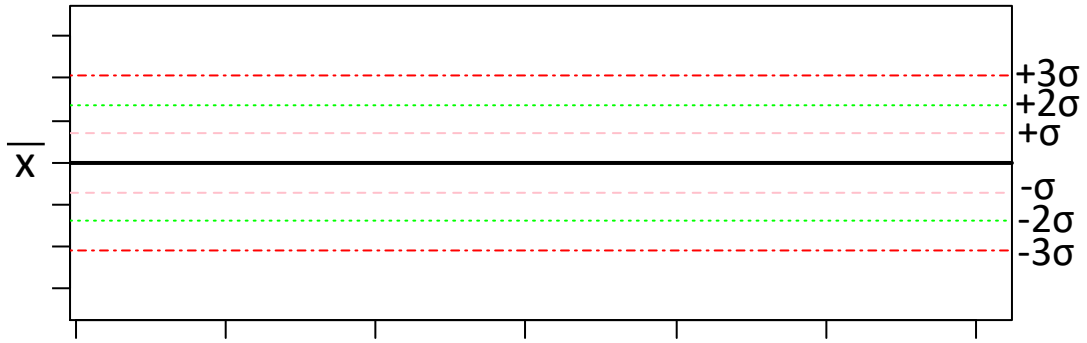
Mettiamo in rapporto visivo (su un grafico) i dati (anche multivariati) rilevati considerando valori di riferimento/norme

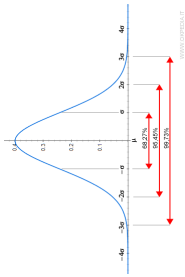
Il concetto di carta di controllo di una variabile

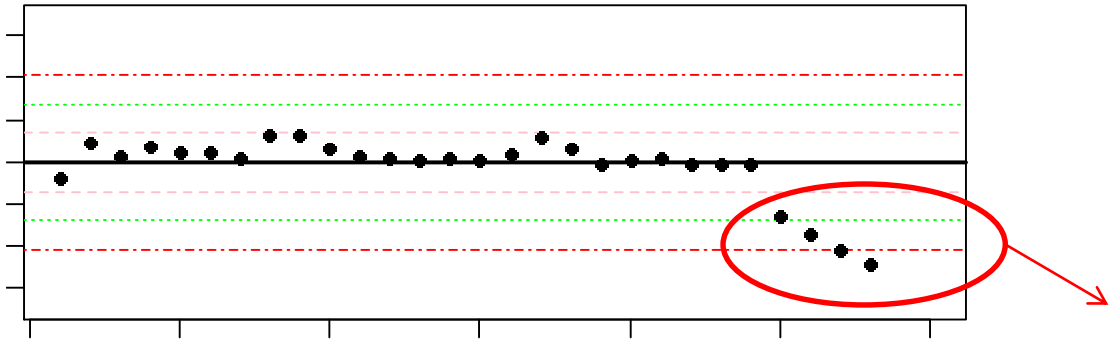
Una carta di controllo è uno strumento grafico per monitorare il/i parametri di un ***processo*** (anche *misure successive su campioni tratti da una popolazione definita*)

Carta di controllo di Shewhart

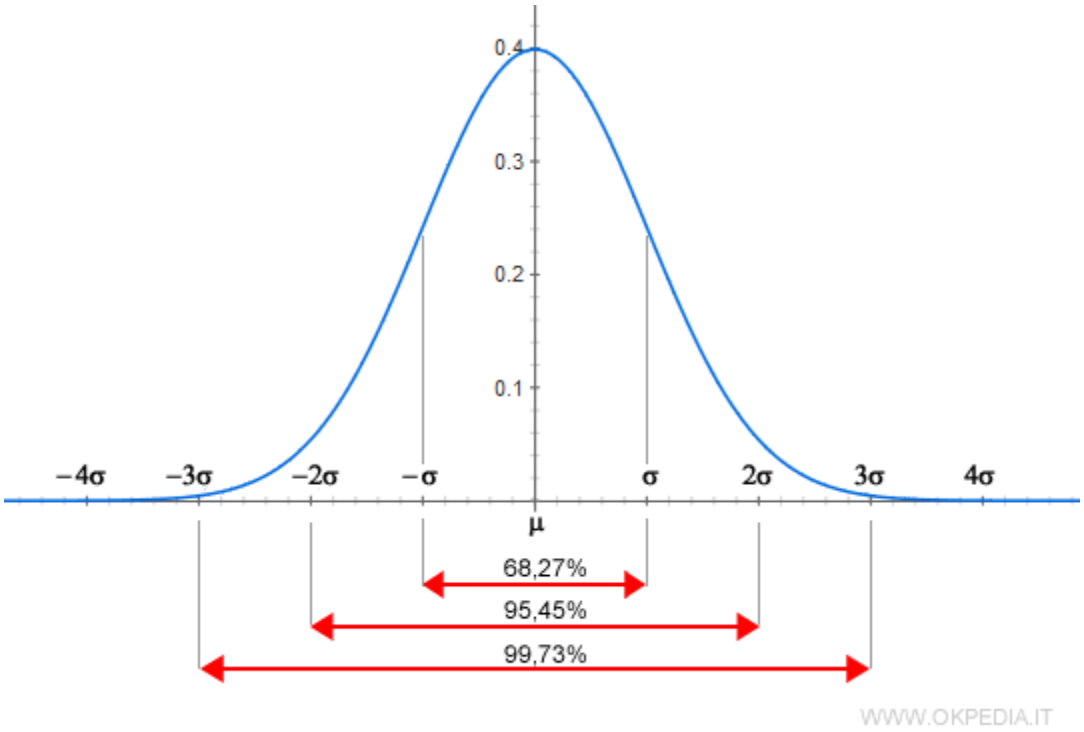
(utilizza la media e la deviazione standard per controllare eventuali derive nei valori di un parametro relativo ad un processo):





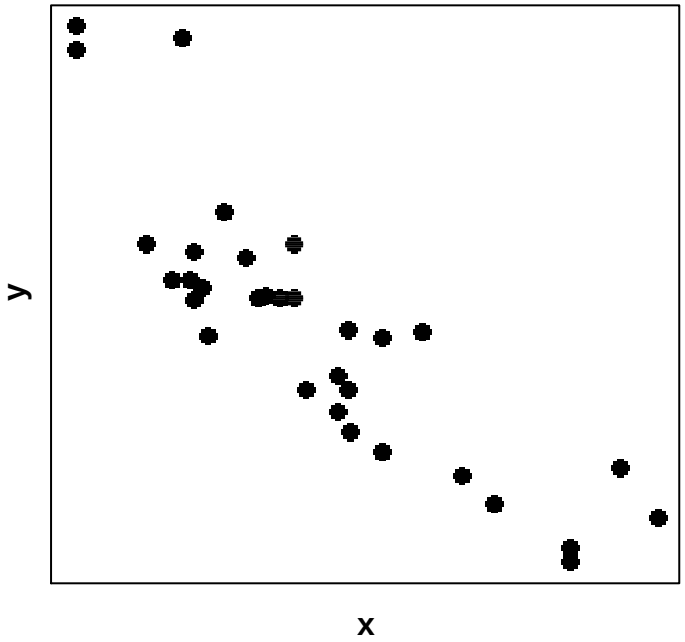


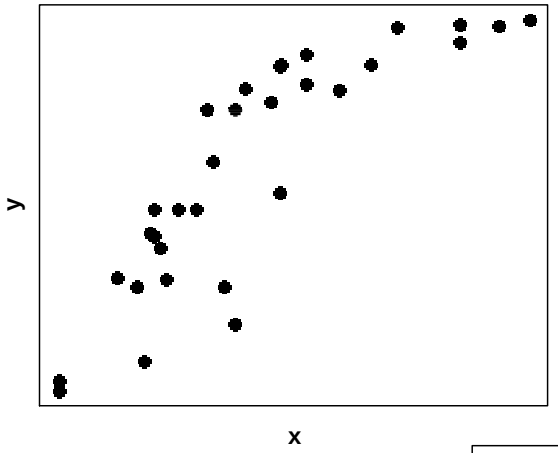
possibile deriva del sistema

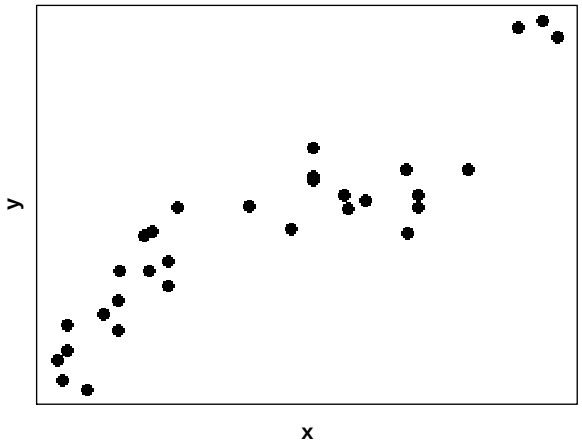

0.4
0.3
0.2
0.1
−4σ
−3σ
−2σ
−σ
σ
2σ
3σ
4σ
μ
68,27%
95,45%
99,73%
WWW.OKPEDIA.IT

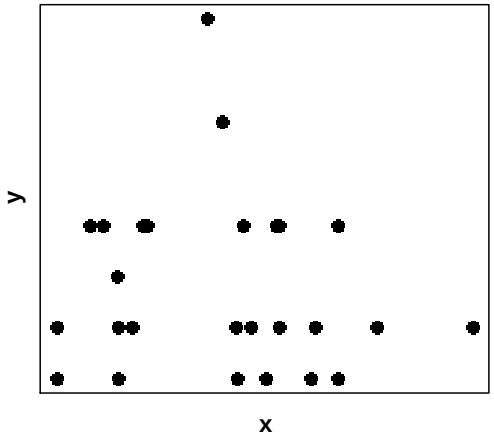
Analisi bivariata

E' finalizzata ad osservare il comportamento relativo di una variabile rispetto ad un'altra









Covarianza

E' un indicatore della variabilità "congiunta" di due variabili.

scarti di due variabili riferite alla medesima osservazione

$$\sigma_{x,y} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n}$$

Si può scrivere anche:

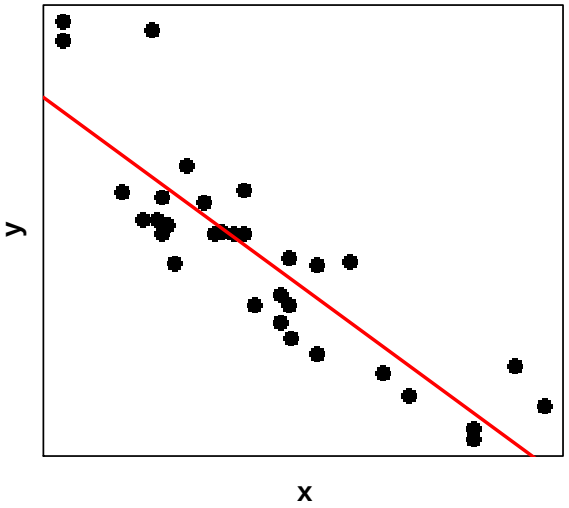
$\overline{x}$ $\overline{y}$

$$\sigma_{x,y} = \frac{1}{n}\sum_{i=1}^{n} x_i\, y_i \; - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right)$$

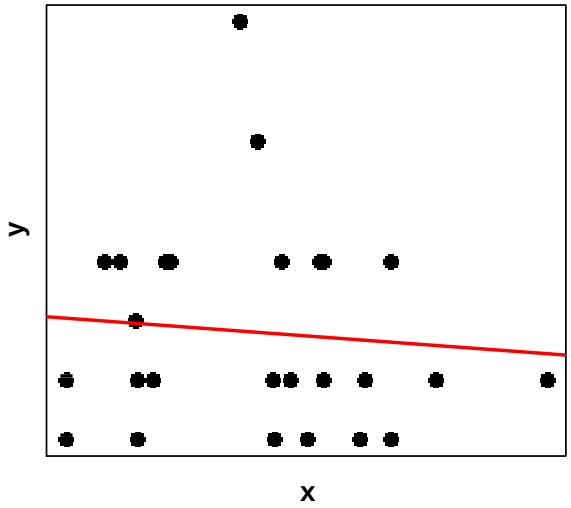
Coefficiente di correlazione di Pearson

E' un indicatore del livello di correlazione **lineare** tra **DUE** variabili, assume valori che vanno da -1 a 1, si indica con $\mathbf{r_{x,y}}$.

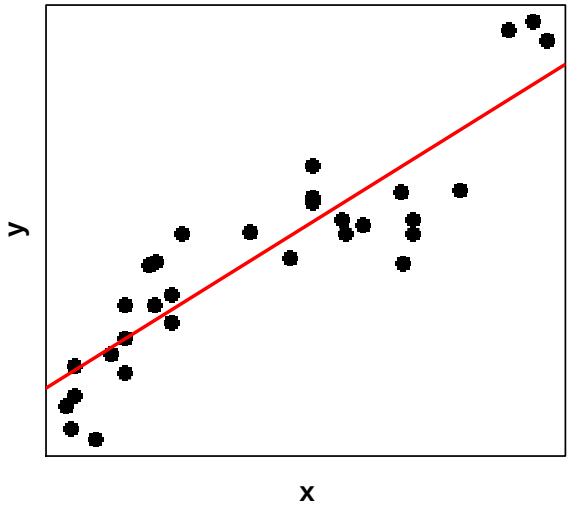
$r_{x,y} < 0$



$r_{x,y} \sim 0$

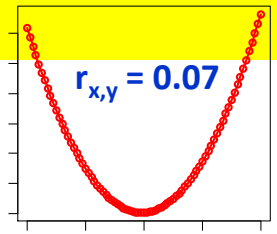


$r_{x,y} > 0$



ATTENZIONE!!!

$r_{x,y} \sim 0$ vuole solamente dire che tra x e y non c'è correlazione lineare (altre sono possibili)!



Indicativamente:

- se il suo valore assoluto è **> 0.75** la correlazione è **forte**;
- se il suo valore assoluto è **> 0.5 e <0.75** la correlazione è **moderata**;
- se il suo valore assoluto è **> 0.25 e <0.5** la correlazione è **debole**;
- se il suo valore assoluto è **< 0.25** la correlazione è **nulla.**

SEGUE

Coefficiente di correlazione di Pearson (2)

$$r_{x,y} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

$$r_{x,y} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2} \cdot \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2}} = \frac{\sigma_{x,y}}{\sqrt{\sigma_x^2 \cdot \sigma_y^2}} = \frac{\sigma_{x,y}}{\sigma_x \cdot \sigma_y}$$

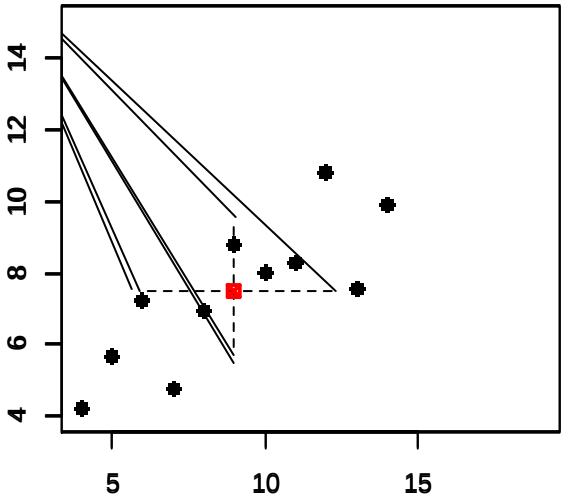
covarianza di x,y

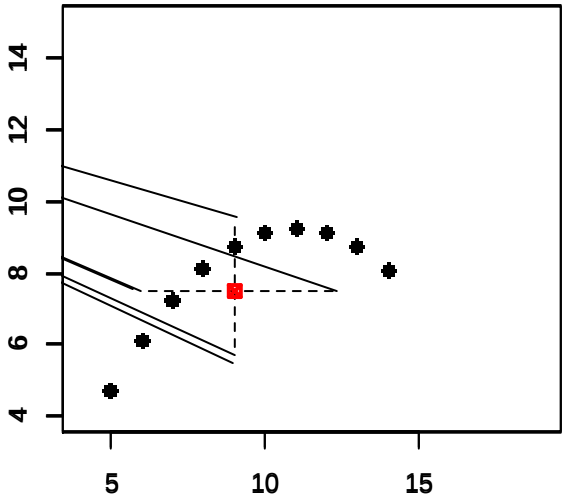
dev.std. di x

dev.std. di y

Quartetto di Anscombe

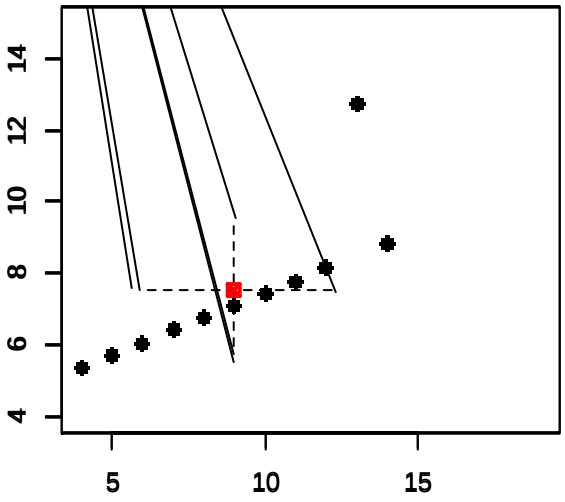
Media x = 9.000
Dev.std. x = 3.317
Media y = 7.500
Dev.std. y = 2.031
$r_{x,y}$ = 0.8164

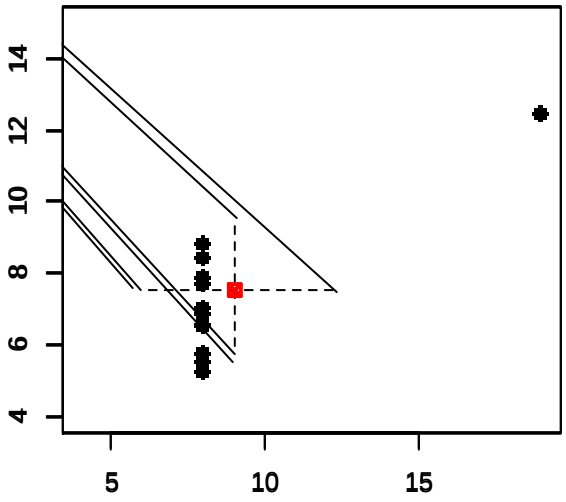




Media x = 9.000
Dev.std. x = 3.317
Media y = 7.501
Dev.std. y = 2.032
$r_{x,y}$ = 0.8162

Media x = 9.000
Dev.std. x = 3.317
Media y = 7.500
Dev.std. y = 2.030
$r_{x,y}$ = 0.8163





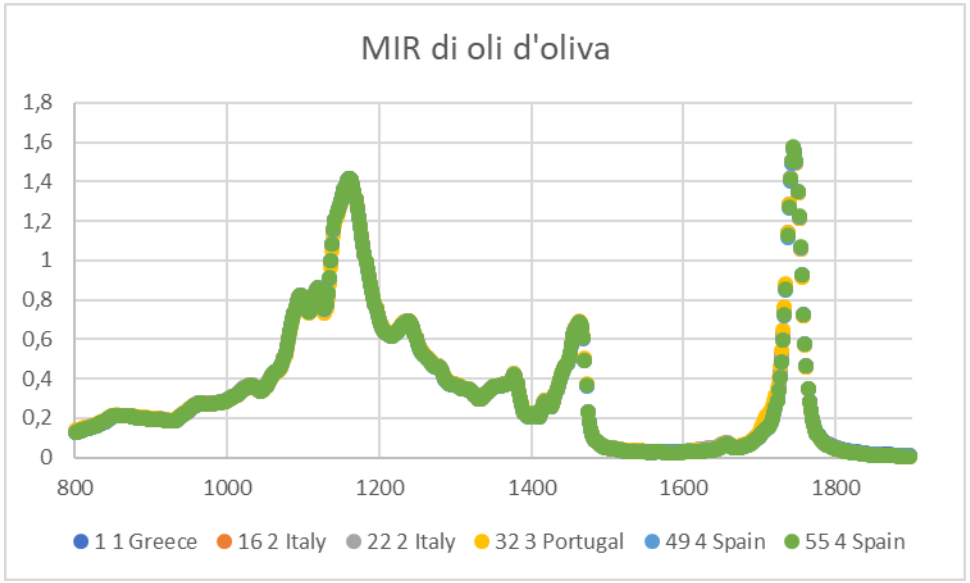
Media x = 9.000
Dev.std. x = 3.317
Media y = 7.501
Dev.std. y = 2.031
$r_{x,y}$ = 0.8165

L'esplorazione visiva dei grafici è FONDAMENTALE!!!

https://doi.org/10.1198/000313007X220057

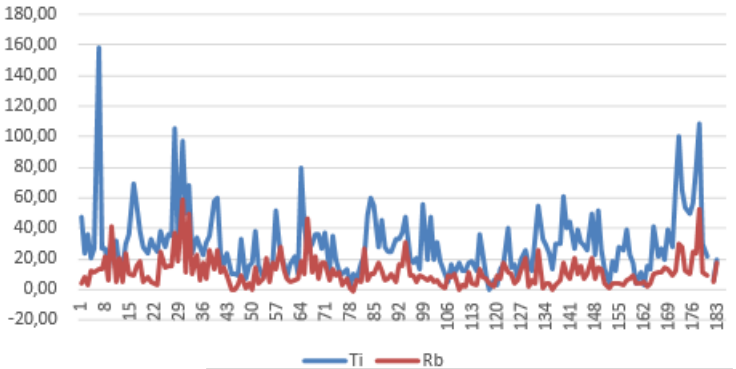
Due variabili / parametri misurati non sono sempre sufficienti

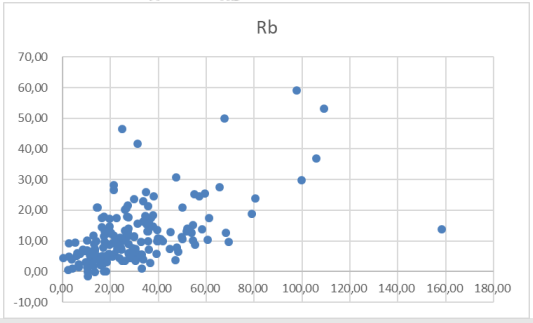
Dai dati bivariati ai dati multivariati



120 Mid-infrared (MIR) spectra. Duplicate acquisitions from 60 authenticated extra virgin olive oils from 4 different countries of origin; raw data matrix size [570 x 120]. Obtained using Fourier transform infrared (FTIR) spectroscopy with attenuated total reflectance (ATR) sampling. As described in "**FTIR spectroscopy and multivariate analysis can distinguish the geographic origin of extra virgin olive oils**" Tapp H.S. et al, J. Agric. Food Chem. 51 (21) 6110-5 (2003)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	
1	Column1	Column2	Column3	Column4	Column5	Column6	Column7	Column8	Column9	Column10	Column11	Column12	Column13	Column14	Column15	Column16	Column17	Col
2	Sample Number:	1	1	2	2	3	3	4	4	5	5	6	6	7	7	8	8	9
3	Group Code:	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
4	Wavenumbers	Greece	Greece	Greece	Greece	Greece	Greece	Greece	Greece	Greece	Greece	Greece	Greece	Greece	Greece	Greece	Greece	Gre
5	798,892	0,12752301	0,12649818	0,13041179	0,13002223	0,12860199	0,12821725	0,12617493	0,12646605	0,127060105	0,126812707	0,127010473	0,128155516	0,12657597	0,126516389	0,127732173	0,126827748	0,1
6	800,8215	0,12794962	0,12713097	0,1306754	0,13040666	0,12878957	0,12828225	0,12673277	0,12691541	0,127551128	0,127460743	0,127513059	0,128140808	0,12715497	0,127141851	0,127919122	0,127103881	0,1
7	802,751	0,12928222	0,12851078	0,13201661	0,13201803	0,13002228	0,12963664	0,12824377	0,12825412	0,128899981	0,128765265	0,128897571	0,128975154	0,128490024	0,128627318	0,128973128	0,128294409	0,1
8	804,6805	0,13117417	0,13033991	0,13382406	0,13400728	0,13201187	0,13179863	0,12989273	0,12995826	0,13060904	0,130639027	0,130888276	0,130859689	0,13025456	0,130513879	0,130898778	0,130351753	0,1
9	806,61	0,13359033	0,13252722	0,1360953	0,13627057	0,13442663	0,13406148	0,1317546	0,13206719	0,132955794	0,133131261	0,13335944	0,133643807	0,132711046	0,133011671	0,133344726	0,132894964	0,1
10	808,5395	0,13642553	0,13530851	0,13894376	0,13887477	0,13706977	0,13667076	0,13452791	0,13465524	0,135758628	0,135663579	0,135990448	0,136300073	0,135354557	0,135696443	0,13612911	0,135619642	0
11	810,469	0,13935783	0,13835292	0,14172278	0,14148113	0,13964566	0,13955847	0,13751151	0,13736041	0,138406848	0,138508136	0,138718434	0,138704405	0,138084047	0,138170288	0,138940003	0,138618977	0,1
12	812,3985	0,14221926	0,14115006	0,14413565	0,14407492	0,14233821	0,14225533	0,14042934	0,14003156	0,141161849	0,141473556	0,14150051	0,141451479	0,140845336	0,140968324	0,141714085	0,141412036	0,1
13	814,328	0,14480547	0,14363018	0,14643138	0,1465531	0,14492324	0,14457669	0,14296365	0,1427058	0,143976738	0,143872271	0,143928118	0,143439756	0,14341072	0,143640981	0,144330303	0,143892394	0,1
14	816,2575	0,14676377	0,14562408	0,14838928	0,14844275	0,14695816	0,1462924	0,14479855	0,14486289	0,146072437	0,145915995	0,145775355	0,145623596	0,145426534	0,145669997	0,146355427	0,145804407	0,1
15	818,187	0,14819234	0,14730907	0,14983736	0,14984542	0,14854536	0,14797596	0,14654969	0,14648832	0,14761742	0,147681335	0,147318474	0,147311361	0,146870054	0,147345111	0,147891757	0,147199096	0,1
16	820,1165	0,14993675	0,14899163	0,15158495	0,151712	0,15018934	0,14979499	0,14835769	0,14862904	0,149306422	0,149497205	0,148985452	0,148976521	0,148556627	0,148854584	0,149395012	0,148979844	0,1
17	822,046	0,15237287	0,15128378	0,15398	0,15426773	0,15260332	0,15191109	0,15065204	0,15122464	0,151735371	0,151591191	0,151270963	0,151805163	0,150643654	0,151180417	0,151601997	0,151310207	0
18	823,9755	0,15505019	0,15403706	0,15672449	0,15687223	0,15516948	0,15468297	0,15389796	0,1537382	0,154784059	0,154308656	0,154030479	0,154457689	0,15343865	0,154128238	0,154149522	0,15407315	0,1
19	825,905	0,15798803	0,15696826	0,15945045	0,15945635	0,15782759	0,15754525	0,15703056	0,15691555	0,158006301	0,157103101	0,157016701	0,157218279	0,156609721	0,156911358	0,156594276	0,156930463	0,1
20	827,8345	0,16080862	0,16015204	0,16258834	0,16240766	0,16063266	0,16076073	0,16005217	0,16017093	0,160993075	0,160068491	0,16016326	0,160235883	0,159583948	0,159926608	0,159460073	0,160068424	0,1
21	829,764	0,16379976	0,16319874	0,16593386	0,16582185	0,16359815	0,16384707	0,16308226	0,16330714	0,163981801	0,163455048	0,163241095	0,163379104	0,162656016	0,162978499	0,162607645	0,16327876	0,1
22	831,6935	0,16724956	0,1661921	0,16909325	0,16915942	0,16701043	0,16688922	0,16633527	0,16676444	0,167266777	0,166992643	0,166488773	0,167026297	0,165873594	0,166189431	0,165903249	0,166421102	0,
23	833,623	0,17081176	0,16939021	0,17239712	0,17227908	0,1703824	0,16996263	0,17005585	0,17007026	0,170468947	0,170313844	0,169942221	0,170424865	0,169150493	0,169594748	0,169404365	0,169684397	0,1
24	835,5525	0,17431873	0,17294206	0,17571468	0,17559527	0,17358774	0,17335513	0,17335174	0,17351435	0,173871151	0,173882254	0,173267474	0,173937471	0,172539479	0,173158003	0,172877275	0,1733436	0,1
25	837,482	0,1779136	0,17708447	0,17949055	0,1794698	0,17743357	0,17748733	0,17713987	0,17750691	0,177825114	0,177995496	0,17727538	0,17840013	0,176562256	0,177201328	0,176697166	0,1774603	0,1
26	839,4115	0,18209116	0,18139716	0,18378728	0,18381641	0,18177611	0,18177181	0,18196762	0,18180918	0,182110688	0,182410978	0,181909134	0,182462321	0,181100017	0,18150301	0,181432924	0,181771667	0,1
27	841,341	0,18670716	0,18592491	0,18850149	0,18856413	0,1864131	0,18646578	0,18663773	0,18650174	0,186973981	0,187008687	0,186560889	0,187253313	0,185973307	0,186239359	0,186681787	0,186552988	0,1
28	843,2705	0,191972	0,19102042	0,19357544	0,19360676	0,19160546	0,19159602	0,19156334	0,19176091	0,192340962	0,191570066	0,191507255	0,192401902	0,191091573	0,191427719	0,191882342	0,19173898	0,1
29	845,2	0,19700805	0,19603465	0,19856573	0,19848271	0,19681511	0,19658184	0,1966703	0,19670938	0,197354875	0,196672954	0,196476017	0,196969376	0,195944016	0,196340482	0,196900879	0,196898164	0





Guiyang_PM2.5_speciation_data_DOI_10.1007/s11631-017-0248-1 (178x42)

Abstract: ED-XRF inorganic speciation of PM2.5 in Guiyang, China. This dataset is associated with the following publication: Liang, L., N. Liu, M. Landis, X. Xu, X. Feng, Z. Chen, L. Shang, and G. Qiu. Chemical characterization and sources of PM2.5 at 12-h resolution in Guiyang, China. Acta Geochimica. Springer, Heidelburg, GERMANY, 37(2): 334-345, (2018).

File Identifier: https://doi.org/10.23719/1419639

https://link.springer.com/content/pdf/10.1007/s11631-017-0248-1.pdf

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF
1	DATE	MON	DAY	YEAR	SITE	SAMPLE ID	UNITS	MASS	MASS	Na	Na	Mg	Mg	Al	Al	Si	Si	P	P	S	S	Cl	Cl	K	K	Ca	Ca	Ti	Ti	Cr	Cr	Mn
2	FINE CONCENTRATIONS							CONC	UNC	CONC	UNC	CONC	UNC	CONC	UNC	CONC	UNC	CONC	UNC	CONC	UNC	CONC	UNC	CONC	UNC	CONC	UNC	CONC	UNC	CONC	UNC	CONC
3	9/ 3/2008	9	3	2008	GY	GZ-003	NG/M3	42374,86	2220,41	864,92	157,14	210,78	51,70	546,68	80,87	967,41	102,16	11,84	17,75	5688,03	423,76	-8,84	6,44	982,46	73,89	501,73	37,75	47,20	5,10	2,83	1,18	86,0
4	9/ 3/2008	9	3	2008	GY	GZ-004	NG/M3	29184,55	1521,03	428,97	116,20	216,14	42,42	546,91	67,27	1676,05	157,06	157,04	18,50	3437,54	256,25	27,07	5,49	1540,19	114,80	663,07	49,49	23,92	2,89	2,62	0,79	163,1
5	9/ 4/2008	9	4	2008	GY	GZ-005	NG/M3	40119,33	2070,86	404,46	104,22	290,57	48,75	547,29	68,95	1187,99	115,91	263,52	27,63	5717,52	426,03	13,03	5,32	923,87	69,30	579,71	43,37	36,54	3,89	3,08	0,91	102,9
6	9/ 4/2008	9	4	2008	GY	GZ-006	NG/M3	39332,27	2003,20	465,16	108,49	225,38	38,91	287,31	46,27	1655,02	154,81	578,38	48,20	5048,09	375,53	2,35	4,52	2081,83	154,87	444,73	33,32	20,19	2,58	1,70	0,70	275,0
7	9/ 5/2008	9	5	2008	GY	GZ-007	NG/M3	48809,84	2530,43	474,58	149,03	308,88	87,05	392,66	70,25	1628,02	157,69	859,70	71,91	5938,19	442,38	-5,87	6,32	1664,16	124,33	533,74	40,14	26,38	3,78	4,37	1,18	208,5
8	9/ 5/2008	9	5	2008	GY	GZ-008	NG/M3	68226,63	3474,08	133,00	129,38	356,61	61,76	533,13	76,25	4000,48	368,08	364,79	37,00	6658,46	495,86	55,13	8,94	1998,15	149,12	1058,51	78,97	158,26	12,73	3,15	1,15	130,8
9	9/ 6/2008	9	6	2008	GY	GZ-009	NG/M3	56489,57	2903,43	145,89	114,56	265,95	54,47	224,42	59,24	3106,93	288,51	141,52	23,97	7545,28	561,55	0,62	6,84	1707,83	127,56	396,92	30,06	27,30	3,85	1,81	1,14	211,6
10	9/ 6/2008	9	6	2008	GY	GZ-013	NG/M3	64468,37	3292,91	541,96	217,84	341,41	66,55	490,98	85,25	3509,10	325,16	107,84	24,91	9654,76	718,02	67,71	10,27	3160,48	235,24	824,22	61,69	27,40	4,10	2,45	1,22	944,2
11	9/ 7/2008	9	7	2008	GY	GZ-014	NG/M3	39539,81	2084,63	274,02	107,16	269,11	50,09	54,69	49,43	1877,13	179,42	38,06	17,67	5245,20	391,02	2,42	6,50	559,82	42,68	343,14	26,02	12,90	3,27	5,06	1,25	153,6
12	9/ 7/2008	9	7	2008	GY	GZ-015	NG/M3	50917,86	2632,57	781,44	192,16	319,89	68,82	378,80	72,84	2430,97	229,53	249,30	30,00	5643,11	420,71	94,56	11,78	4884,84	363,05	1173,57	87,57	31,26	4,35	5,19	1,31	683,7
13	9/ 8/2008	9	8	2008	GY	GZ-010	NG/M3	29809,22	1631,04	230,35	90,31	142,84	42,52	231,30	53,55	1236,32	122,94	143,16	20,43	3283,84	245,56	-7,40	5,67	776,64	58,68	532,95	40,04	31,68	4,14	2,05	1,11	107,8
14	9/ 8/2008	9	8	2008	GY	GZ-011	NG/M3	41141,79	2163,47	410,58	186,92	397,34	103,39	422,21	72,68	5025,50	461,55	340,05	33,96	3457,03	258,61	43,55	8,12	2226,31	166,01	998,91	74,56	14,48	3,46	4,83	1,19	618,5
15	9/ 9/2008	9	9	2008	GY	GZ-012	NG/M3	25721,79	1443,81	175,77	78,35	115,94	34,71	-38,59	40,75	1694,82	162,07	94,93	16,89	2718,35	203,62	11,34	5,50	476,24	36,47	333,91	25,29	14,94	3,08	-0,05	0,99	103,4
16	9/ 9/2008	9	9	2008	GY	GZ-016	NG/M3	48069,24	2493,97	771,35	200,52	423,10	66,38	205,76	57,88	2891,95	271,23	324,26	33,48	4651,83	347,08	120,06	12,49	2280,16	170,00	936,31	69,93	30,02	4,15	1,22	1,17	424,5
17	9/10/2008	9	10	2008	GY	GZ-017	NG/M3	38214,05	2020,37	156,42	85,42	149,87	41,59	120,10	50,88	1220,62	122,14	94,05	19,82	5730,90	426,92	-0,96	6,08	978,45	73,57	453,01	34,14	35,86	4,28	0,32	1,06	140,8
18	9/10/2008	9	10	2008	GY	GZ-018	NG/M3	46036,83	2396,59	797,26	191,25	428,48	78,37	1058,60	121,16	1907,15	182,89	145,18	23,58	5746,12	428,15	16,58	7,36	1966,25	146,75	1041,47	77,71	69,45	6,59	2,05	1,16	147,5
19	9/11/2008	9	11	2008	GY	GZ-019	NG/M3	64082,49	3269,90	530,03	133,46	598,71	90,48	781,12	101,48	1868,77	180,01	77,12	23,39	9052,60	673,30	-1,79	7,29	1974,55	147,37	1335,85	99,51	54,37	5,54	4,17	1,25	140,3
20	9/11/2008	9	11	2008	GY	GZ-020	NG/M3	76390,93	3889,80	287,33	104,46	424,05	68,99	449,92	78,82	5587,76	511,64	747,81	65,62	9914,87	737,52	8,91	9,54	3007,05	223,94	583,98	43,98	37,80	4,65	3,79	1,26	601,3
21	9/12/2008	9	12	2008	GY	GZ-021	NG/M3	55757,79	2823,01	263,31	75,20	120,17	34,59	302,29	59,20	1991,53	189,54	205,16	25,48	9081,40	675,33	3,14	4,95	1076,66	80,49	473,69	35,48	27,72	3,15	2,27	0,80	172,6
22	9/12/2008	9	12	2008	GY	GZ-022	NG/M3	69736,31	3550,46	8,58	119,60	187,03	62,86	49,98	75,73	4601,75	422,42	205,11	29,09	10400,79	773,34	8,63	8,01	1139,65	85,55	588,17	44,14	24,21	3,72	3,18	1,18	280,1
23	9/13/2008	9	13	2008	GY	GZ-023	NG/M3	94786,07	4788,96	168,03	102,72	263,89	57,69	200,66	74,04	4667,39	429,02	161,55	30,65	15502,37	1152,03	-11,47	7,79	1241,34	93,11	575,57	43,24	33,10	4,38	7,03	1,39	364,6
24	9/13/2008	9	13	2008	GY	GZ-024	NG/M3	55764,08	2867,63	249,50	113,12	284,53	54,23	369,01	68,30	5509,07	503,41	354,04	37,95	11320,87	841,45	5,80	6,81	1371,69	102,67	508,92	38,28	28,00	3,83	7,32	1,30	306,5
25	9/14/2008	9	14	2008	GY	GZ-025	NG/M3	47004,61	2440,69	125,14	104,51	117,11	52,73	219,57	59,48	2014,62	190,92	139,45	23,12	7669,88	570,68	-1,66	5,97	814,46	61,42	344,92	26,15	24,87	3,66	2,22	1,07	84,5
26	9/14/2008	9	14	2008	GY	GZ-026	NG/M3	77067,17	3911,00	237,16	128,97	576,64	88,59	785,28	107,44	4890,10	448,91	309,07	35,92	10523,66	782,61	42,78	10,79	3740,98	278,28	768,44	57,58	38,01	4,63	2,44	1,23	471,8
27	9/15/2008	9	15	2008	GY	GZ-027	NG/M3	84403,67	4270,76	474,43	160,72	451,09	78,23	555,01	88,56	2013,04	192,21	356,84	39,55	14211,23	1056,03	-1,13	7,64	2454,87	182,95	667,23	50,05	27,48	3,85	6,93	1,33	298,2
28	9/15/2008	9	15	2008	GY	GZ-028	NG/M3	99827,20	5035,42	41,22	270,15	578,66	85,78	682,30	93,66	3947,62	364,34	237,91	34,81	16245,95	1207,16	87,65	11,65	2820,59	210,09	1023,70	76,44	35,96	4,57	5,78	1,37	1336,9
29	9/16/2008	9	16	2008	GY	GZ-029	NG/M3	75303,96	3821,52	330,97	114,32	266,35	52,31	380,53	72,25	2475,17	233,02	332,89	37,14	11682,97	868,47	0,60	9,21	2477,99	184,65	764,77	57,25	35,18	4,45	2,74	1,22	321,6
30	9/16/2008	9	16	2008	GY	GZ-030	NG/M3	109818,95	5530,56	870,05	315,31	738,47	111,85	993,69	125,40	6423,82	588,85	434,81	46,48	12163,81	904,50	357,99	29,65	4766,26	354,40	3748,48	278,38	105,99	9,19	8,04	1,47	478,4
31	9/17/2008	9	17	2008	GY	GZ-031	NG/M3	70067,85	3563,62	356,10	146,46	231,17	51,99	284,00	69,73	3433,38	317,98	167,13	27,37	10706,54	796,00	11,32	7,53	1645,94	123,04	694,67	52,06	34,50	4,36	2,98	1,19	258,9
32	9/17/2008	9	17	2008	GY	GZ-032	NG/M3	126593,14	6370,03	1144,70	266,99	747,21	110,80	809,53	122,78	10146,73	924,82	551,05	56,30	14486,00	1077,01	80,70	13,67	8015,05	595,31	3478,88	258,48	97,69	8,89	3,84	1,50	960,9
33	9/18/2008	9	18	2008	GY	GZ-033	NG/M3	56967,26	2909,02	183,68	87,00	172,61	43,79	429,33	71,32	1977,27	187,26	163,29	24,79	9020,22	670,70	-0,99	6,42	1275,97	95,49	665,60	49,81	49,74	5,01	3,22	1,07	236,9
34	9/18/2008	9	18	2008	GY	GZ-034	NG/M3	100781,14	5079,57	726,00	218,18	528,65	89,71	781,57	111,54	5982,54	549,30	659,82	60,91	12771,20	949,44	50,24	10,84	6903,82	512,77	2480,02	184,36	67,77	6,59	9,27	1,48	1064,1
35	9/19/2008	9	19	2008	GY	GZ-035	NG/M3	59918,47	3071,98	257,78	98,68	150,85	45,89	205,07	63,63	2847,14	265,50	150,15	25,64	9377,87	697,45	-1,65	6,78	1276,91	95,69	485,06	36,54	24,19	3,75	3,47	1,18	245,8
36	9/19/2008	9	19	2008	GY	GZ-036	NG/M3	72835,42	3701,26	422,09	134,96	378,64	66,67	352,34	73,79	5121,74	468,95	159,55	27,01	10536,33	783,41	5,44	7,50	2307,27	172,00	890,59	66,55	33,72	4,35	12,68	1,57	537,5
37	9/20/2008	9	20	2008	GY	GZ-037	NG/M3	46292,66	2405,21	176,10	86,49	185,86	44,11	249,94	57,67	1599,70	154,84	138,04	23,55	8069,09	600,26	0,92	6,36	879,34	66,21	459,21	34,57	27,59	3,86	2,91	1,11	143,0
38	9/20/2008	9	20	2008	GY	GZ-038	NG/M3	42629,01	2231,26	385,82	112,42	456,45	74,53	158,13	59,15	2850,86	265,68	158,58	22,92	4817,10	359,31	23,45	7,33	1465,76	109,62	549,49	41,30	22,68	3,74	2,98	1,16	560,2
39	9/21/2008	9	21	2008	GY	GZ-039	NG/M3	26408,22	1471,93	253,43	87,08	163,00	45,40	159,12	50,82	1766,93	168,38	350,41	33,26	2764,42	207,06	13,35	5,78	770,60	58,16	444,17	33,44	30,59	3,90	0,94	0,99	173,5
40	9/21/2008	9	21	2008	GY	GZ-040	NG/M3	35878,57	1910,49	210,89	99,91	286,17	58,06	432,34	82,28	2703,51	258,84	251,47	30,04	3687,29	275,67	73,06	10,05	3263,35	242,83	1304,19	97,19	34,95	4,44	2,23	1,15	865,0
41	9/22/2008	9	22	2008	GY	GZ-041	NG/M3	38048,40	2009,46	311,91	93,30	354,59	59,55	407,29	66,54	1744,62	169,76	177,67	23,94	4355,67	325,67	-3,92	6,10	1303,08	97,60	838,71	62,66	58,20	5,67	1,40	1,11	199,0

Matrice di correlazione

JEM

Survey of environmental complex systems: pattern recognition of physicochemical data describing coastal water quality in the Gulf of Trieste

Pierluigi Barbieri,[a] Gianpiero Adami,[a] Sergio Predonzani,[b] Edoardo Reisenhofer[a] and Desiré Luc Massart[c]

[a]*Department of Chemical Sciences, University of Trieste, via L . Giorgieri 1, I-34127, Trieste, Italy*
[b]*Laboratory of Marine Biology, Strada Costiera 336, I-34010 Santa Croce-Trieste, Italy*
[c]*ChemoAC, Pharmaceutical Institute, Vrije Universiteit Brussel, Laarbeeklaan 103, B-1090 Brussels, Belgium*

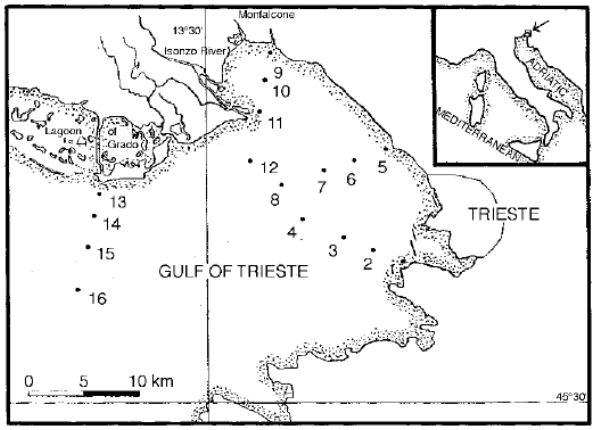


Fig. 1 Map of the sampling area.

Table 1 Basic statistics for the 909 valid cases

	TEMP/ °C	SAL/ practical salinity units	O_2/ μmol dm^{-3}	N-NH_3/ μmol dm^{-3}	N-NO_2/ μmol dm^{-3}	N-NO_3/ μmol dm^{-3}	Si$(OH)_4$/ μmol dm^{-3}	Chlo.a/ μg dm^{-3}	Phaeop./ μg dm^{-3}
Mean	14.74	36.06	243.24	0.90	0.28	4.67	3.35	0.47	0.47
s	5.25	2.74	34.11	1.10	0.31	9.30	4.41	0.42	0.64
Min	5.85	17.58	18.30	0.05	0.02	0.02	0.02	0.01	0.01
Max	26.99	37.90	305.80	13.73	2.60	125.38	44.65	2.90	9.72

Table 2 Correlation matrix for 909 cases

	TEMP	SAL	O_2	N-NH_3	N-NO_2	N-NO_3	Si$(OH)_4$	Chlo.a	Phaeop.
TEMP	1.0000								
SAL	−0.2541	1.0000							
O_2	−0.5302	−0.1108	1.0000						
N-NH_3	0.1851	−0.2283	−0.3033	1.0000					
N-NO_2	−0.3291	−0.1139	0.1382	0.1261	1.0000				
N-NO_3	0.1295	−0.7243	0.0999	0.2437	0.1002	1.0000			
Si$(OH)_4$	0.0612	−0.4569	−0.1968	0.3029	0.1040	0.4550	1.0000		
Chlo.a	0.0024	−0.1350	0.0102	−0.0615	−0.0653	0.0400	−0.0386	1.0000	
Phaeop.	−0.1547	−0.0765	0.0800	−0.0487	0.1134	0.0138	0.0024	0.6798	1.0000

Comprimere l'informazione: analisi delle componenti principali e metodi fattoriali

Nozioni di algebra delle matrici

Moltiplicazione matrice-vettore: $M \cdot \vec{v} = \vec{v}_r$

$$\begin{pmatrix} a & b & c \\ d & e & f \end{pmatrix} \cdot \begin{pmatrix} g \\ h \\ i \end{pmatrix} = \begin{pmatrix} a \cdot g + b \cdot h + c \cdot i \\ d \cdot g + e \cdot h + f \cdot i \end{pmatrix} = \begin{pmatrix} j \\ k \end{pmatrix}$$

Dimensioni: n x m m x 1 n x 1

Attenzione!!! $A \cdot B \neq B \cdot A$

Moltiplicazione matrice-matrice: $A \cdot B = C$

$$\begin{pmatrix} a & b \\ c & d \\ e & f \end{pmatrix} \cdot \begin{pmatrix} g & h \\ i & j \end{pmatrix} = \begin{pmatrix} a \cdot g + b \cdot i & a \cdot h + b \cdot j \\ c \cdot g + d \cdot i & c \cdot h + d \cdot j \\ e \cdot g + f \cdot i & e \cdot h + f \cdot j \end{pmatrix} = \begin{pmatrix} k & l \\ m & n \\ o & p \end{pmatrix}$$

Dimensioni: n x m m x r n x r

Significato geometrico

Moltiplicazione matrice-vettore: $A \cdot \vec{x} = \vec{v}$

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} a \cdot x_1 + b \cdot x_2 \\ c \cdot x_1 + d \cdot x_2 \end{pmatrix} = \begin{pmatrix} e \\ f \end{pmatrix}$$

Dimensioni: n x m m x 1 n x 1

Si può pensare come:

$$a \cdot x_1 + b \cdot x_2 = e$$

$$c \cdot x_1 + d \cdot x_2 = f$$

e si può scomporre in:

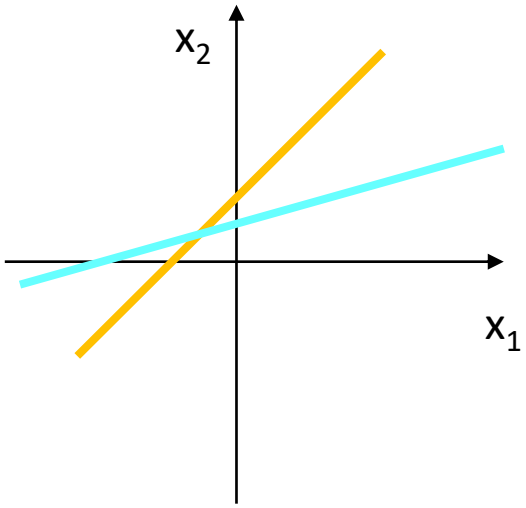
$$x_1 \cdot \begin{pmatrix} a \\ c \end{pmatrix} + x_2 \cdot \begin{pmatrix} b \\ d \end{pmatrix} = \begin{pmatrix} e \\ f \end{pmatrix}$$

Significato geometrico (2)

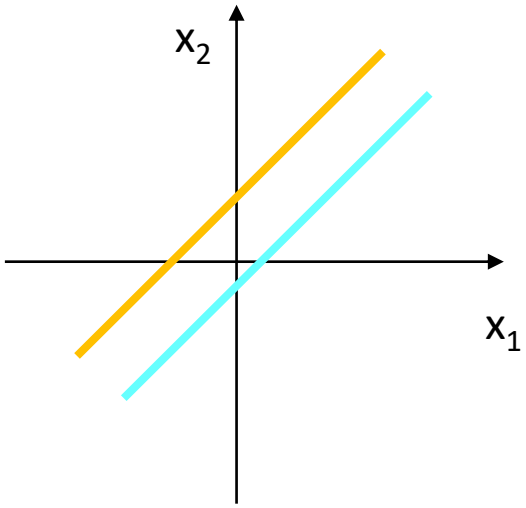
$\vec{A} \cdot \vec{x} = \vec{v}$

$a \cdot x_1 + b \cdot x_2 = e$ → **1** $x_2 = \dfrac{e - a \cdot x_1}{b}$

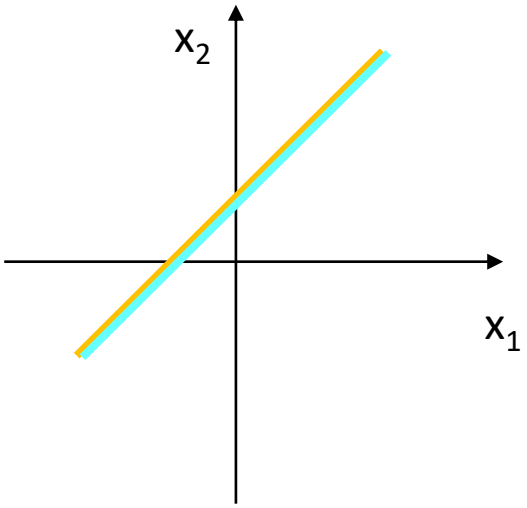
$c \cdot x_1 + d \cdot x_2 = f$ → **2** $x_2 = \dfrac{f - c \cdot x_1}{d}$



1 soluzione



nessuna soluzione



infinite soluzioni

La matrice identità

E' una matrice quadrata diagonale in cui tutti gli elementi sono uguali a zero, tranne che quelli presenti sulla diagonale, che sono uguali a 1.

Si indica con I_n

$$I_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \qquad I_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Dimensioni: n x m con n=m

E' tale per cui:

$$I_n \cdot \vec{v} = \vec{v}$$

$$I_n \cdot A = A$$

$$A \cdot I_n = A$$

Quindi "funziona" come la moltiplicazione per 1 in algebra classica

Nota: ovviamente le dimensioni di I devono essere "compatibili" per effettuare le operazioni di moltiplicazione matrice-vettore o matrice-matrice.

In R:

diag(numerointero)

```
> diag(3)
     [,1] [,2] [,3]
[1,]    1    0    0
[2,]    0    1    0
[3,]    0    0    1
>
```

Matrice invertibili e matrici singole

Mentre di qualsiasi numero ≠ da zero si può ottenere l'inverso, per le matrici non è sempre così.

Una matrice è invertibile se è quadrata (n x n) **ed** esiste una matrice B tale che:

$$A \cdot B = I_n$$

$$B \cdot A = I_n$$

Se queste **condizioni** vengono **raggiunte** allora si dice che **A è invertibile** e si ha $\mathbf{B = A^{-1}}$, cioè B è la matrice inversa di A

Se **non esiste** una matrice B che soddisfi tali condizioni allora si dice che **A è una matrice singola**, cioè non invertibile.

Nota: la matrice inversa di I_n è uguale a I_n

Risolvere equazioni matrice-vettore

In algebra "classica":

$$a \cdot x = b$$

$$\frac{1}{a} \cdot a \cdot x = \frac{1}{a} \cdot b$$

$$1 \cdot x = \frac{1}{a} \cdot b$$

$$x = \frac{b}{a}$$

$\vec{x}$ è anche detto vettore dei "pesi" (***weights***) da assegnare agli elementi di A per riuscire ad ottenere $\vec{b}$

In algebra lineare:

$$A \cdot \vec{x} = \vec{b}$$

$$A^{-1} \cdot A \cdot \vec{x} = A^{-1} \cdot \vec{b}$$

$$I \cdot \vec{x} = A^{-1} \cdot \vec{b}$$

$$\vec{x} = A^{-1} \cdot \vec{b}$$

quindi è necessario l'inverso di a!!!

quindi è necessaria l'inversa di A!!!

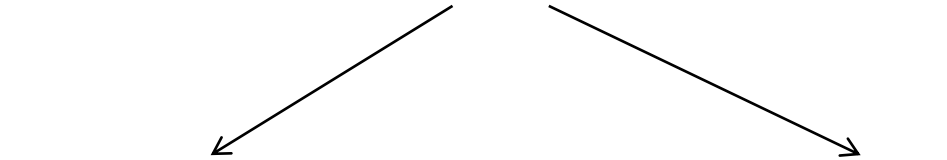
Nota: A è una matrice quadrata.

Risolvere equazioni matrice - vettore in R - casi reali

Problema: quando i dataset (cioè le matrici) sono prodotti da casi reali è molto difficile che le matrici siano sempre quadrate.

$$A \cdot \vec{x} = \vec{b}$$

Dimensioni di A: n x m



Caso 1: più osservazioni che variabili

Se la matrice A ha più righe che colonne, di conseguenza ci sono più equazioni disponibili rispetto alle soluzioni possibili.

$$\begin{pmatrix} 5 & 2 \\ -3 & 4 \\ 2 & 6 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 4 \\ 3 \\ 7 \end{pmatrix}$$

$$5 \cdot x_1 + 2 \cdot x_2 = 4$$

$$-3 \cdot x_1 + 4 \cdot x_2 = 3$$

$$2 \cdot x_1 + 6 \cdot x_2 = 7$$

In questo caso la terza equazione è la somma delle prime due. Non è utile, ma non impedisce di risolvere l'equazione

L'equazione si dice CONSISTENTE

$$\begin{pmatrix} 5 & 2 \\ -3 & 4 \\ 2 & 6 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 4 \\ 3 \\ -1 \end{pmatrix}$$

$$5 \cdot x_1 + 2 \cdot x_2 = 4$$

$$-3 \cdot x_1 + 4 \cdot x_2 = 3$$

$$2 \cdot x_1 + 6 \cdot x_2 = -1$$

In questo caso la terza equazione è incompatibile con le prime due. Quindi impedisce di ottenere una soluzione

L'equazione si dice INCONSISTENTE

Caso 1: rappresentazione grafica

$$\begin{pmatrix} 5 & 2 \\ -3 & 4 \\ 2 & 6 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 4 \\ 3 \\ 7 \end{pmatrix}$$

$$\begin{pmatrix} 5 & 2 \\ -3 & 4 \\ 2 & 6 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 4 \\ 3 \\ -1 \end{pmatrix}$$

$5 \cdot x_1 + 2 \cdot x_2 = 4$

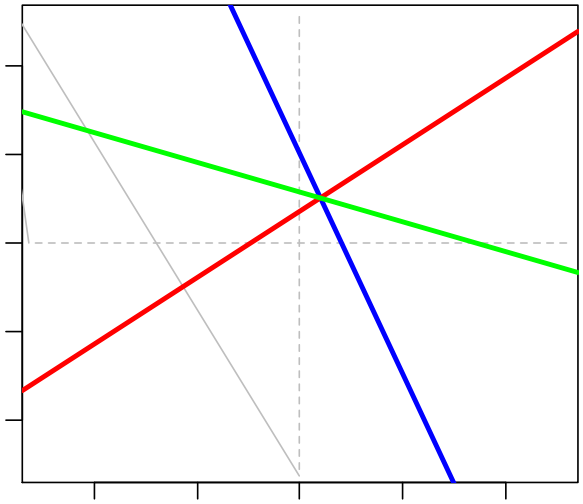
$-3 \cdot x_1 + 4 \cdot x_2 = 3$

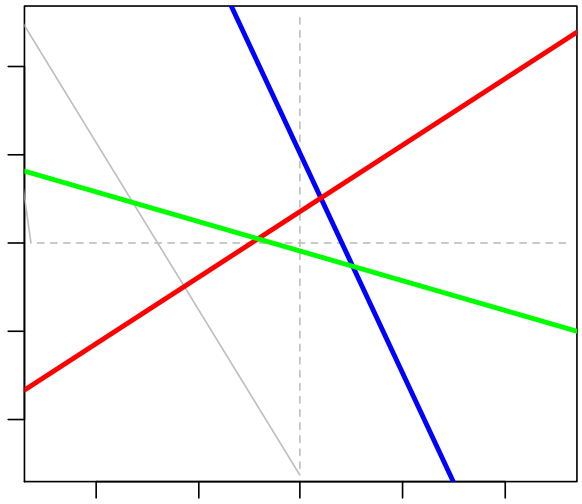
$2 \cdot x_1 + 6 \cdot x_2 = 7$

$5 \cdot x_1 + 2 \cdot x_2 = 4$

$-3 \cdot x_1 + 4 \cdot x_2 = 3$

$2 \cdot x_1 + 6 \cdot x_2 = -1$





Caso 2: meno osservazioni che variabili

Se la matrice A ha meno righe che colonne, di conseguenza ci sono meno equazioni disponibili rispetto alle soluzioni possibili.

$$\begin{pmatrix} 3 & -5 & 1 \\ 9 & -1 & 2 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 5 \\ -1 \end{pmatrix}$$

$$3 \cdot x_1 - 5 \cdot x_2 + 1 \cdot x_3 = 5$$

$$9 \cdot x_1 - 1 \cdot x_2 + 2 \cdot x_3 = -1$$

Manca una condizione per riuscire ad avere una soluzione univoca.

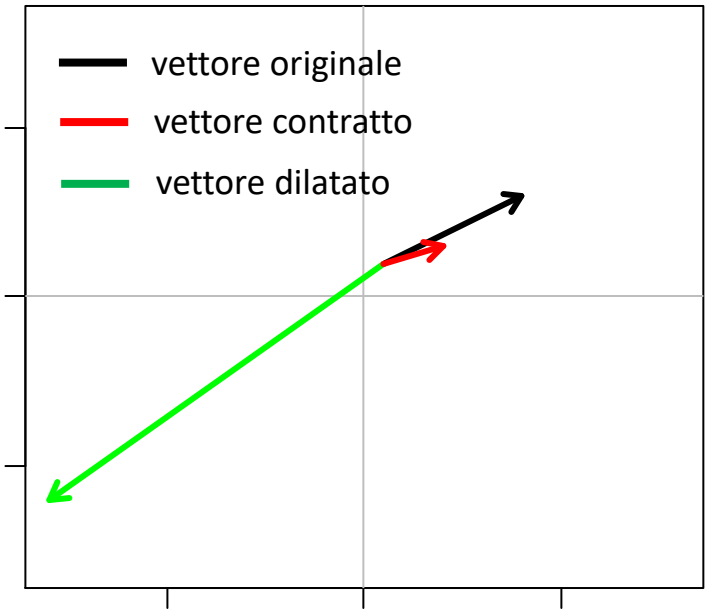
Ci sono diversi approcci per cercare di ottenere una soluzione univoca

Interpretazione geometrica di $A \cdot \vec{x} = \vec{b}$

Una matrice moltiplicata per un vettore può indurre diversi effetti sul vettore:

- Rotazioni;
- Riflessioni;
- Contrazioni;
- Dilatazioni;
- Proiezioni;
- Combinazioni delle precedenti.

Nota: uno **scalare** moltiplicato per un vettore, cioè $\mathbf{a} \cdot \vec{b}$, può avere come unico effetto di contrarre o dilatare tutte le componenti del vettore allo stesso modo



In R: `2*nomevettore`

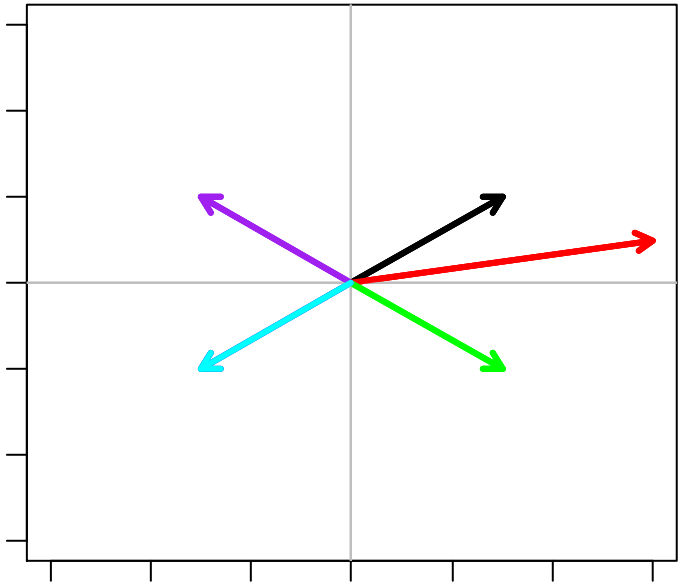
Interpretazione geometrica di $A \cdot \vec{x} = \vec{b}$

$$\begin{pmatrix} 2 & 0 \\ 0 & 0.5 \end{pmatrix} \cdot \begin{pmatrix} 3 \\ 2 \end{pmatrix} = \begin{pmatrix} 6 \\ 1 \end{pmatrix}$$

$$\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \cdot \begin{pmatrix} 3 \\ 2 \end{pmatrix} = \begin{pmatrix} 3 \\ -2 \end{pmatrix}$$

$$\begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 3 \\ 2 \end{pmatrix} = \begin{pmatrix} -3 \\ 2 \end{pmatrix}$$

$$\begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix} \cdot \begin{pmatrix} 3 \\ 2 \end{pmatrix} = \begin{pmatrix} -3 \\ -2 \end{pmatrix}$$

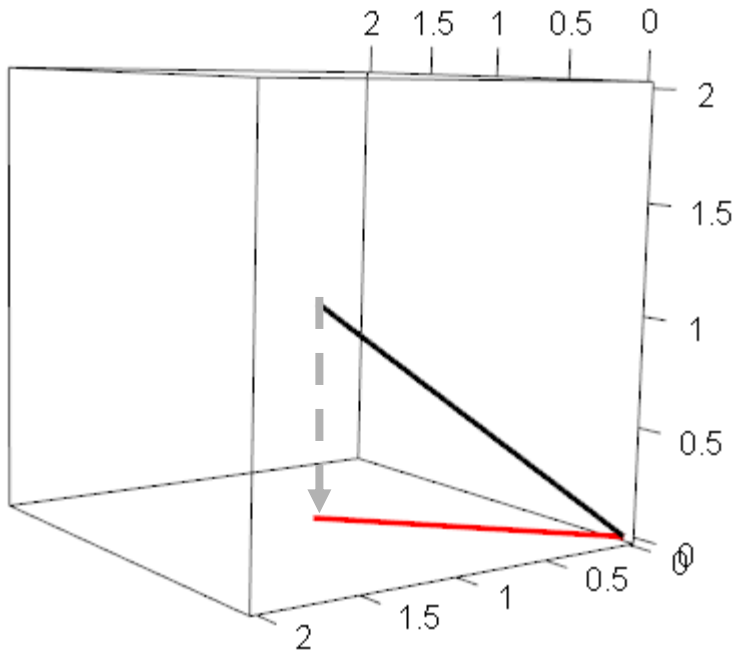


- vettore originale
- vettore contratto e dilatato
- vettore riflesso su asse x
- vettore riflesso su origine
- vettore riflesso su asse y

Interpretazione geometrica di $A \cdot \vec{x} = \vec{b}$

Proiezione:

$$\begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix}$$



vettore originale

vettore proiettato

Autovalori (*eigenvalues*) e autovettori (*eigenvectors*)

$$A \cdot \vec{v} = \lambda \cdot \vec{v}$$

(con $\vec{v} \neq \vec{0}$)

λ è un autovalore di **A** con un autovettore associato $\vec{\mathbf{v}}$

L'operazione di moltiplicazione matrice-vettore $A \cdot \vec{v}$ produce lo stesso effetto dell'operazione di moltiplicazione scalare-vettore $\lambda \cdot \vec{v}$

Nota: non è necessario che A sia una matrice diagonale

La coppia autovalore-autovettore (*eigenpair*) ha su $\vec{v}$ lo stesso effetto della matrice identità

$$I \cdot \vec{v} = \vec{v}$$

Autovalori (*eigenvalues*) e autovettori (*eigenvectors*) (2)

$$A \cdot \vec{v} = \lambda \cdot \vec{v}$$

$$\begin{pmatrix} 2 & 3 \\ 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 0 \end{pmatrix} = 2 \cdot \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

Quindi **2** è un autovalore di **A** accoppiato all'autovettore $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$

Dal punto di vista geometrico: per una matrice A che può operare una rotazione, riflessione, proiezione o combinazione di queste su un generico vettore, l'autovettore è un vettore che rimane "fisso" rispetto a questa operazione (cioè rimane sulla stessa "linea").

Un autovettore può essere **riscalato** a piacere (cioè essere moltiplicato per uno scalare) perché la direzione (linea) del vettore rimane la stessa indifferentemente dalla sua grandezza (*magnitude*). Quindi un autovettore è un multiplo scalare di se stesso.

per questo si definisce "**auto**", a causa della sua "**indipendenza**"

Proprietà delle soluzioni

$$A \cdot \vec{v} = \lambda \cdot \vec{v}$$

- ✓ Risolvere l'equazione sopra equivale a risolvere l'equazione: $A \cdot \vec{v} - \lambda \cdot \vec{v} = \vec{0}$

- ✓ Una matrice **n x n** può avere al massimo **n** autovalori (alcuni di essi però possono anche essere uguali);
- ✓ Una matrice di numeri reali può anche avere uno o più autovalori complessi (ma saranno comunque sempre presenti a coppie);
- ✓ La matrice $(\lambda \cdot I - A)$ non è invertibile.

In R:

```
> E<-eigen(nomematrice)
> E$values
> E$vectors
```

Altre proprietà

$$A \cdot \vec{v} = \lambda \cdot \vec{v}$$

Per un set di autovalori DISTINTI di A:

$$\lambda_1 + \lambda_2 + \ldots + \lambda_n$$

il corrispondente set di autovettori:

$$\vec{v}_1 + \vec{v}_2 + \ldots + \vec{v}_n$$

forma una **base** (cioè un insieme di vettori linearmente indipendenti) per uno **spazio a n-dimensioni** in cui qualsiasi vettore espresso come combinazione lineare degli autovettori può esservi rappresentato:

$$\overrightarrow{\text{VettoreGenerico}} = a_1 \cdot \vec{v}_1 + a_2 \cdot \vec{v}_2 + \ldots + a_n \cdot \vec{v}_n$$

Altre proprietà (2)

$$\overrightarrow{\text{VettoreGenerico}} = a_1 \cdot \vec{v}_1 + a_2 \cdot \vec{v}_2 + \ldots + a_n \cdot \vec{v}_n$$

Dato che $A \cdot \vec{v}_j = \lambda_j \cdot \vec{v}_j$ (per j da 1 a n), si può trasformare anche in:

$$A \cdot \overrightarrow{\text{VettoreGenerico}} = a_1 \cdot \lambda_1 \cdot \vec{v}_1 + a_2 \cdot \lambda_2 \cdot \vec{v}_2 + \ldots + a_n \cdot \lambda_n \cdot \vec{v}_n$$

Quindi gli *eigenpairs* **trasformano** una moltiplicazione matrice-vettore in una somma di moltiplicazioni scalare-vettore

In altre parole gli autovettori generano gli "assi" lungo i quali la moltiplicazione della matrice per un vettore generico semplicemente dà un "peso" (*weight*) a quel vettore utilizzando gli autovalori.

L'Analisi delle Componenti Principali (o *Principal Component Analysis*)

- ✓ E' un'analisi statistica non parametrica;
- ✓ Rivela strutture nascoste a livello multidimensionale che si trovano all'interno dei dati;
- ✓ Rivela eventuali RIDONDANZE (o collinearità) tra le variabili (cioè se alcune variabili sono correlate);
- ✓ Le strutture nascoste che vengono rivelate hanno generalmente un numero basso di dimensioni (2 o 3) quindi sono visualizzabili.

Principal Component Analysis (2)

- Per una matrice A (n x m) si può sempre ottenere la matrice trasposta A^T (m x n) scambiando le righe con le colonne;
- Il prodotto $A^T \cdot A$ è una matrice quadrata di dimensioni m x m.

- Se per ogni colonna della matrice A la media di colonna è stata sottratta ad ogni elemento della colonna (rispettivamente), allora nella seguente matrice

$$B = \frac{A^T \cdot A}{n - 1}$$

ogni elemento i,j è la COVARIANZA tra le variabili che si trovano alle colonne i e j della matrice A. Quindi sulla diagonale di B ci sono le varianze di ogni colonna di A.

Var	Cov	Cov
Cov	Var	Cov
Cov	Cov	Var

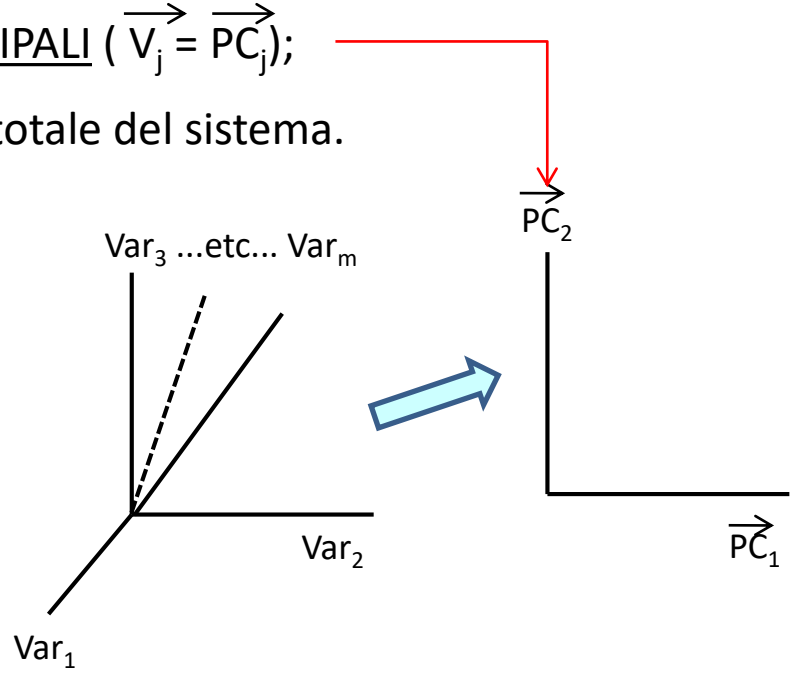
Proprietà di B

$$B = \frac{A^T \cdot A}{n - 1}$$

$$\begin{pmatrix} \text{Var} & \text{Cov} & \text{Cov} \\ \text{Cov} & \text{Var} & \text{Cov} \\ \text{Cov} & \text{Cov} & \text{Var} \end{pmatrix}$$

- Gli autovalori di B sono numeri reali e i corrispondenti autovettori sono ortogonali nello spazio a m dimensioni (cioè generano degli assi);
- La varianza totale del sistema A è data dalla somma degli autovalori di B;
- Gli autovettori di B sono detti COMPONENTI PRINCIPALI ($\overrightarrow{V_j} = \overrightarrow{PC_j}$);
- La direzione v_j spiega la λ_j - frazione della varianza totale del sistema.

Quindi se **alcuni autovalori** spiegano buona parte della **varianza** totale del sistema, basta usare i **relativi autovettori** per rappresentare le osservazioni (righe di A) nel nuovo sistema di coordinate. Il resto di solito è piccola variabilità o "rumore"



La normalizzazione

- Se per ogni colonna della matrice A la media di colonna è stata sottratta ad ogni elemento della colonna (rispettivamente), allora nella seguente matrice

$$B = \frac{A^T \cdot A}{n - 1}$$

$$\begin{pmatrix} \text{Var} & \text{Cov} & \text{Cov} \\ \text{Cov} & \text{Var} & \text{Cov} \\ \text{Cov} & \text{Cov} & \text{Var} \end{pmatrix}$$

ogni elemento i,j è la COVARIANZA tra le variabili che si trovano alle colonne i e j della matrice A. Quindi sulla diagonale di B ci sono le varianze di ogni colonna di A.

Questa è una normalizzazione necessaria perché le componenti principali sono gli autovettori della matrice di covarianza

---

Di solito è opportuno anche utilizzare la deviazione standard per la normalizzazione utilizzando il così detto metodo **Z-score**:

$$Z_{Var(i,j)} = \frac{Var_{(i,j)} - \overline{Var_j}}{\sigma_{Varj}}$$

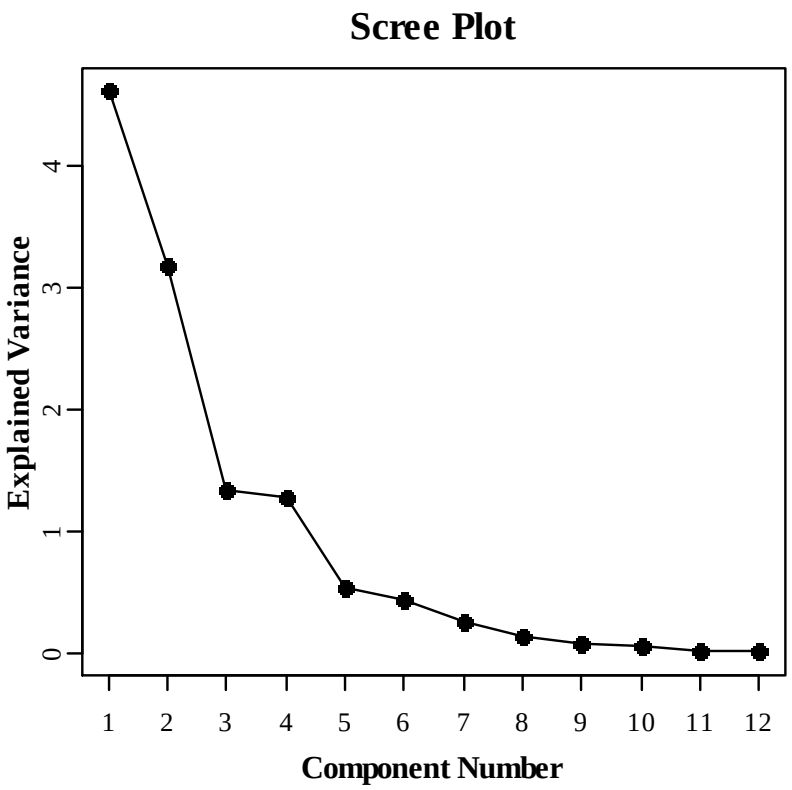
Questo fattore rende **la media di Var = 0**

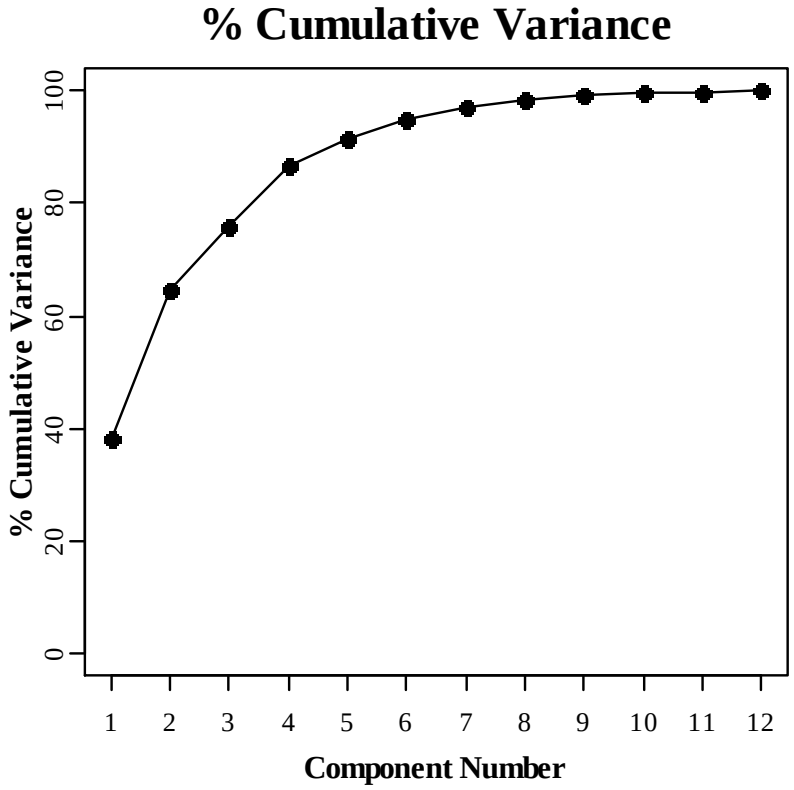
Questo fattore rende la **varianza di Var = 1**

$$\begin{pmatrix} 1 & \text{Cov} & \text{Cov} \\ \text{Cov} & 1 & \text{Cov} \\ \text{Cov} & \text{Cov} & 1 \end{pmatrix}$$

Risultati della PCA: i grafici della varianza spiegata

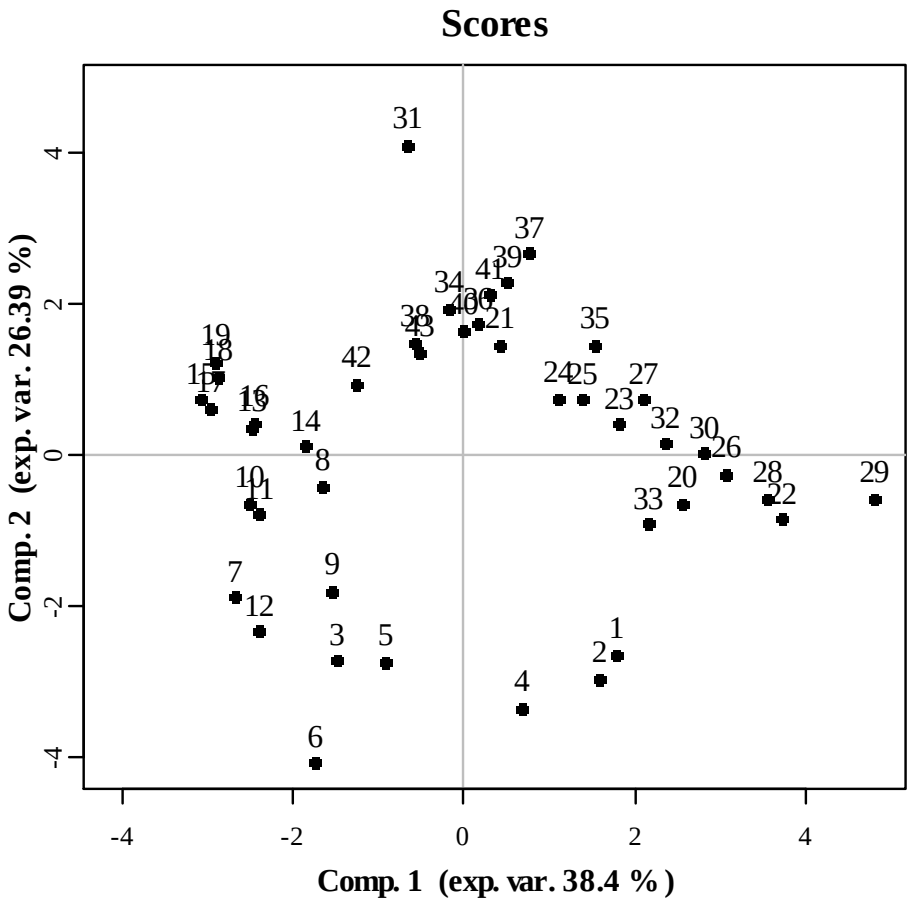
Gli autovalori di ogni componente principale (PC) rappresentano la varianza spiegata da quella componente

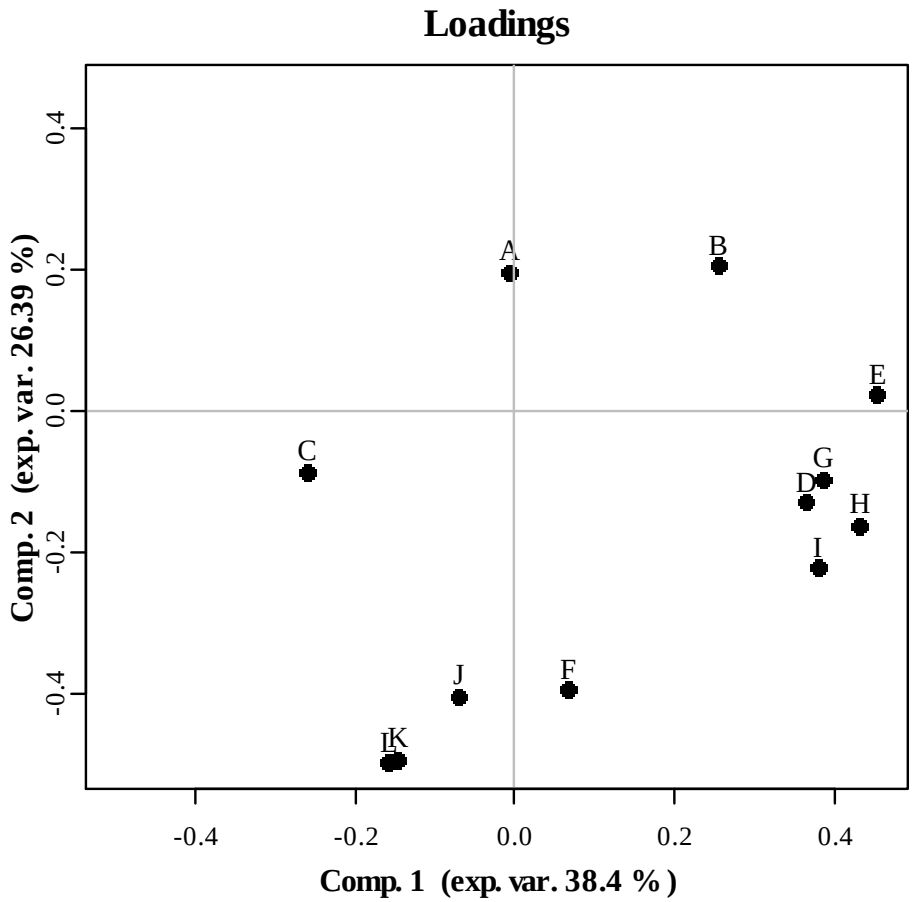




Risultati della PCA: i grafici degli Score e dei Loading

In un grafico a 2 dimensioni si pongono in ascissa una PC e in ordinata un'altra PC (di solito si inizia da 1 e 2, che rappresentano la maggior parte della varianza del set di dati)

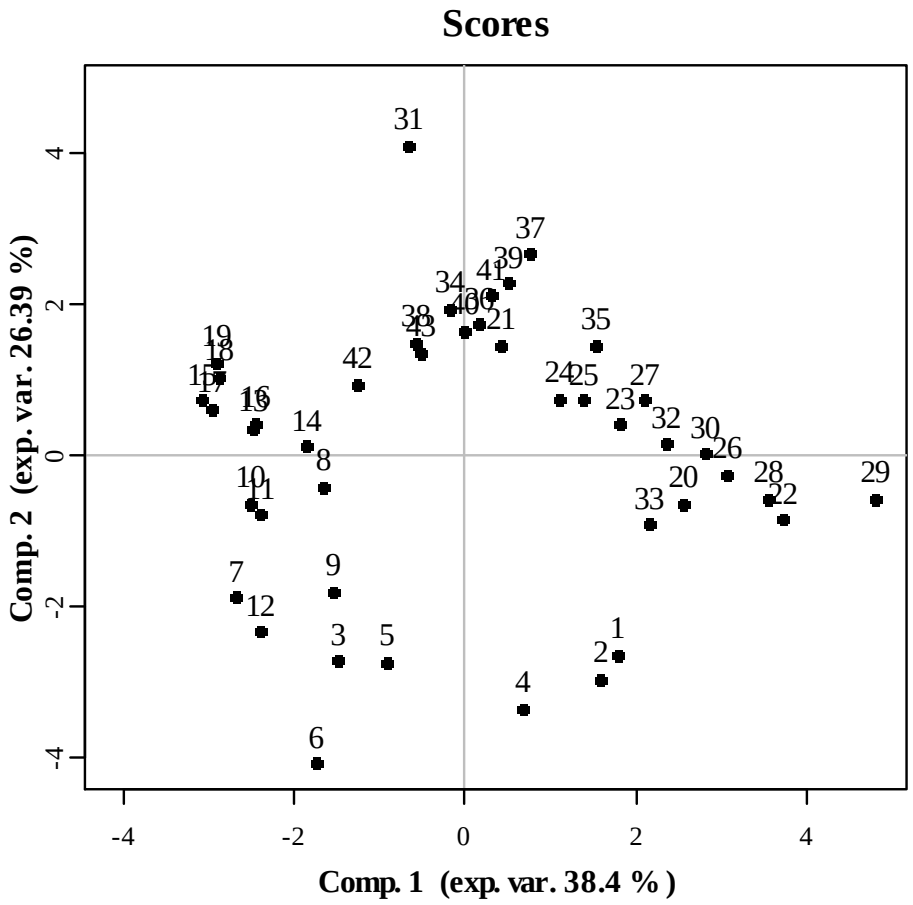




SEGUE

Risultati della PCA: i grafici degli Score e dei Loading (2)

Il grafico degli **Score** rappresenta le osservazioni nel nuovo sistema di variabili



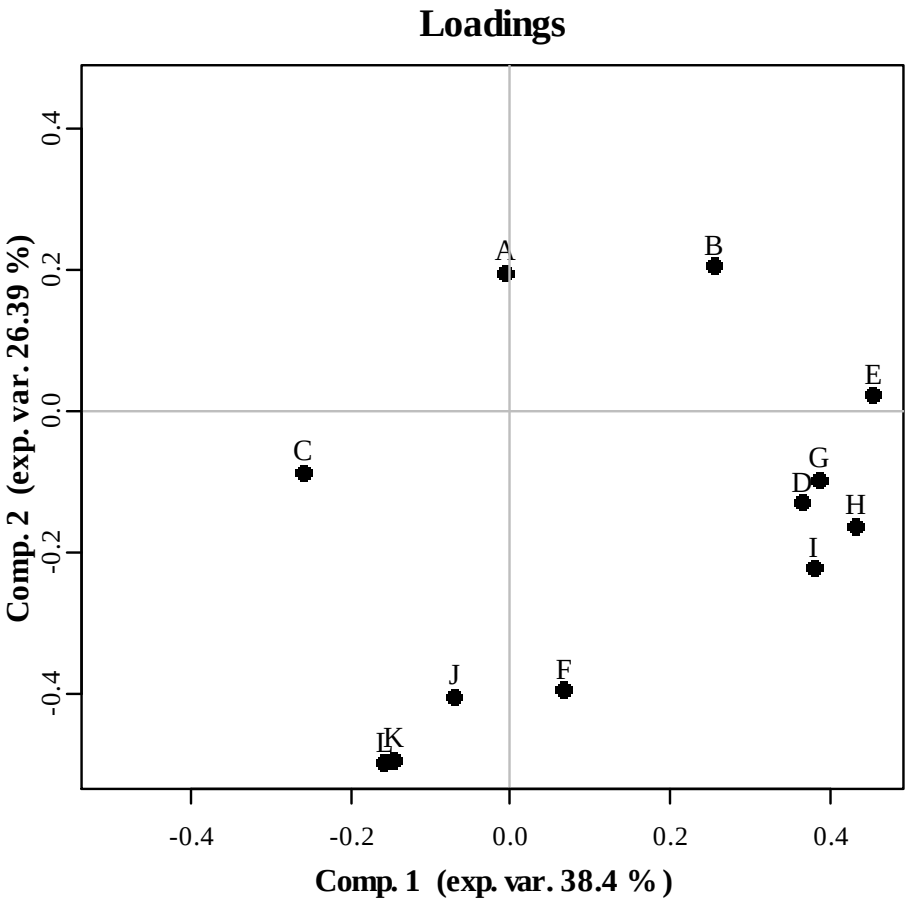
- I punti (quindi le osservazioni) che sono prossime all'origine sono osservazioni che hanno valori prossimi alla media per la maggior parte delle variabili;
- Le osservazioni che si trovano molto lontano dall'origine possono essere o dati estremi del dataset o *outlier*;
- Osservazioni simili si trovano vicine nel nuovo spazio ovvero osservazioni dissimili si trovano lontane tra loro.

SEGUE

Risultati della PCA: i grafici degli Score e dei Loading (3)

Il grafico dei **Loading** rappresenta i coefficienti (pesi) della combinazione lineare delle variabili originali che servono per generare le componenti principali.

$$\vec{v}_j = \overrightarrow{PC_j} = w_1 \cdot \overrightarrow{Var_1} + w_2 \cdot \overrightarrow{Var_2} + \ldots + w_m \cdot \overrightarrow{Var_m}$$

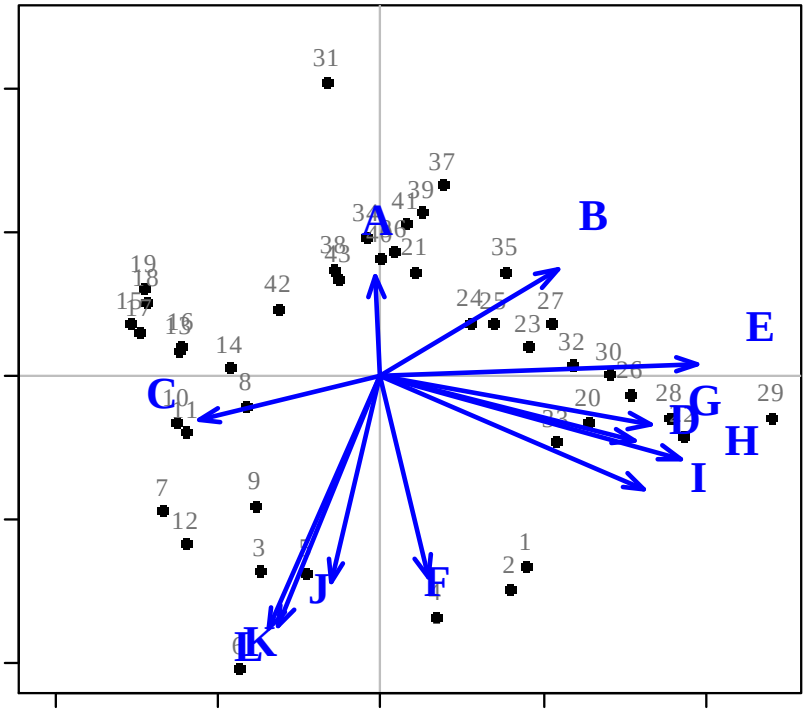


- Variabili originali che contribuiscono poco alla direzione di una componente principale mostrano loading prossimi a zero per quella componente;
- Variabili originali che contribuiscono in modo simile alla direzione di una componente principale sono generalmente correlate e mostrano loading di valore circa uguale;
- Variabili originali correlate positivamente si trovano vicine nel grafico, mentre variabili correlate negativamente si trovano opposte tra loro rispetto ad una diagonale.

Fonti: Hartmann, K., Krois, J., Waske, B. (2018): E-Learning Project SOGA: Statistics and Geospatial Data Analysis. Department of Earth Sciences, Freie Universitaet Berlin

Risultati della PCA: il biplot

E' dato dalla sovrapposizione del grafico dei **Loading** a quello degli **Score**, centrati sull'origine. I **Loading** vengono rappresentati da frecce che si dipartono dall'origine.

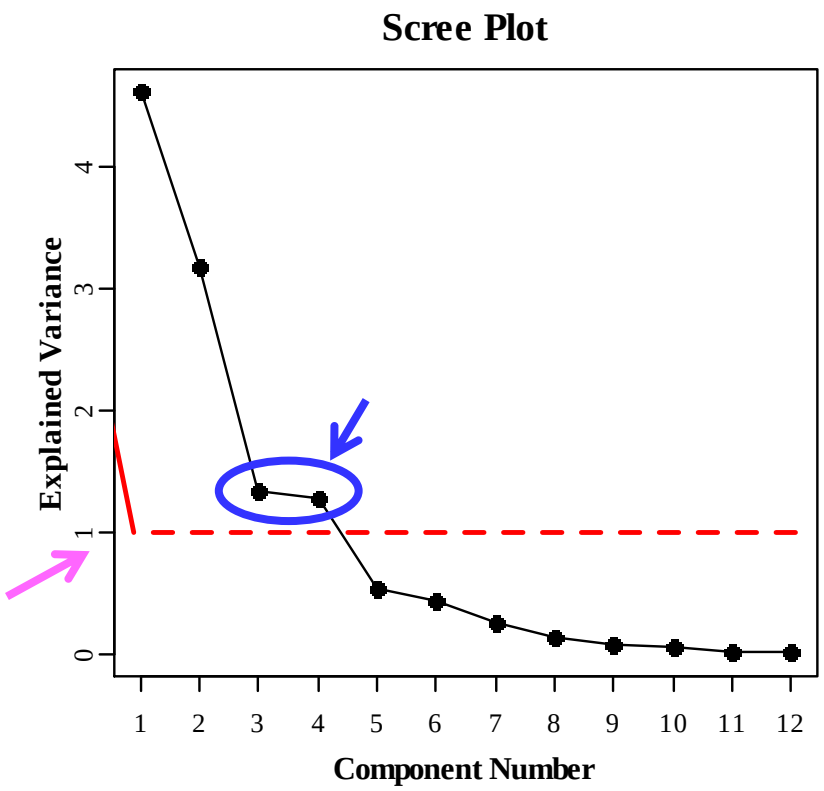


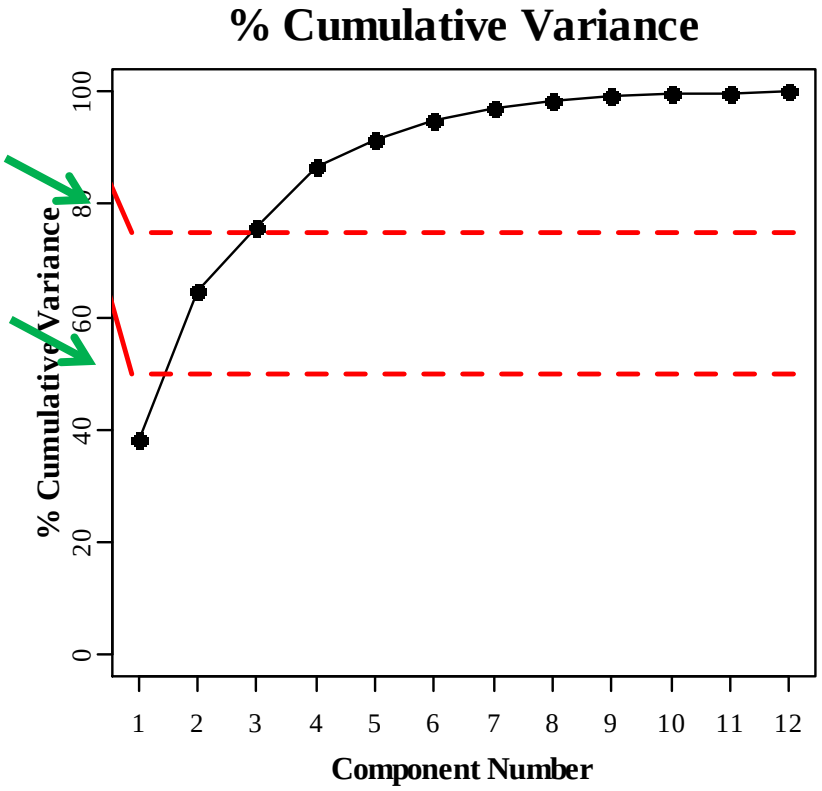
Inoltre si visualizza quali variabili originali "spostano" di più quali osservazioni dall'origine (cioè dal "baricentro" del dataset)

- Più un vettore è <u>parallelo</u> all'asse di una PC più la variabile originale associata <u>contribuisce pressoché esclusivamente a quella PC</u>;
- <u>Più è lungo</u> un vettore <u>maggiore è la variabilità della variabile originale associata</u> che viene spiegata dalle due PC su cui è costruito il grafico;
- Di conseguenza <u>se un vettore è corto</u> vuol dire che la <u>variabilità</u> della variabile originale associata <u>viene rappresentata meglio da altre PC</u> non utilizzate per costruire il grafico;
- <u>L'angolo</u> tra due vettori rappresenta la relativa <u>correlazione</u> delle variabili originali associate:
  - angolo ~ 0° → alta correlazione positiva;
  - angolo ~ 90° → pressoché non correlate;
  - angolo ~ 180° → alta correlazione negativa

Fonti: Hartmann, K., Krois, J., Waske, B. (2018): E-Learning Project SOGA: Statistics and Geospatial Data Analysis. Department of Earth Sciences, Freie Universitaet Berlin

Quante PC utilizzare per rappresentare il dataset?





1. Esplorazione visiva dello scree plot: valutare la presenza di "un gomito" nel grafico

2. Valutazione di quante PC spiegano almeno il 50% (o il 75%) della varianza del dataset

3. Valutazione di quante PC mostrano varianza > 1 (poichè quelle < 1 tecnicamente spiegano meno varianza di una variabile originale) - *Criterio di Kaiser-Guttman*

Dati mancanti (NA) e dati < Limit of Detection (LOD), che fare?

➢ Dati mancanti:

Soluzione 1: eliminazione — **In R:** `na.omit(nomeoggetto)`

Soluzione 2: sostituzione con dati modellati

Esempi:

EM-PCA (iterative PCA)
NIPALS (non-linear iterative partial least squares)

In R:

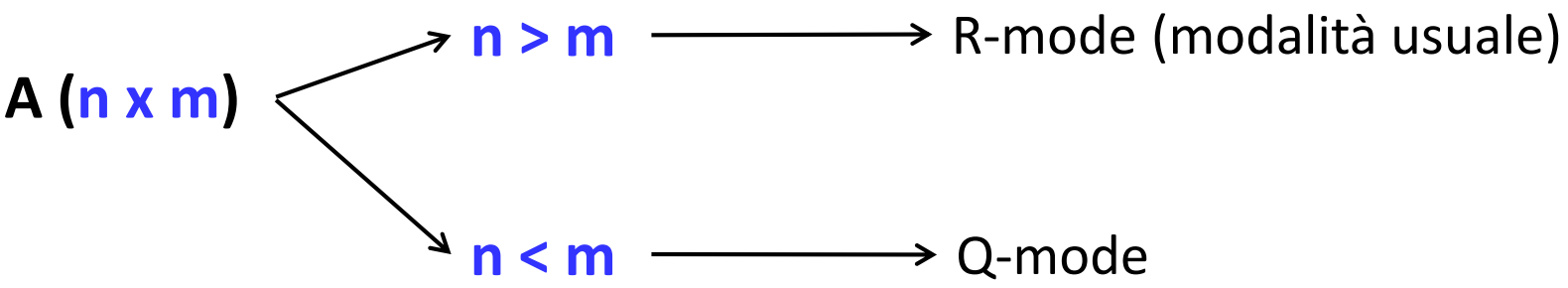
- Il pacchetto "**missMDA**" con le sue funzioni consente di gestire i dati mancanti con diversi algoritmi;
- Il pacchetto "**pcaMethods**" (in Bioconductor packages) consente di calcolare una PCA con diversi algoritmi che gestiscono dati mancanti.

➢ Dati < LOD:

Suggerimento: non conviene porre = a 0 (zero) questi dati, ma conviene sostituirli o con il valore del LOD o con un valore più piccolo del LOD che sia approssimabile a zero ma che non sia zero (es. 10^{-5} o 10^{-7} , etc...)

Fonti: S. Dray; J. Josse "Principal component analysis with missing values: a comparative survey of methods",Plant Ecol (2015) 216:657–667,DOI 10.1007/s11258-014-0406-z

Se ci sono più variabili che osservazioni?



Procedere nel modo seguente:

1 - Scalare e centrare il dataset;

In R: **t(nomeoggetto)**

2 - Generare la matrice trasposta del dataset (ovvero scambiare righe e colonne);

3 - Calcolare la PCA

I grafici degli **Score** rappresentano i **Loading** e viceversa, l'interpretazione è simile a quella già vista