Carnegie Mellon University

MSE

Data Analytics for Materials Science

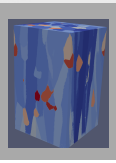
27-737

A.D. (Tony) Rollett, Amit Verma, Richard A. LeSar (Iowa State Univ.)

Dept. Materials Sci. Eng., Carnegie Mellon University

Principal Component Analysis (PCA)

Lecture 8, part 1



Revised: 2nd Mar. 2021

http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/118-principal-component-analysis-in-r-prcomp-vs-princomp/

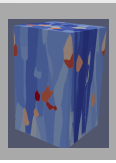
packages: FactoMineR, factoextra, ggplot2, scatterplot3d, yacca, car, CC, cca

https://www.rdocumentation.org/packages/factoextra/versions/1.0.7

https://www.benjaminbell.co.uk/2018/02/principal-components-analysis-pca-in-r.html
https://www.benjaminbell.co.uk/2018/03/principal-components-analysis-pca-in-r-part-2.html

Chapter 9 in Jobson, Vol. 2.

Hastie et al. only refer to PCA in passing …

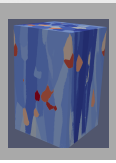


A common type of dataset: 44 semiconductor compounds and 6 descriptors.

We have seen what regression can do to help understand the data.

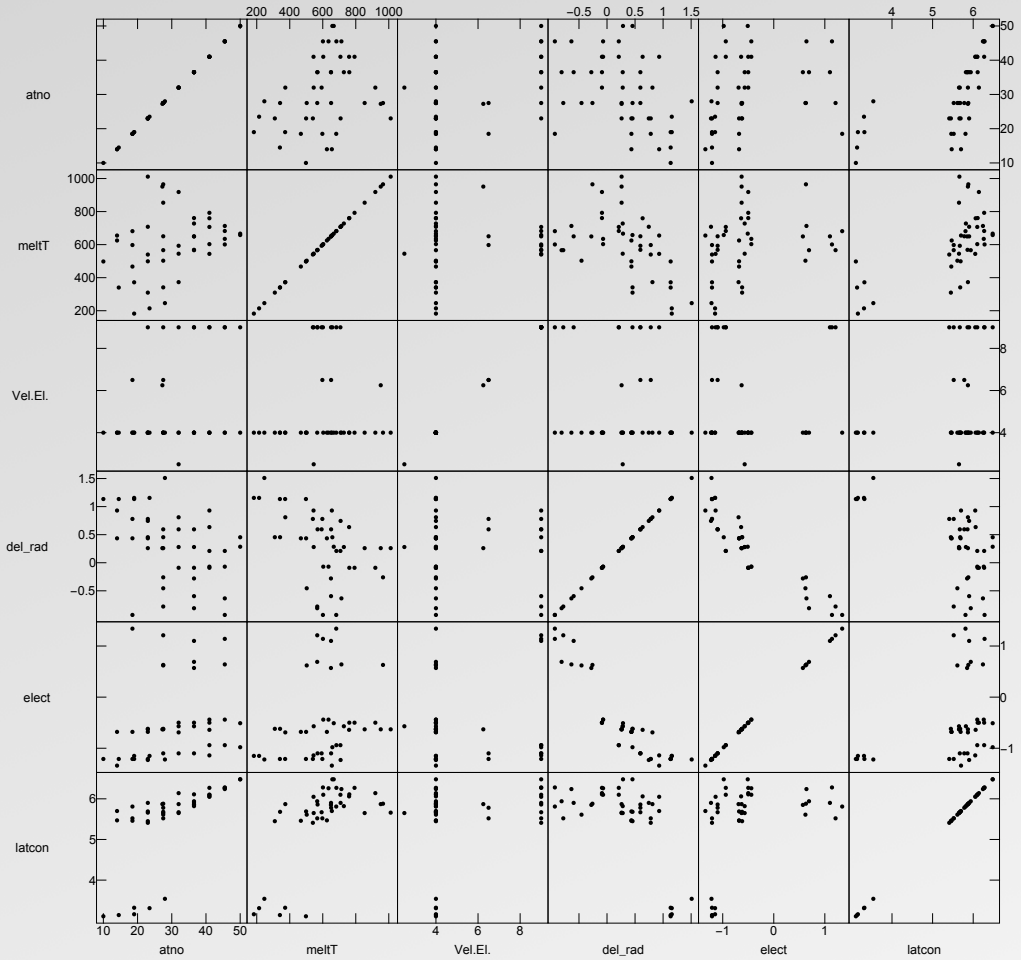
What other tools do we have?

	atomic number	melting pt. (C)	# valence e-	Δ in radii	electro-negativity	lattice const (Å)		atomic number	melting pt. (C)	# valence e-	Δ in radii	electro-negativity	lattice const (Å)
AlN	10	498	4	1.135	-1.21	3.11	$(ZnMg)_{0.5}$ S	18.5	598	6.5	0.78	-1.21	5.52
AlP	14	625	4	0.435	-0.68	5.47	$(SSe)_{0.5}$ Mg	18.5	682	4	-0.93	1.34	5.81
AlAs	23	1012	4	0.26	-0.63	5.66	$(SSe)_{0.5}$ Z	27.5	567	9	-0.78	1.21	5.52
AlSb	32	919	4	-0.09	-0.5	6.14	$(ZnMg)_{0.5}$ Se	27.5	651	6.5	0.595	-1.1	5.78
GaN	19	183	4	1.155	-1.15	3.16	$(ZnCd)_{0.5}$ Se	36.5	569	9	0.595	-1.1	5.86
GaP	23	310	4	0.455	-0.62	5.45	$(SeTe)_{0.5}$ Zn	36.5	650	9	-0.595	1.1	5.9
GaAs	32	545	2.5	0.28	-0.57	5.65	$(SeTe)_{0.5}$ Cd	45.5	601	9	-0.93	1.14	6.28
GaSb	41	603	4	-0.07	-0.44	6.1	$(ZnCd)_{0.5}$ Te	45.5	683	9	0.21	-0.94	6.27
InN	28	246	4	1.51	-1.22	3.54	$(AlGa)_{0.5}$ P	18.5	467	4	0.435	-0.68	5.46
InP	32	373	4	0.81	-0.69	5.87	$(PAs)_{0.5}$ Ga	27.5	503	4	-0.455	0.62	5.61
InAs	41	760	4	0.635	-0.64	6.06	$(AlGa)_{0.5}$ As	27.5	854	4	0.26	-0.63	5.65
InSb	50	667	4	0.285	-0.51	6.48	$(GaIn)_{0.5}$ P	27.5	341	4	0.455	-0.62	5.68
ZnS	23	540	9	0.78	-1.21	5.41	$(AlIn)_{0.5}$ P	23	499	4	0.435	-0.68	5.69
ZnSe	32	593	9	0.595	-1.1	5.67	$(AlG)_{0.5Zn}$As	27.5	965	4	-0.26	0.63	5.88
ZnTe	41	707	9	0.21	-0.94	6.1	$(AlZn)_{0.5}$ As	27.25	951	6.25	0.26	-0.63	5.87
CdSe	41	544	9	0.93	-1.14	6.05	$(AsSb)_{0.5}$ Ga	36.5	650	4	-0.28	0.57	5.85
CdTe	50	658	9	0.454	-0.98	6.48	$(AlGa)_{0.5}$ Sb	36.5	761	4	-0.09	-0.5	6.11
MgS	14	655	4	0.93	-1.34	5.7	$(GaIn)_{0.5}$ Sb	36.5	728	4	0.28	-0.57	5.82
MgSe	23	708	4	0.745	-1.23	5.9	$(PAs)_{0.5}$ In	36.5	566	4	-0.81	0.69	5.94
$(AlGa)_{0.5}$ N	14.5	340	4	1.135	-1.21	3.14	$(AlIn)_{0.5}$ Sb	41	793	4	-0.09	-0.5	6.27
$(AlIn)_{0.5}$ N	19	372	4	1.135	-1.21	3.32	$(GaIn)_{0.5}$ Sb	45.5	635	4	-0.07	-0.44	6.26
$(GaIn)_{0.5}$ N	23.5	214	4	1.155	-1.15	3.31	$(AsAb)_{0.5}$ In	45.5	713	4	-0.635	0.64	6.24



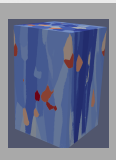
dataset-semiconductors-from-KrishnaRajan-talk-on-PCA-RALeSar.xlsx

Semiconductor compounds: Courtney of Krishna Rajan

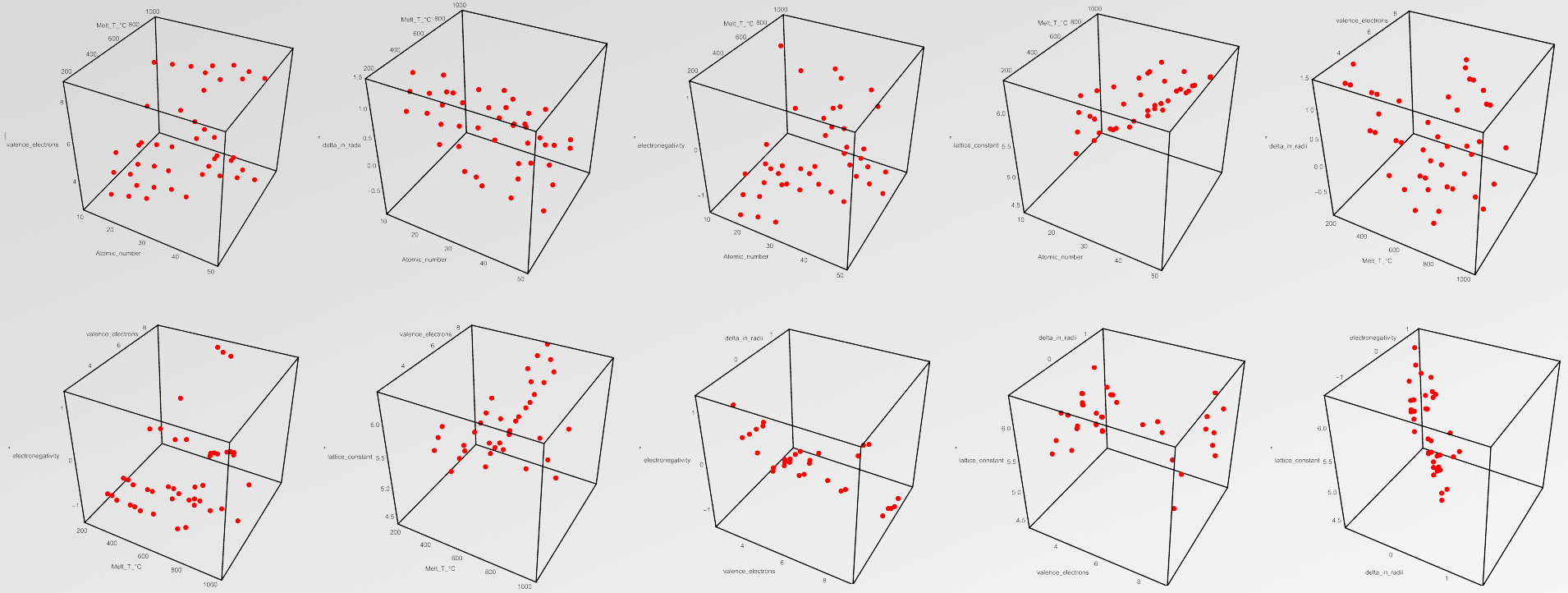


scatterplot matrix

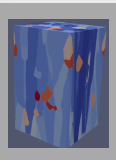
- no information about multidimensional correlations



Scatterplot matrix



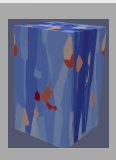
3D plots are even less useful



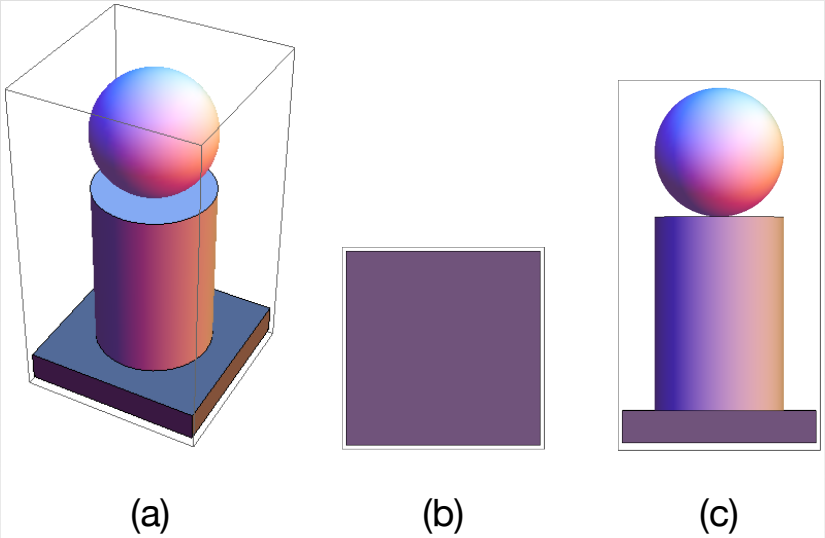
Triples of variables

2D and 3D plots describe the data, but do not tell us much about the 6-dimensional data

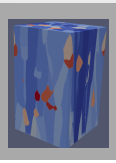
- we can try to *reduce the dimensionality* of the data to make it more accessible for analysis
  - the goal will be to reduce the dimensionality without losing the content of all the data
- we focus today on *principal component analysis*
- For those who are familiar with *eigenanalysis*, we apply this very standard numerical analysis to the covariance (correlation) matrix to rotate and diagonalize it. Read up on the method on Wikipedia: Eigenvalues and eigenvectors



In PCA, we find a reduced-dimension representation of the data that *maximizes the variance of the data.*



In this simple picture, (c) provides more more of the variance, and thus more information about the data, information than (b). In effect, we project the data shown in (a) in two different directions to get (b) and (c). PCA is, however, more complex than just projection.



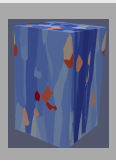
mean: $\bar{x} = \frac{1}{N}\sum_{n=1}^{N} x_i$

bias-corrected variance: $S_x = \frac{1}{N-1}\sum_{n=1}^{N} \left(x_i - \bar{x}\right)^2$

bias-corrected standard deviation: $\sigma_x = \sqrt{S_x}$

bias-corrected covariance: $C_{ij} = \frac{1}{N-1}\sum_{n=1}^{N} \left(x_{in} - \bar{x}_i\right)\left(x_{jn} - \bar{x}_j\right)$

Here, the means have been subtracted but the range of each variable can be very different, i.e., the variance can differ between each variable (which may be a problem).



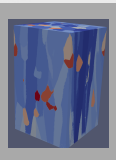
Statistics relative to a sample mean $\bar{x}$

Our focus will be on maximizing the total variance of the data captured in a reduced dimensional representation.

We need not only the variance of individual types of data, but also whatever *correlations* exist between data types, do the data types track with each other.

However, we need to rescale the data so we can capture those correlations accurately — the correlation matrix we described in Lecture 5.

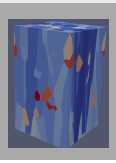
We will build our (eigenanalysis) solution on the *principal components* of either the covariance (means subtracted) or the correlation (means subtracted *and* normalized by variance) matrix.



To put all variables on the same scale and to eliminate issues with the different units in each data type, we *autoscale* the data:

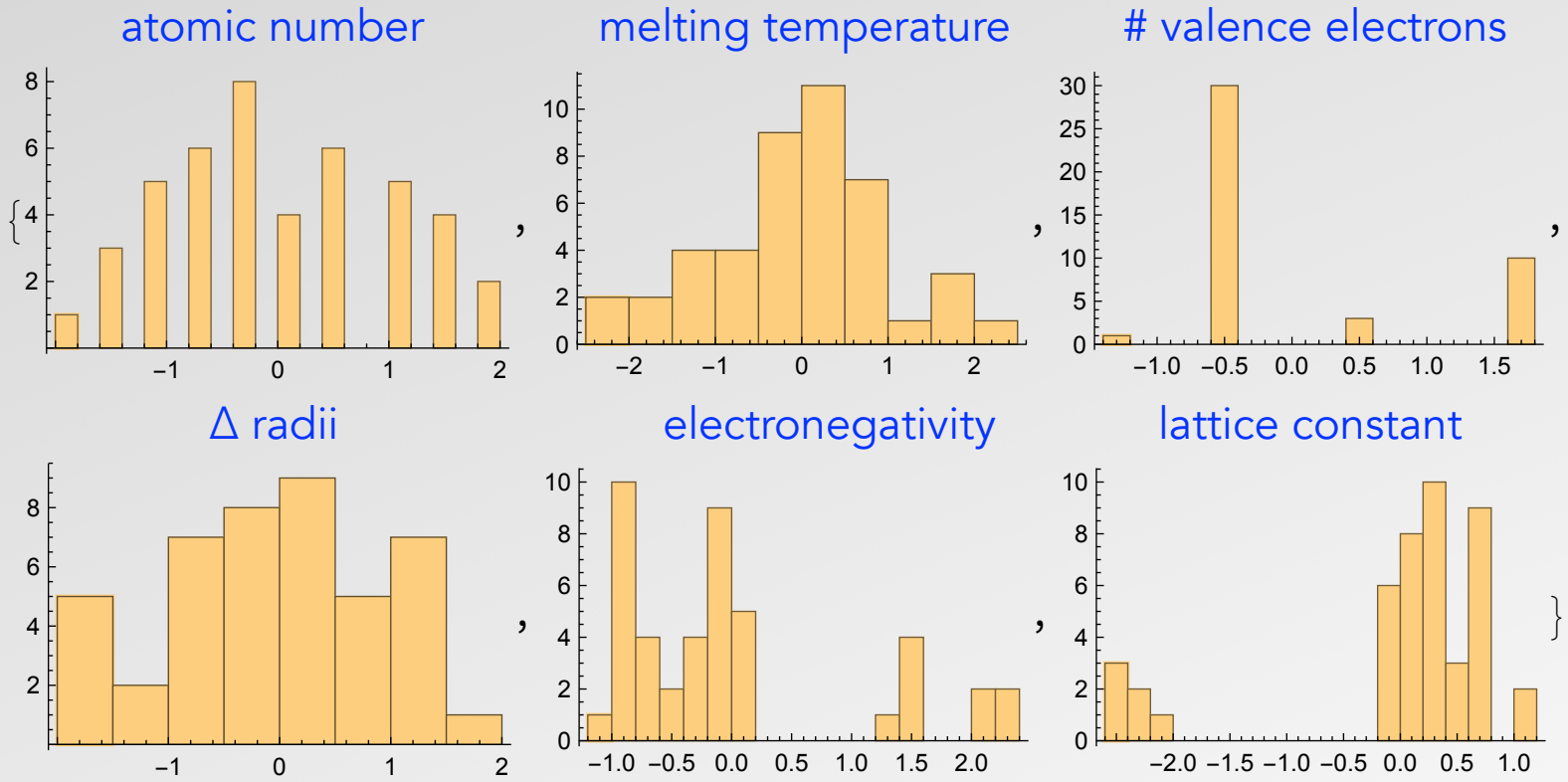
$X'_{ki} = \dfrac{X_{ki} - \overline{X_i}}{\sigma_i}$, for each data entry with i = the type of data (e.g., *atomic number*).

- The vector of values of X'_{ki} will be called $\mathbf{X}'_i$ and we have: $\overline{\mathbf{X}'} = 0$ and $\sigma_{\mathbf{X}'_i} = 1$
- Create the data matrix as: $\mathbf{A}_N = \left[\mathbf{X}'_1 \quad \mathbf{X}'_2 \quad \mathbf{X}'_3 \quad \mathbf{X}'_4 \quad \mathbf{X}'_5 \quad \mathbf{X}'_6\right]$ in which each $\mathbf{X}'_i$ is a 44 component column vector of the autoscaled data type $\mathbf{X}'_i$.
- Autoscaling ensures that the variances are weighted the same for each data type and avoids complications of having various units.

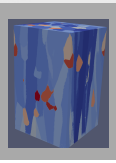


Autoscaling the data

$\mathbf{A}_N =$



Autoscaling ensures that the variances are weighted the same for each data type and are unitless.



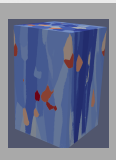
Autoscaled data

Calculate the *correlation matrix* (measure of the variance between variables) for the autoscaled data $X'_{ki} = \dfrac{X_{ki} - \overline{X_i}}{\sigma_i}$ with $\overline{X'_i} = 0$ and $\sigma_{X'_i} = 1$

$$C_{ij} = \frac{1}{N-1}\sum_{k=1}^{N} X'_{ki}X'_{kj} = \frac{1}{N-1}\sum_{k=1}^{N} \frac{\left(X_{ki} - \overline{X_i}\right)\left(X_{kj} - \overline{X_j}\right)}{\sigma_i \sigma_j}$$

$$C_{ii} = \frac{1}{N-1}\sum_{k=1}^{N} X'^{2}_{ki} = \frac{1}{N-1}\sum_{k=1}^{N} \frac{\left(X_{ki} - \overline{X_i}\right)^2}{\sigma_i^2} = \frac{S_i}{S_i} = 1$$

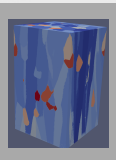
C is a measure of the correlation between the variables. We can also construct a similar *covariance matrix* that omits the normalization by variance (next slide). In general, the values in the the various columns/rows of a correlation matrix look rather similar to each other because of the scaling.



Calculate the *covariance matrix* (measure of the variance between variables) for the autoscaled data $X'_{ki} = X_{ki} - \overline{X}_i$ with $\overline{X}'_i = 0$.

$$C_{ij} = \frac{1}{N-1}\sum_{k=1}^{N} X'_{ki}X'_{kj} = \frac{1}{N-1}\sum_{k=1}^{N}\left(X_{ki} - \overline{X}_i\right)\left(X_{kj} - \overline{X}_j\right)$$

C is a measure of the covariance between the variables. In fact, we already showed this formula a few slides back so this is a reminder! In general, the values in the the various columns/rows of a covariance matrix do not resemble each other because of the different scales of each variable.

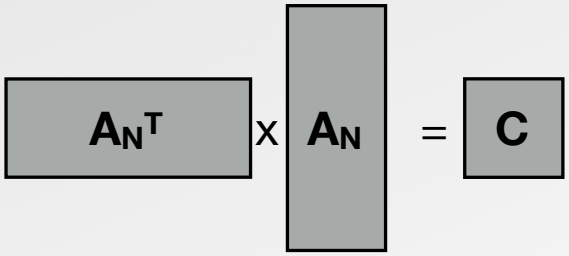


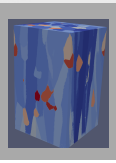
While we could sum the data as shown on the previous slide, it is actually easier to use what we've learned about matrices:

Since $\mathbf{A}_N = \begin{bmatrix} \mathbf{X}_1' & \mathbf{X}_2' & \mathbf{X}_3' & \mathbf{X}_4' & \mathbf{X}_5' & \mathbf{X}_6' \end{bmatrix}$, we can calculate $\mathbf{C}$ as

$$\mathbf{C} = \frac{\mathbf{A}_N^T \mathbf{A}_N}{N-1}$$

$\mathbf{A}_N$ is has dimensions of $p \times N$ and $\mathbf{A}_N^T$ has dimensions of N X p. The product has dimensions of p X p.



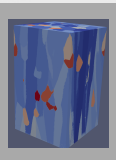


The covariance matrix for this dataset is:

	at.no.	MeltT	# val. e-	Δ radii	elecneg	latcon
at.no.	1.	0.301065	0.333507	−0.423666	0.260447	0.644653
MeltT	0.301065	1.	0.0800864	−0.475718	0.257429	0.649356
# val. e-	0.333507	0.0800864	1.	−0.102481	0.0249477	0.258756
Δ radii	−0.423666	−0.475718	−0.102481	1.	−0.91109	−0.615628
elecneg	0.260447	0.257429	0.0249477	−0.91109	1.	0.367428
latcon	0.644653	0.649356	0.258756	−0.615628	0.367428	1.

Note that since $\mathbf{C}$ is symmetric and has a unit values along the diagonal:

- the eigenvectors of $\mathbf{C}$ are orthogonal and normalized (unit vectors)
- the C_{ij} are the correlations between data types i and j



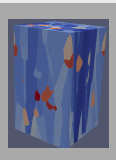
To proceed, we will use *Principal Component Analysis* (PCA), e.g., in mechanical engineering it is called *Proper Orthogonal Decomposition*. It uses *eigenanalysis*, which is a well-known tool in many other names in other fields and is sometimes referred to as spectral decomposition.

PCA is an orthogonal linear transformation that transforms the data to a new basis such that the greatest variance by some scalar projection of the data comes to lie on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on.[9]

In essence, PCA fits a *p*-dimensional ellipsoid to the data, in which each axis is a principal component of the covariance matrix $\mathbf{C}$ and whose length is proportional to the variance of the data in that direction.

The principal components (PCs) are the *eigenvectors* of $\mathbf{C}$ (which are unit vectors because $\mathbf{C}$ is an orthonormal symmetric matrix) multiplied by the associated *eigenvalue*. The eigenvalues of $\mathbf{C}$ are the variances associated each PC.

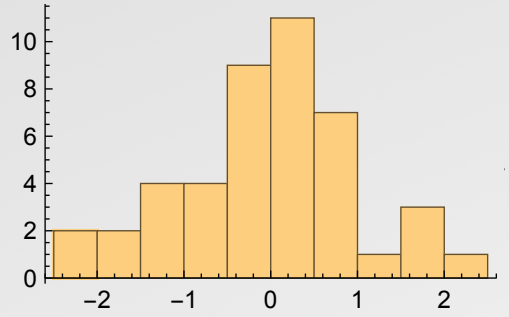
Most often, we operate on the (auto-scaled) *correlation matrix* as opposed to the *covariance* matrix (but the latter is possible: just be careful about numerical issues).



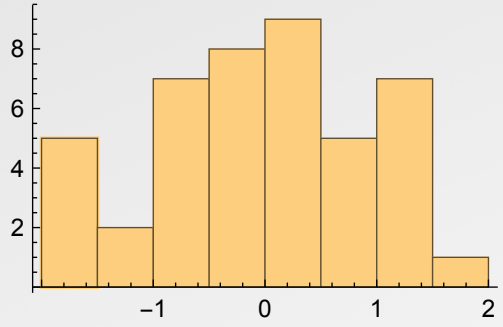
As an example, consider the correlation matrix between two of the data types: “melting temperature" (X_2') and “difference in radii” (X_4').

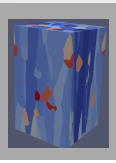
The variation in the individual datasets can be seen in the histograms:

melting temperature



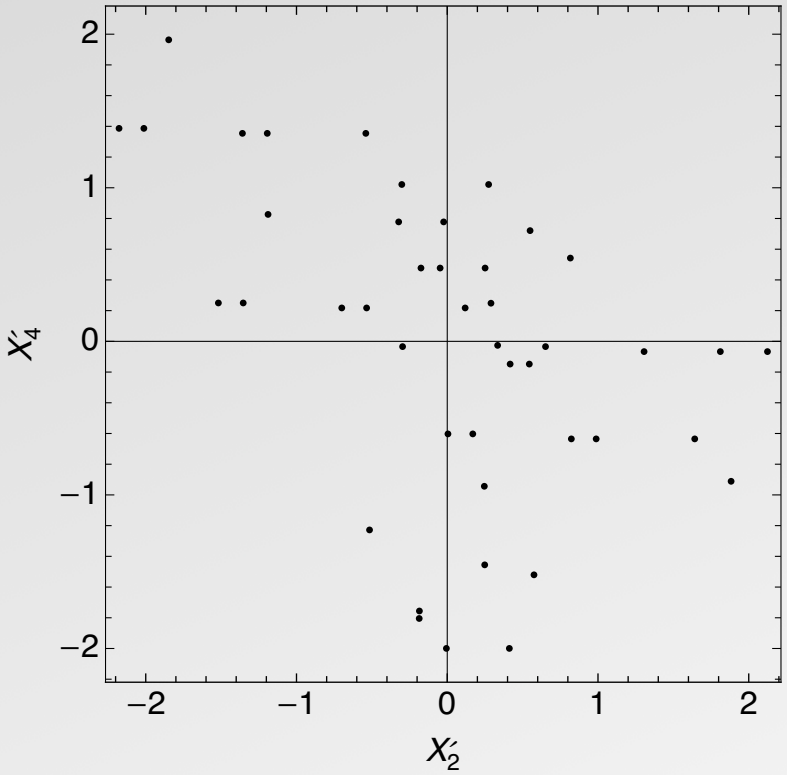
difference in radii





Eigenvalues and eigenvectors: example

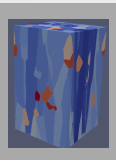
The correlation between the data types is apparent in a plot of X_4' (difference in radii) versus X_2' (melting temperature)



There is a clear relationship between these datatypes, with a general trend along a line with a slope of -1 through the origin.

From the correlation matrix, we can read off the covariance between these two variables as

$C_{24} = -0.475718.$



Eigenvalues and eigenvectors: example

The net correlation matrix for these two data types (i.e., just this pair of variables, to make it a 2D example) thus looks like

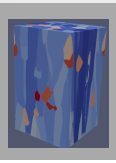
$$\mathbf{C}' = \begin{bmatrix} 1 & C_{24} \\ C_{24} & 1 \end{bmatrix} \qquad C_{24} = -0.475718$$

The eigensystem equation is $\mathbf{C}'\mathbf{v} = \lambda\mathbf{v}$, which we start by rearranging to get $\left(\mathbf{C}' - \lambda\mathbf{I}\right)\mathbf{v} = 0$ and then solve using

$$\det\left(\mathbf{C}' - \lambda\mathbf{I}\right) = 0$$

$$\begin{vmatrix} 1-\lambda & C_{24} \\ C_{24} & 1-\lambda \end{vmatrix} = (1-\lambda)^2 - C_{24}^2 = 0$$

$$\Rightarrow \lambda_1 = 1 - C_{24} \text{ and } \lambda_2 = 1 + C_{24}$$



Eigenvalues and eigenvectors: example

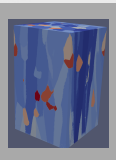
Inserting the values for λ into $\mathbf{C}'\mathbf{v} = \lambda\mathbf{v}$ and solving for the components of $\mathbf{v}$, we find:

$$\lambda_1 = 1.47572 \text{ with } \hat{\mathbf{v}}_1 = \begin{bmatrix} -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \end{bmatrix}$$

$$\lambda_2 = 0.524282 \text{ with } \hat{\mathbf{v}}_2 = \begin{bmatrix} -\frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} \end{bmatrix}$$

Note that $\mathbf{v}_1 \cdot \mathbf{v}_2 = 0$

The eigenvectors form an orthonormal basis.



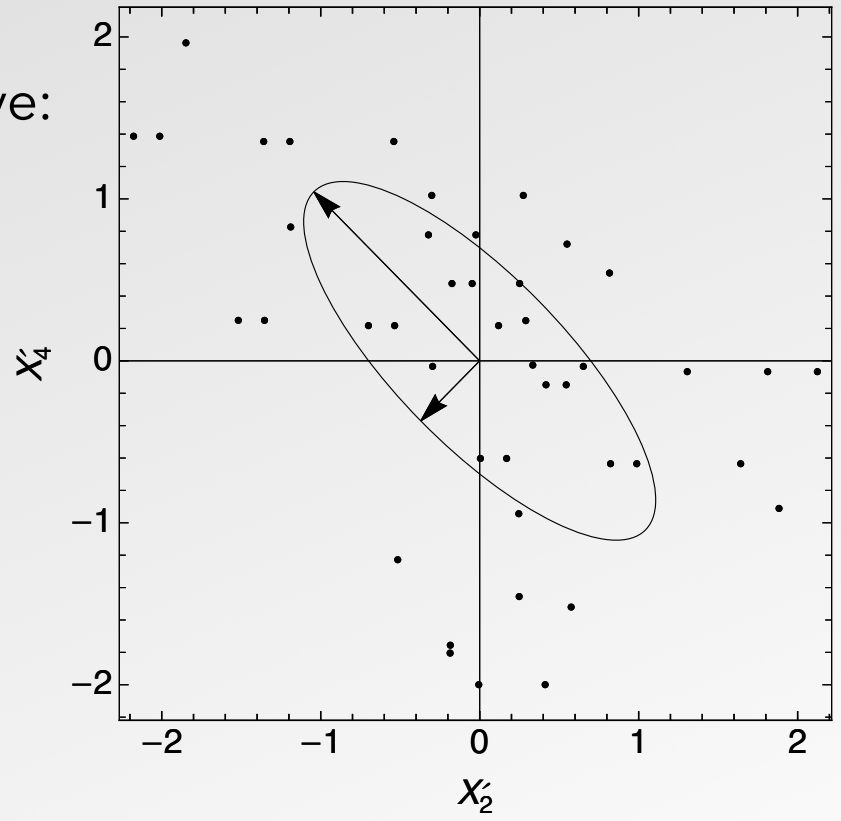
Eigenvalues and eigenvectors: example

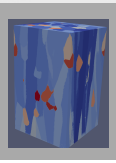
The vectors $\mathbf{V}_i = \lambda_i \hat{\mathbf{v}}_i$ lie along the eigenvectors and have the magnitude of the eigenvalues.

Adding them to the plot of the data we have:

$\mathbf{V}_1$ points along the direction of the greatest variance. $\mathbf{V}_2 \perp \mathbf{V}_1$ and points in the direction of the next biggest variance.

The variance varies along an ellipse aligned along $\mathbf{V}_1$ with semi major axis of λ_1 and semi minor axis of λ_2.





Eigenvalues and eigenvectors

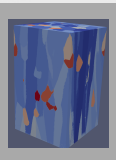
We now find the eigenvectors and eigenvalues of the covariance matrix **C**: the *principal components*, which create an orthogonal basis.

From the correlation matrix given above, we find:

$$\lambda = \begin{pmatrix} 3.06229 \\ 1.21896 \\ 0.870424 \\ 0.586331 \\ 0.220362 \\ 0.0416367 \end{pmatrix} \qquad \mathbf{P} = \begin{pmatrix} -0.393417 & -0.391968 & -0.0774367 & 0.706188 & -0.431937 & -0.0178061 \\ -0.38676 & -0.0474299 & 0.654928 & -0.451536 & -0.457014 & -0.0805578 \\ -0.170358 & -0.644791 & -0.527402 & -0.525786 & -0.0246437 & -0.00331295 \\ 0.506673 & -0.351709 & 0.203926 & 0.0689972 & -0.0348986 & -0.756319 \\ -0.411283 & 0.508746 & -0.407694 & -0.0727362 & -0.107181 & -0.623724 \\ -0.49066 & -0.214087 & 0.281282 & 0.104463 & 0.768931 & -0.179255 \end{pmatrix}$$

The eigenvectors are orthonormal: $\mathbf{P}^T\mathbf{P} = \mathbf{I}$

Note: the sum of the eigenvalues = the number of data types: $\sum_{i=1}^{p} \lambda_k = p$



PCA on the full dataset

The eigenvalues measure the amount of variance within each principal component.

$$\boldsymbol{\lambda} = \left[3.06229,\ 1.21896,\ 0.870424,\ 0.586331,\ 0.220362,\ 0.041637\right]$$

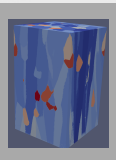
As noted $\sum_{i=1}^{6} \lambda_k = 6.$

We define the *fractional variance* of each principal component as (its eigenvalue normalized by the sum of the eigenvalues):

$$\lambda_i^f = \frac{\lambda_i}{\sum_{i=1}^{6} \lambda_k} = \frac{\lambda_i}{6}$$

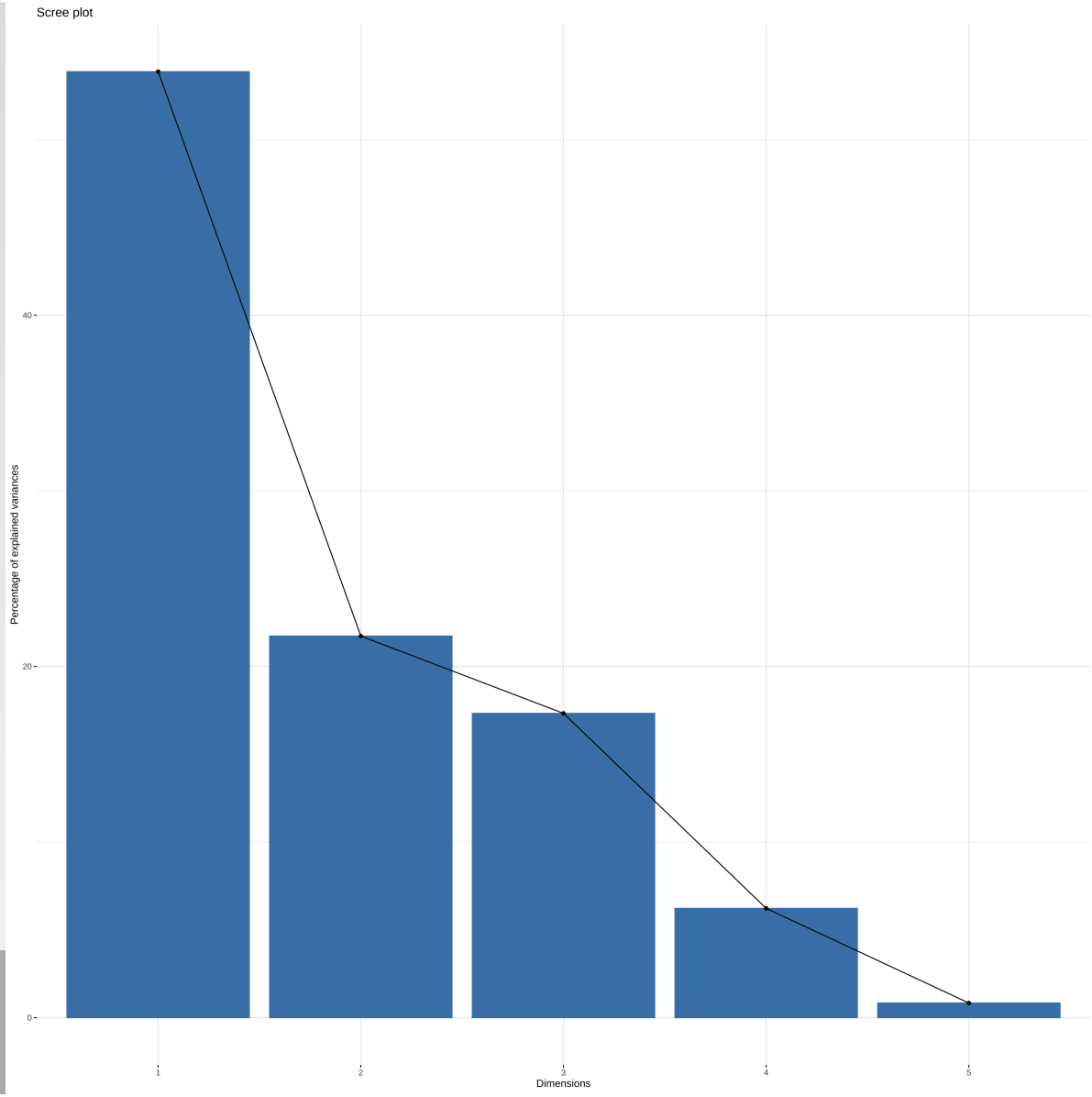
The fractional variance is:

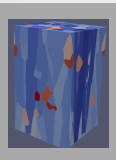
$$\boldsymbol{\lambda}^f = \left[0.510381,\ 0.20316,\ 0.145071,\ 0.0977218,\ 0.036727,\ 0.00693945\right]$$



Variance and the principal values

The screeplot shows the variance explained by each PC





Screeplot

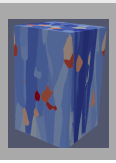
Cumulative fractional variance measures how much variance is included as new PCs are included in the description of the data:

$$\Lambda_k = \sum_{i=1}^{k} \lambda_i^f = (0.510381,\ 0.713541,\ 0.858612,\ 0.956334,\ 0.993061\ 1.)$$

71 % of the variance is contained in the first 2 principal components and 86 % in the first three principal components (PCs).

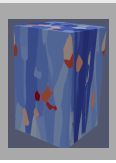
We can develop *reduced dimensional* representations of the data using only the first few principal components (PCs).

The aim of PCA is find linear combinations of the variables that explain a large fraction of the data, preferably only two [principal components]!



Cumulative fractional variance

For Wednesday, read up on biplots, which are a very useful way to compare the datapoints against the variables for pairs of PCs



Biplots - study ahead