university of groningen

faculty of behavioural and social sciences

Date 08.10.2010 |

The Generic Subspace Clustering Model

Marieke Timmerman[1] & Eva Ceulemans[2]

[1]Heymans Institute for Psychology, University of Groningen, The Netherlands - m.e.timmerman@rug.nl

[2]Center for Methodology of Educational Research, Catholic University of Leuven, Belgium

Partitioning of High dimensional data

› Problems with recovery of partition:
  - With increasing dimensionalities, sufficient sample sizes increase strongly
  - Hampered by including variables that hardly or do not reflect the partition

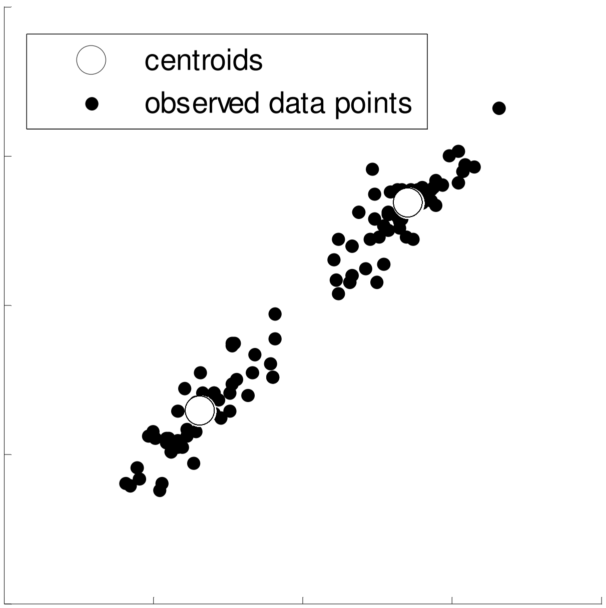
Partitioning of High dimensional data

- Problems with recovery of partition:
  - With increasing dimensionalities, sufficient sample sizes increase strongly
  - Hampered by including variables that hardly or do not reflect the partition

- Approaches to avoid recovery problems:
  - Variable importance: weighting of variables in analysis
  - Variable selection: exclude variables from analysis
  - Subspace clustering: identify clusters in some subspace(s) of the variables

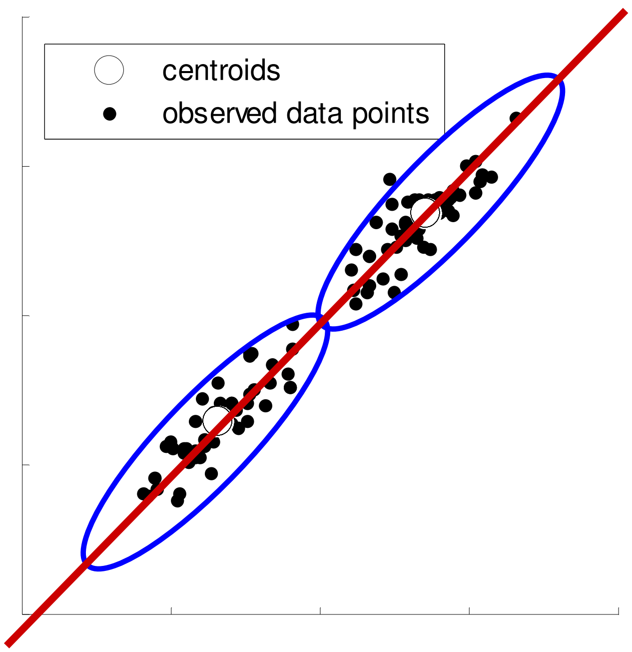
Subspace clustering

› Assumption:
- Clusters are located in some subspace(s) of the variables



Subspace clustering

- Tasks:
  - Identify subspace(s)
  - Identify partitioning



Subspace clustering

› Models
  - Stochastic (e.g., mixtures of factor analyzers)
  - **Deterministic** (e.g., reduced k-means)

Partitioning of objects

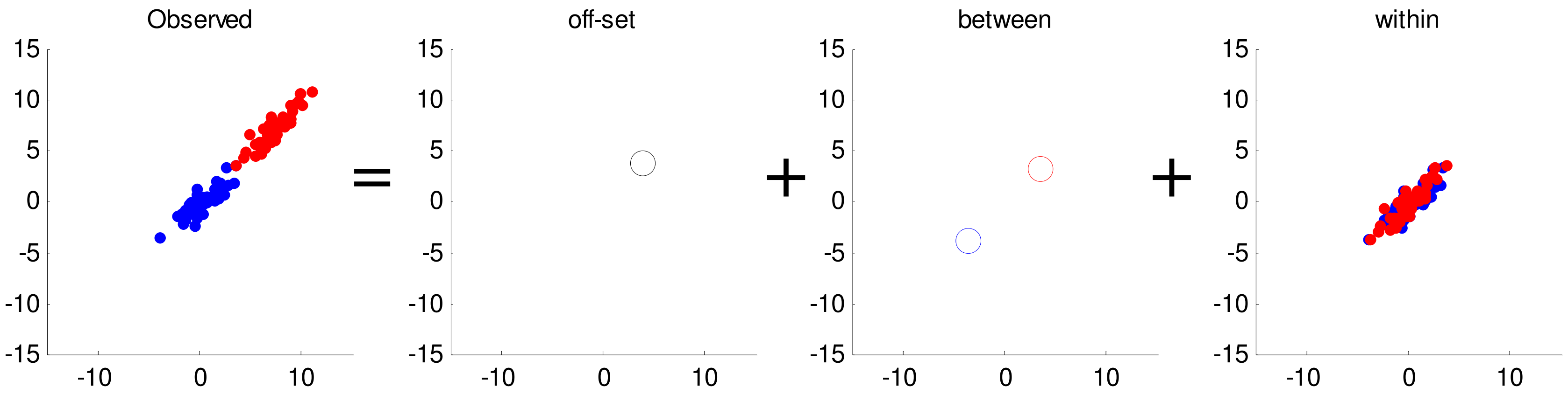
$$\boxed{\mathbf{x}_i = \mathbf{m} + \sum_{c=1}^{C} u_{ic}\left(\mathbf{b}^c + \mathbf{w}_i^c\right)}$$

$\mathbf{x}_i$ $(J\times1)$	observed scores of object i on J variables
$\mathbf{m}$ $(J\times1)$	off-set term
u_{ic}	binary cluster membership indicator: $u_{ic}=1$ if object i belongs to cluster c, and $u_{ic}=0$ otherwise
$\mathbf{b}^c$ $(J\times1)$	centroids of cluster c, with $\sum_{i=1}^{I}\sum_{c=1}^{C} u_{ic}\mathbf{b}^c = \mathbf{0}$
$\mathbf{w}_i^c$ $(J\times1)$	within-cluster residuals of object i in cluster c, with $\sum_{i=1}^{I}\sum_{c=1}^{C} u_{ic}\mathbf{w}_i^c = \mathbf{0}$, and $\mathbf{w}_i^c = \mathbf{0}$ if $u_{ic} = 0$.

Partitioning

$$\mathbf{x}_i = \mathbf{m} + \sum_{c=1}^{C} u_{ic}\left(\mathbf{b}^c + \mathbf{w}_i^c\right)$$

$$\mathbf{x}_i = \text{off-set} + \text{between-part} + \text{within-part}$$



Generic subspace clustering model

$$\boxed{\mathbf{x}_i = \mathbf{m} + \sum_{c=1}^{C} u_{ic}\left(\mathbf{A}_\mathbf{b}\mathbf{f}_\mathbf{b}^c + \mathbf{A}_\mathbf{w}^c\mathbf{f}_{\mathbf{w},i}^c\right) + \mathbf{e}_i^c}$$

$\mathbf{A}_\mathbf{b}$ $(J\times Q_b)$	between-loading matrix
$\mathbf{f}_\mathbf{b}^c$ $(Q_b\times 1)$	between-component scores of cluster c
$\mathbf{A}_\mathbf{w}^c$ $(J\times {Q_w}^c)$	within-loading matrix of cluster c
$\mathbf{f}_{\mathbf{w},i}^c$ $({Q_w}^c\times 1)$	within-component scores of object i in cluster c
$\mathbf{e}_i^c$ $(J\times 1)$	error of object i

Generic subspace clustering model

$$\boxed{\mathbf{x}_i = \mathbf{m} + \sum_{c=1}^{C} u_{ic}\left(\mathbf{A_b}\mathbf{f}_\mathbf{b}^c + \mathbf{A}_\mathbf{w}^c\mathbf{f}_{\mathbf{w},i}^c\right) + \mathbf{e}_i^c}$$

Constraints:

$$\sum_{i=1}^{I}\sum_{c=1}^{C} u_{ic}\mathbf{f}_\mathbf{b}^c = \mathbf{0} \qquad \rightarrow \quad \mathbf{A_b}\mathbf{f}_\mathbf{b}^c : \text{ model for between-part}$$

$$\sum_{i=1}^{I}\sum_{c=1}^{C} u_{ic}\mathbf{f}_{\mathbf{w},i}^c = \mathbf{0} \qquad \rightarrow \quad \mathbf{A}_\mathbf{w}^c\mathbf{f}_{\mathbf{w},i}^c : \text{ model for within-part}$$

with $\mathbf{f}_{\mathbf{w},i}^c = \mathbf{0}$ if $u_{ic} = 0$

$\mathbf{A_b}'\mathbf{A_b} = \mathbf{I}$ and $\mathbf{A}_\mathbf{w}^{c\,\prime}\mathbf{A}_\mathbf{w}^c = \mathbf{I}$

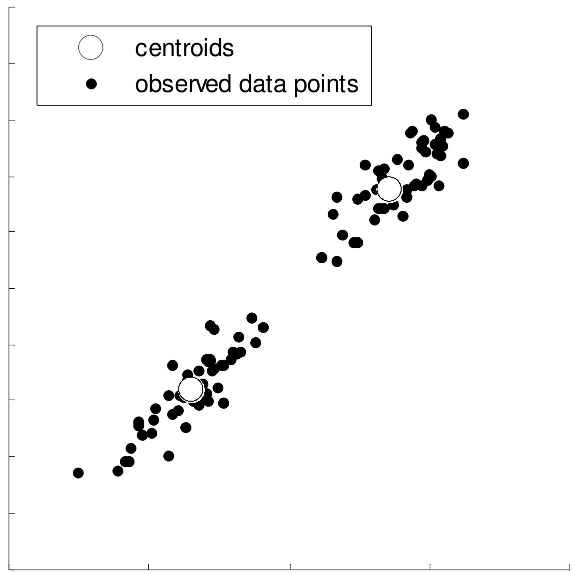
Generic subspace clustering model

$$\boxed{\mathbf{x}_i = \mathbf{m} + \sum_{c=1}^{C} u_{ic}\left(\mathbf{A}_{\mathbf{b}}\mathbf{f}_{\mathbf{b}}^{c} + \mathbf{A}_{\mathbf{w}}^{c}\mathbf{f}_{\mathbf{w},i}^{c}\right) + \mathbf{e}_i^{c}}$$

- Between-part:
  - in full space (Q_b=J), or in any subspace
- For each cluster *c*, within-part:
  - in full space ($Q_w{}^c$=J), or in any subspace

Generic subspace clustering model illustrated

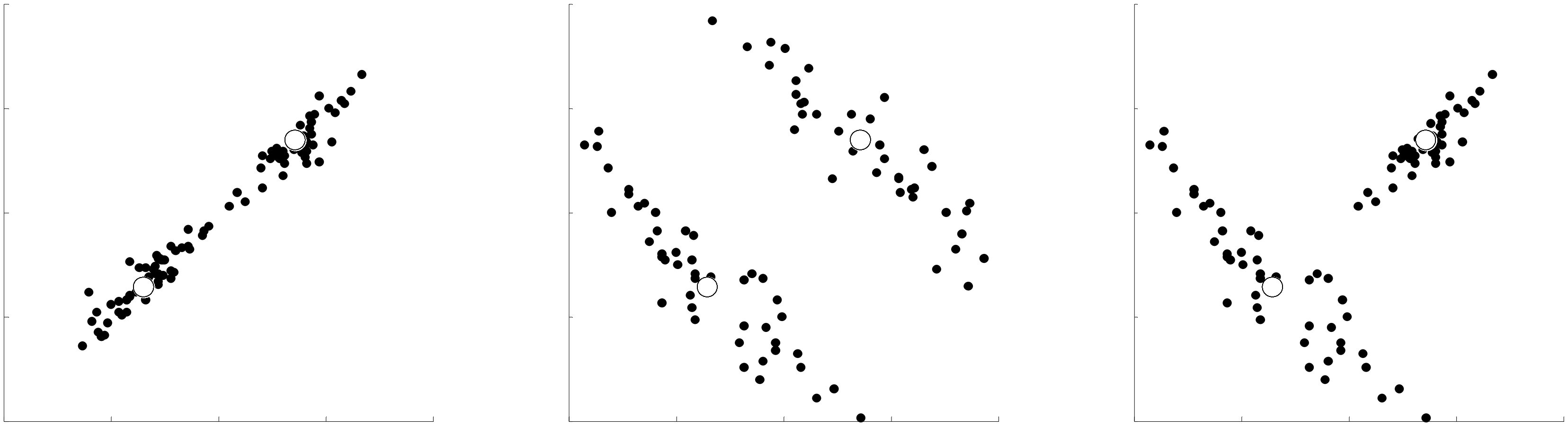
$$\mathbf{x}_i = \mathbf{m} + \sum_{c=1}^{C} u_{ic}\left(\mathbf{A}_{\mathbf{b}}\mathbf{f}_{\mathbf{b}}^{c} + \mathbf{A}_{\mathbf{w}}^{c}\mathbf{f}_{\mathbf{w},i}^{c}\right) + \mathbf{e}_i^{c}$$



- *C* clusters?
  - 1 between subspace
  - *C* within subspaces

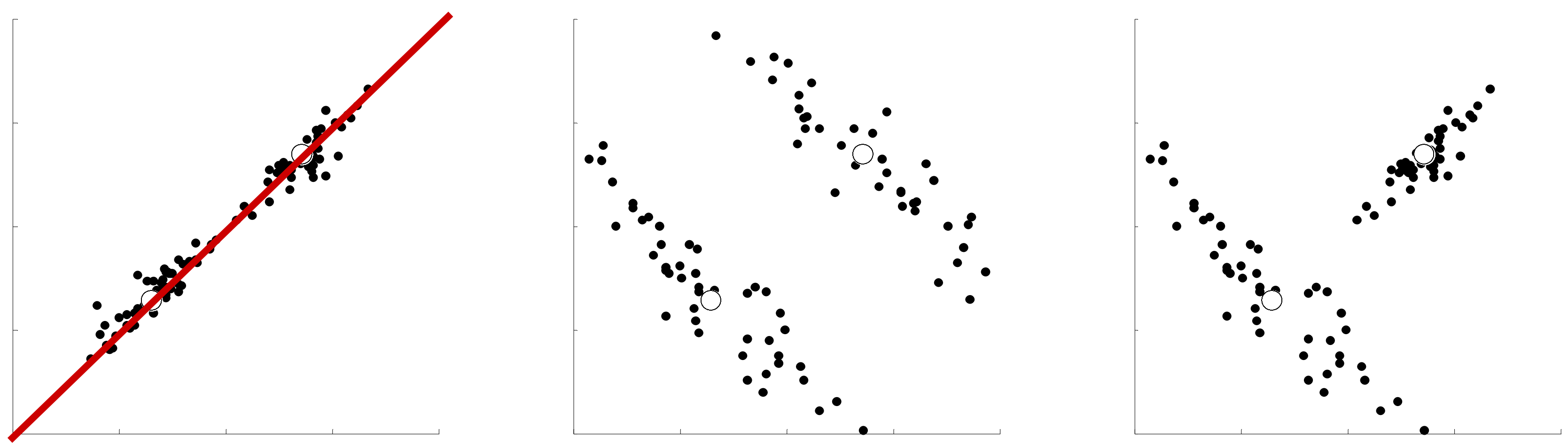
Generic subspace clustering model illustrated

2 observed variables, 2 clusters



Generic subspace clustering model illustrated

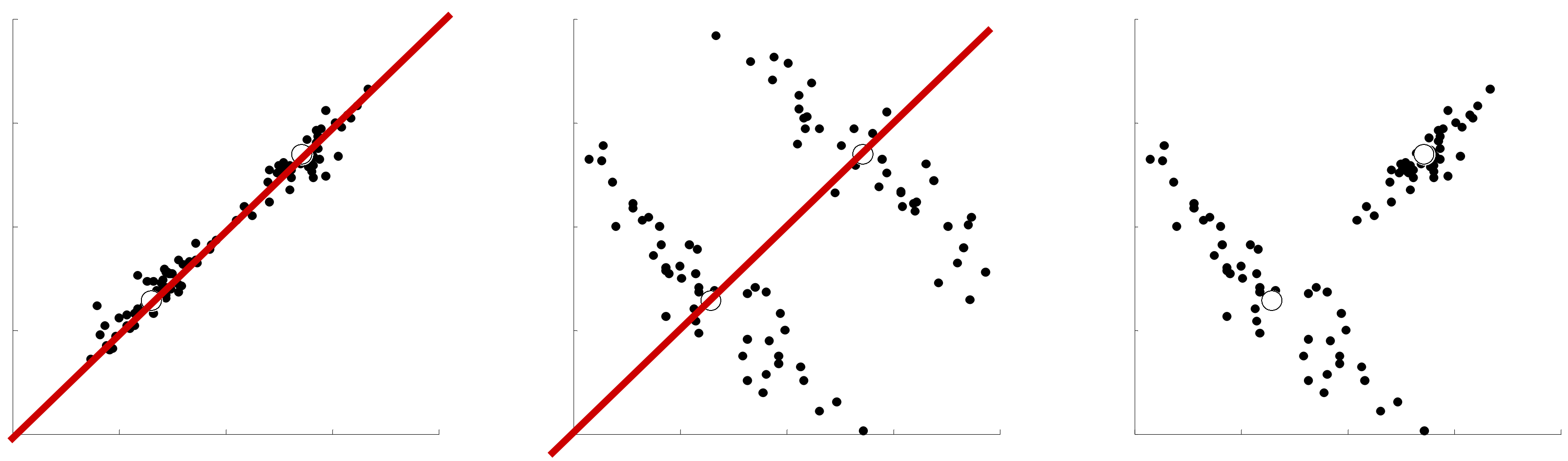
subspace between =
subspace within cluster 1 =
subspace within cluster 2



Generic subspace clustering model illustrated

subspace between =
subspace within cluster 1 =
subspace within cluster 2

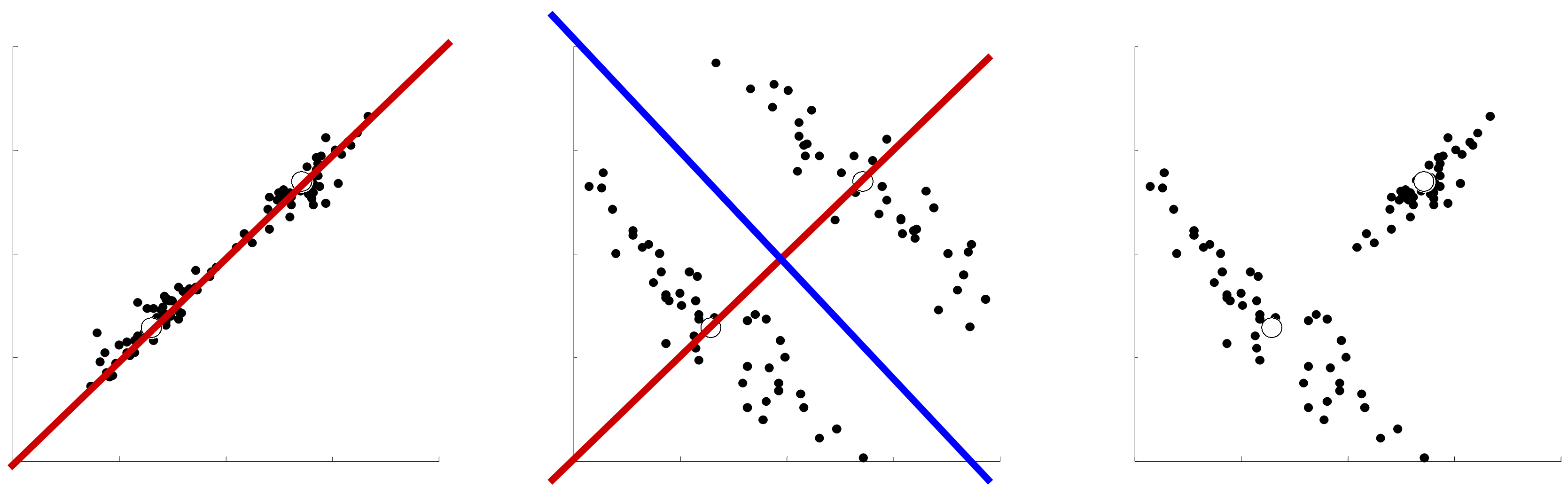
subspace between ≠
{subspace within cluster 1 =
subspace within cluster 2}



Generic subspace clustering model illustrated

subspace between =
subspace within cluster 1 =
subspace within cluster 2

subspace between $\neq$
{subspace within cluster 1 =
subspace within cluster 2}

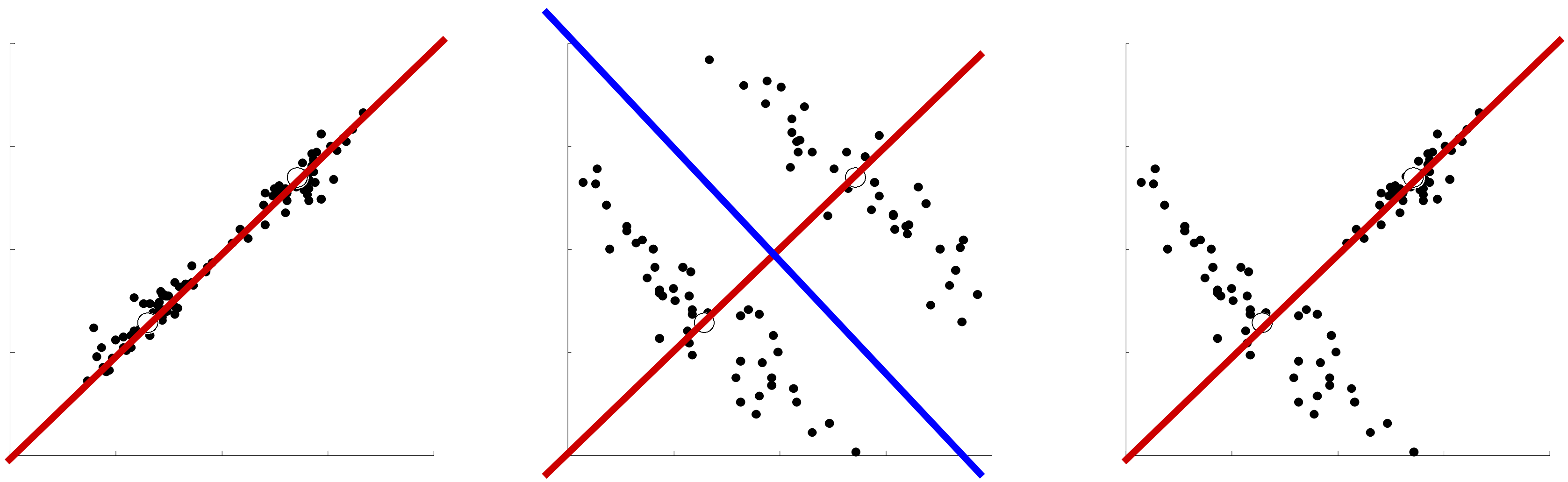


Generic subspace clustering model illustrated

subspace between =
subspace within cluster 1 =
subspace within cluster 2

subspace between ≠
{subspace within cluster 1 =
subspace within cluster 2}

{subspace between =
subspace within cluster 2} ≠
subspace within cluster 1

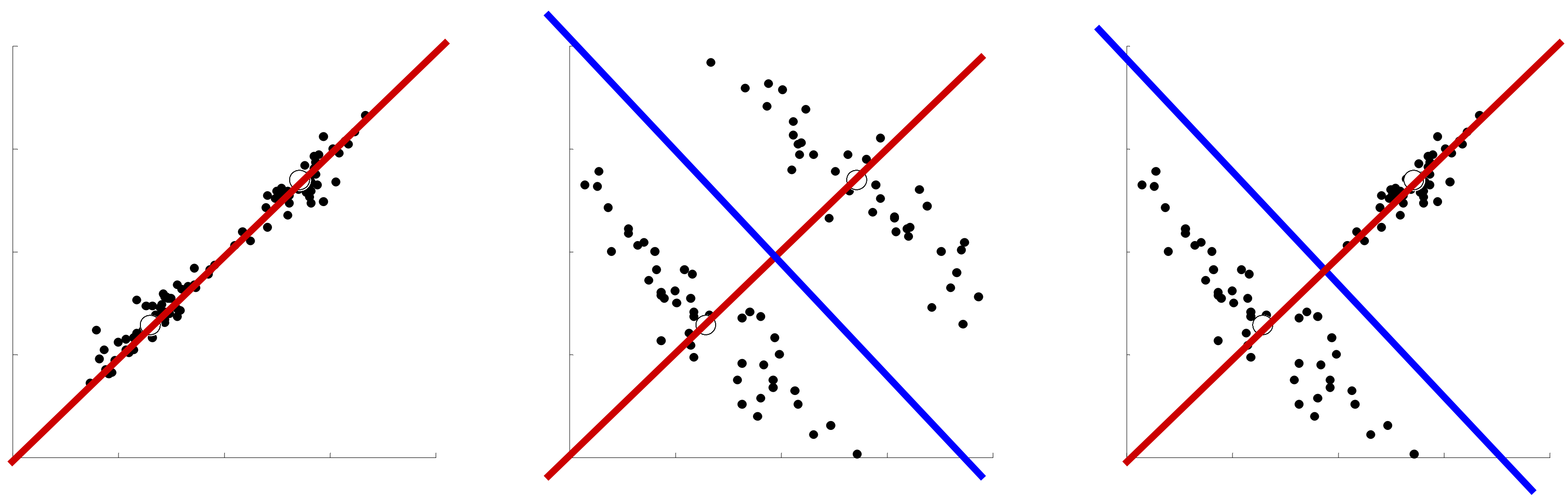


Generic subspace clustering model illustrated

subspace between =
subspace within cluster 1 =
subspace within cluster 2

subspace between ≠
{subspace within cluster 1 =
subspace within cluster 2}

{subspace between =
subspace within cluster 2} ≠
subspace within cluster 1



Generic subspace clustering model

$$\mathbf{x}_i = \mathbf{m} + \sum_{c=1}^{C} u_{ic}\left(\mathbf{A}_\mathbf{b}\mathbf{f}_\mathbf{b}^c + \mathbf{A}_\mathbf{w}^c\mathbf{f}_{\mathbf{w},i}^c\right) + \mathbf{e}_i^c$$

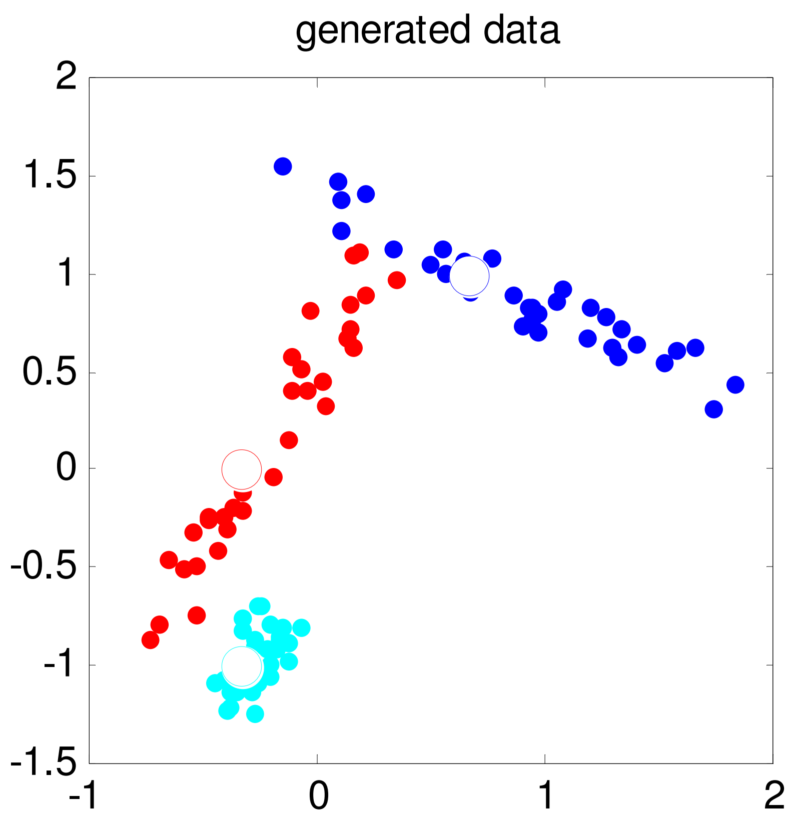
- Very general model
- Various previously proposed models as special cases:
  I. between-part in full space / subspace; within-parts of clusters in subspace / zero
  II. Within-part in subspace(s), with (partial) equalities across clusters

Special cases (I)

- between-part in full space or subspace
- within-parts of clusters in subspace or zero

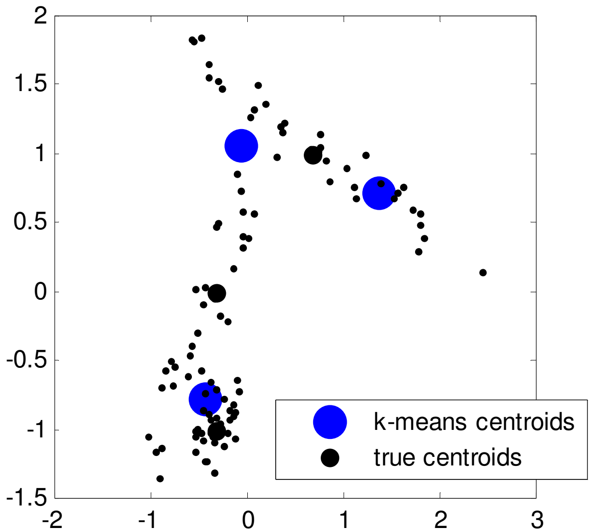
Model	**k-means clustering**	**Projection Pursuit clustering = Reduced k-means**	**PCA-based clustering with class-specific hyperplanes**
between-part	full space	subspace	full space
within-part	zero	zero	for each cluster in a subspace, dimension equal across clusters
Author(s), year	MacQueen, 1967	Bock, 1987; De Soete & Carroll, 1994	Bock, 1987

Illustration of Special cases (I)



k-means

between: full space
within: zero

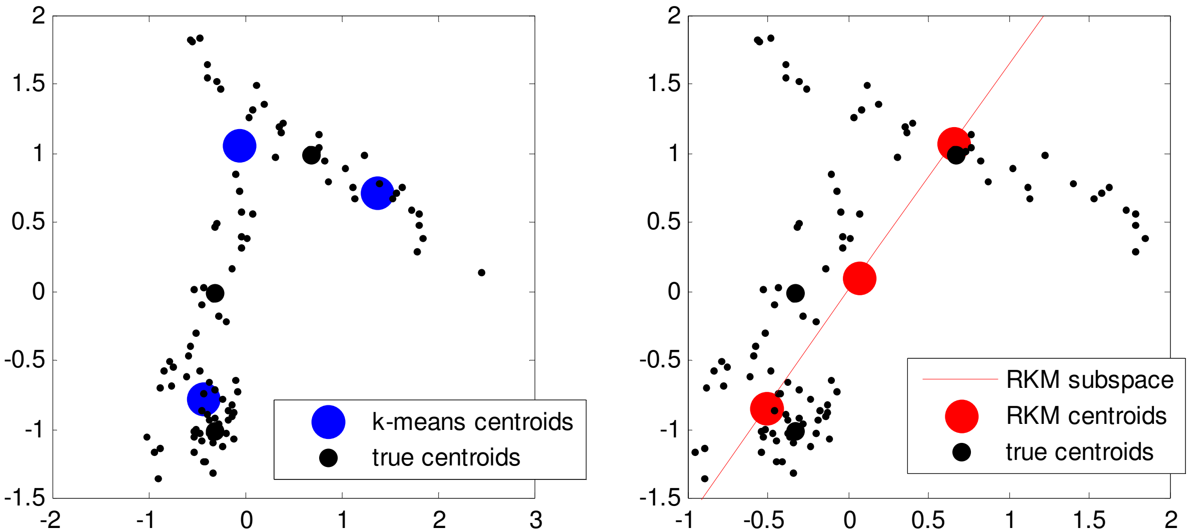


k-means

between: full space
within: zero

Reduced k-means

between: subspace
within: zero



k-means

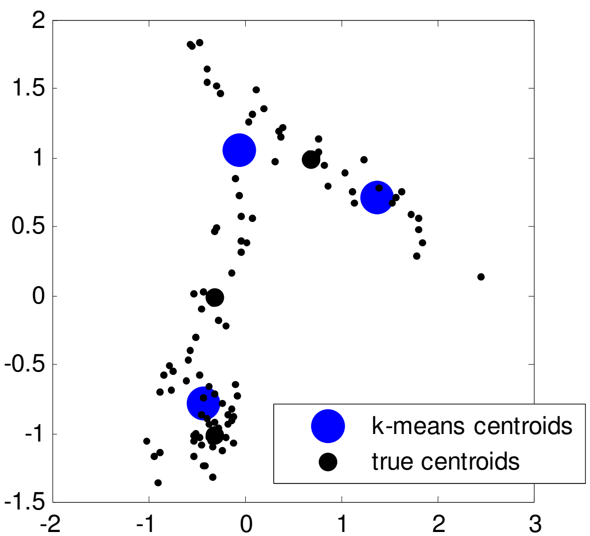
between: full space
within: zero

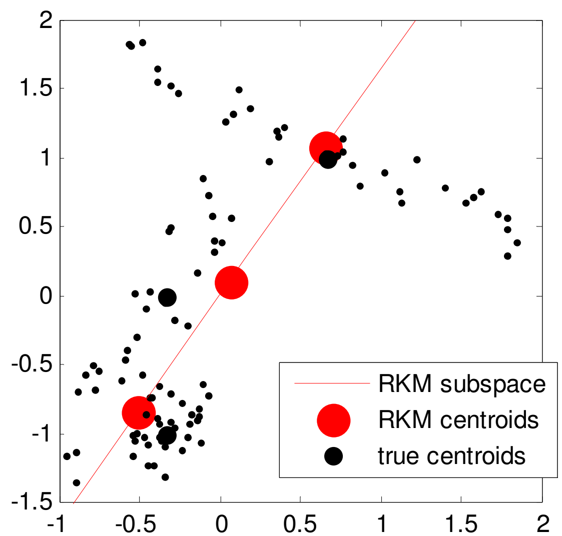
Reduced k-means

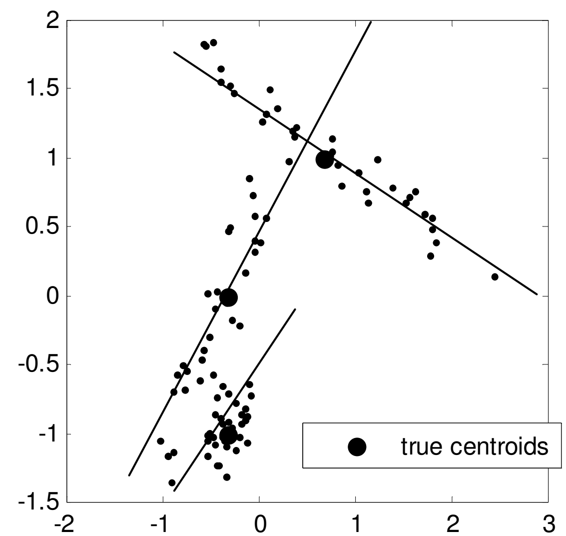
between: subspace
within: zero

PCA-based clustering with class-specific hyperplanes

between: full space
within: subspace per cluster



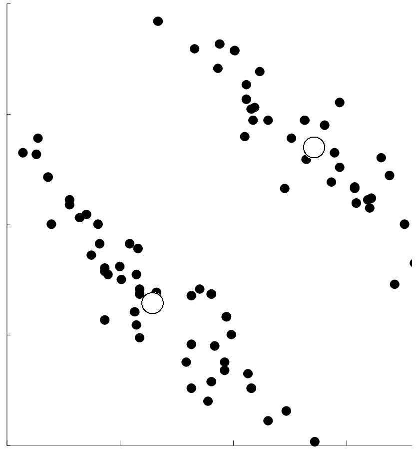




Special cases (II)

- Model for within-part of object *i* in cluster *c:* $\boxed{\mathbf{A}_{\mathbf{w}}^{c}\mathbf{f}_{\mathbf{w},i}^{c}}$

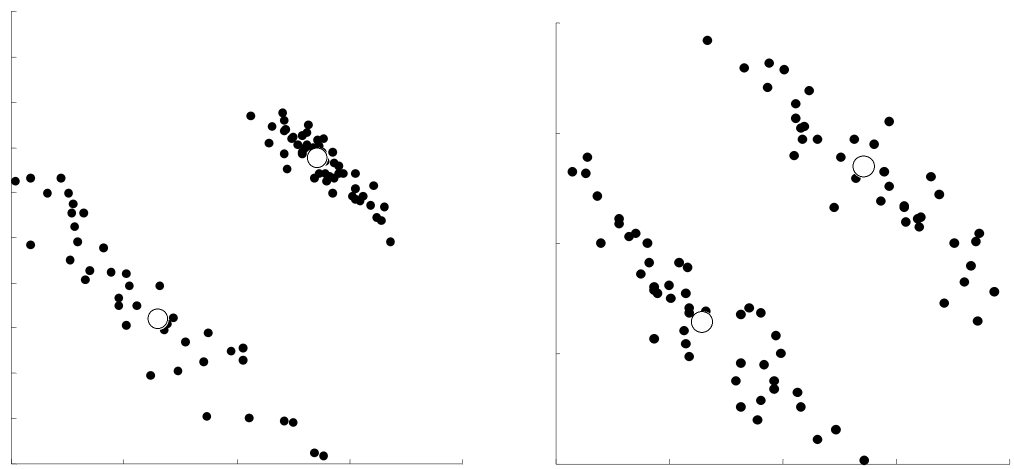
- Within-parts in subspaces?
  Models for within-parts may be (partly) equal to each other



Special cases (II)

- Within-parts in subspace(s)?
  Models for within-parts may be (partly) equal to each other

- Similarities across clusters in
  - subspace
  - (and) shape
  - (and) size



Similarities across clusters in subspace

› PCA-clustering with common and class-specific dimensions (Bock, 1987)

$$\mathbf{A}_{\mathbf{W}}^{c} = [\mathbf{A}_{\mathbf{W}} \mid \mathbf{A}_{\mathbf{W}}^{c*}]$$

with

$\mathbf{A}_{\mathbf{W}}$ the common loading matrix

$\mathbf{A}_{\mathbf{W}}^{c*}$ the class-specific loading matrix

Similarities across clusters in subspace, size and shape

› Borrowed from stochastic models (Banfield & Raftery, 1993):

- similarity in subspace
  - via constraints on $\mathbf{A}_{\mathbf{w}}^{c*} = \mathbf{A}_{\mathbf{w}}$
- similarity in size and/or shape
  - via constraints on variances of within-componentscores $(\mathbf{f}_{\mathbf{w},i}^{c})$ per cluster

Generic subspace clustering model

$$\boxed{\mathbf{x}_i = \mathbf{m} + \sum_{c=1}^{C} u_{ic}\left(\mathbf{A}_\mathbf{b}\mathbf{f}_\mathbf{b}^c + \mathbf{A}_\mathbf{w}^c\mathbf{f}_{\mathbf{w},i}^c\right) + \mathbf{e}_i^c}$$

- Well-examined deterministic models:
  - k-means clustering (no subspace at all)
  - reduced k-means (subspace for between-part)

Generic subspace clustering model

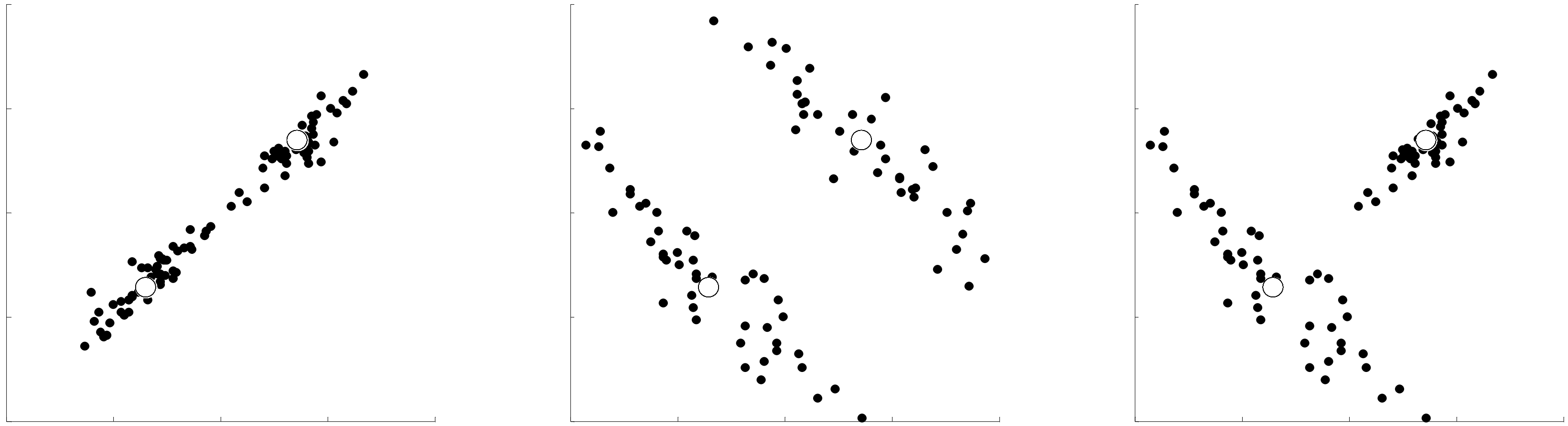
$$\mathbf{x}_i = \mathbf{m} + \sum_{c=1}^{C} u_{ic} \left(\mathbf{A_b} \mathbf{f_b^{\mathit{c}}} + \mathbf{A_w^{\mathit{c}}} \mathbf{f_{w,\mathit{i}}^{\mathit{c}}} \right) + \mathbf{e}_i^c$$

- Well-examined deterministic models:
  - k-means clustering (no subspace at all)
  - reduced k-means (subspace for between-part)

- Hardly examined so far:
  - models with subspaces for the within-parts

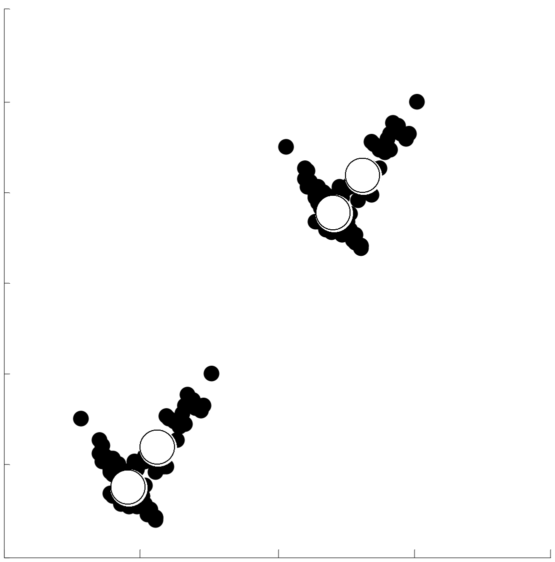
Future of Generic subspace clustering model

- Elaborate models with subspaces for the within-parts
  - fitting procedures
  - obtain insight into additional value of those constraints



- Note:
  Different models may cover different properties of clusters

- Note:
  Different models may cover different properties of clusters
  - Example:
    - cluster centroids optimally separated, or
    - clusters of equal subspace, size and shape



Future of Generic subspace clustering model

- Key issue: Model selection