Multi-Slice Clustering for 3-order tensors

December 8, 2021

Dina Faneva Andriantsiory
Laboratoire d'Informatique de Paris Nord
Université Sorbonne Paris Nord

based on arXiv:2109.10803
published in 20th IEEE International Conference on Machine Learning and Applications, 2021

A joint work with Joseph Ben Geloun and Mustapha Lebbah

Outline of talk

1. Introduction
2. Reviewing the tensor biclustering
3. Multi-Slice Clustering (MSC)
   - Signal and noise
   - ϵ-similarity
   - Slice selection theorems
   - MSC algorithm
   - Experiments
4. Conclusion and Perspectives

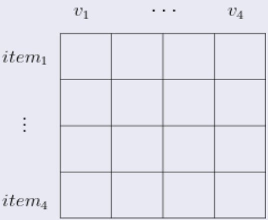
Table of Contents

1. Introduction
2. Reviewing the tensor biclustering
3. Multi-Slice Clustering (MSC)
   - Signal and noise
   - ϵ-similarity
   - Slice selection theorems
   - MSC algorithm
   - Experiments
4. Conclusion and Perspectives

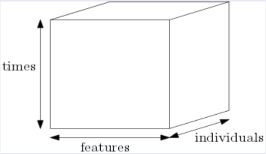
N-th order tensor

First order tensor : Vector

Second order tensor : Matrix

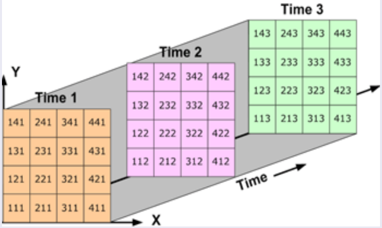


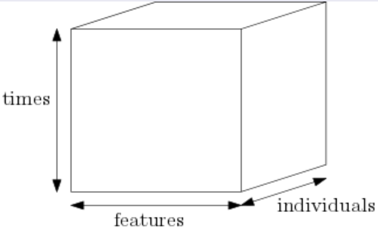
Third order tensor



3-order data (tensor)

Data structure

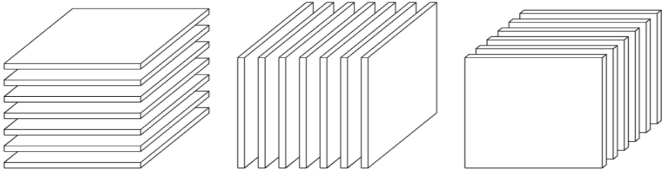




Tensor Manipulation [Kolda and Bader (2009)]

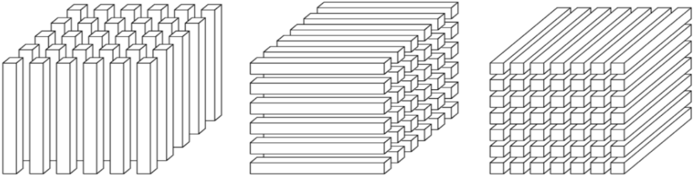
Let $\mathcal{T} \in \mathbb{R}^{m_1 \times m_2 \times m_3}$ the data tensor:

Slices



(a) Horizontal slices: $T_{i::}$ (b) Lateral slices: $T_{:j:}$ (c) Frontal slices: $T_{::k}$

Fibers (Trajectories)

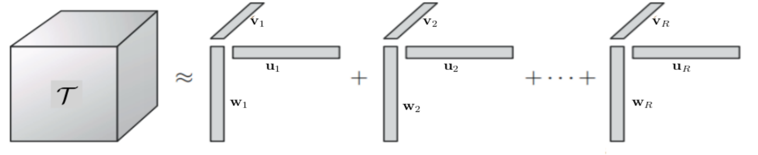


(a) Mode-1 fibers: $T_{:jk}$ (b) Mode-2 fibers $T_{i:k}$ (c) Mode-3 fibers $T_{ij:}$

CANDECOMP/PARAFAC (CP) decomposition [Kiers (2000)]

It is a factorization of tensor $\mathcal{T} \in \mathbb{R}^{m_1 \times m_2 \times m_3}$ into sum of component rank-one tensor.

$$\mathcal{T} \approx \sum_{r=1}^{R} \mathbf{w}_r \otimes \mathbf{u}_r \otimes \mathbf{v}_r,$$



- R is the rank of the tensor decomposition.
- $\forall i, j, k,\ T_{ijk} \approx \sum_{r=1}^{R} \mathbf{w}_r(i)\mathbf{u}_r(j)\mathbf{v}_r(k)$

Definition [W A Barbakh et al. (2009)]

Clustering is a partitioning of a data set or objects into meaningful groups or clusters. The cluster can be defined as a collection of objects which resemble to each other and are dissimilar or different to the objects in the other clusters.

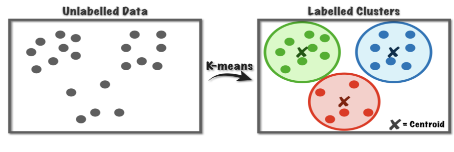
Clustering

Definition [W A Barbakh et al. (2009)]

Clustering is a partitioning of a data set or objects into meaningful groups or clusters. The cluster can be defined as a collection of objects which resemble to each other and are dissimilar or different to the objects in the other clusters.

Examples: clustering algorithm for matrix datasets

- k-means [Hartigan and Wong(1979)]

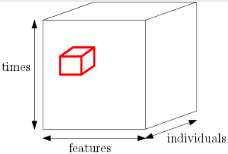


- Biclustering [Yuval Kluger(2003)]

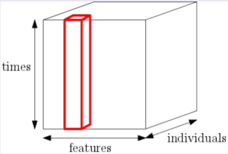
Clustering

Tensor clustering

Triclustering [Zhao and Zaki (2005)]



Tensor biclustering [F. Soheil et al. (2017)]



Multi-way clustering [Wang and Zeng (2019)]

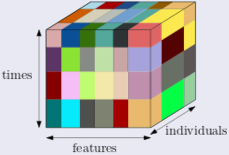


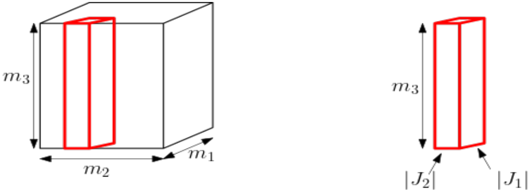
Table of Contents

1. Introduction
2. Reviewing the tensor biclustering
3. Multi-Slice Clustering (MSC)
   - Signal and noise
   - ϵ-similarity
   - Slice selection theorems
   - MSC algorithm
   - Experiments
4. Conclusion and Perspectives

Tensor biclustering

The main goal:

The tensor biclustering problem computes a subset of individuals and a subset of features whose signal trajectories over time lie in a low-dimensional subspace, modeling similarity among the signal trajectories (fibers).



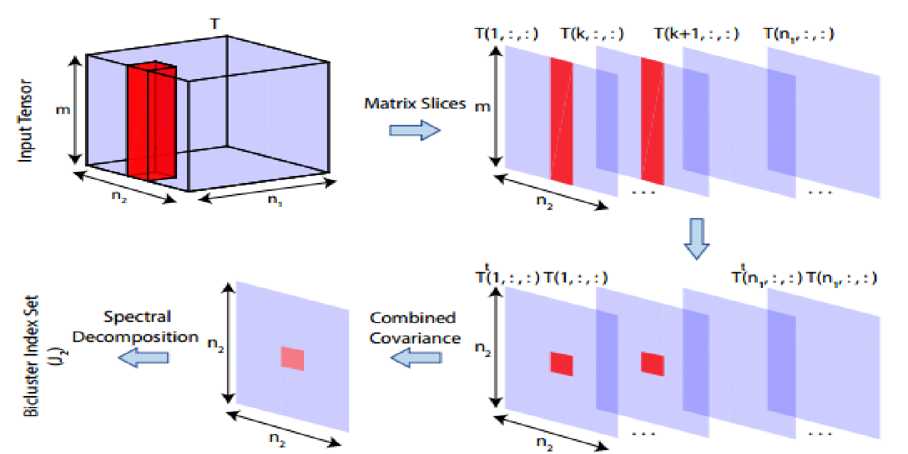
Fours methods for tensor biclustering:

1. Tensor Folding+Spectral
2. Tensor Unfolding+Spectral
3. Thresholding Sum of Squared and Individual Trajectory Lengths
4. Thresholding Sum of Squared Trajectory Lengths

Algorithm [F. Soheil et al. (2017)]

- The input : $\mathcal{T}$, k_1, k_2
- The output : J_1 and J_2 (with $|J_1| = k_1$ and $|J_2| = k_2$)

Tensor Folding+Spectral



Hyperparameter in clustering algorithms:

Usual hyperparameters

- Cluster size :
  - Tensor Biclustering [F. Soheil et al. (2017)].
- The number of clusters :
  - Heterogeneous Tensor Decomposition for Clustering via Manifold Optimization [Y. Sun et al. (2016)].
  - Multi-way clustering via Tensor Block Model [Wang et al. (2019)]
  - Tensor latent block model for co-clustering [B. Rafika et al. (2020)]

Hyperparameter in clustering algorithms:

Usual hyperparameters

- Cluster size :
  - Tensor Biclustering [F. Soheil et al. (2017)].
- The number of clusters :
  - Heterogeneous Tensor Decomposition for Clustering via Manifold Optimization [Y. Sun et al. (2016)].
  - Multi-way clustering via Tensor Block Model [Wang et al. (2019)]
  - Tensor latent block model for co-clustering [B. Rafika et al. (2020)]

Our hyperparameter

- Threshold error to gauge the similarity: ϵ

Table of Contents

1. Introduction
2. Reviewing the tensor biclustering
3. Multi-Slice Clustering (MSC)
    - Signal and noise
    - ϵ-similarity
    - Slice selection theorems
    - MSC algorithm
    - Experiments
4. Conclusion and Perspectives

- Decomposition in signal and noise: We assume $\mathcal{T} \in \mathbb{R}^{m_1 \times m_2 \times m_3}$:

$$\mathcal{T} = \mathcal{X} + \mathcal{Z} = \gamma \mathbf{w} \otimes \mathbf{u} \otimes \mathbf{v} + \mathcal{Z} \quad (1)$$

  - $\mathcal{X}$ is the signal tensor, γ is the signal strength.
  - $\mathcal{Z}$ is the noise tensor: each entry of $\mathcal{Z} \sim i.i.d.\mathcal{N}(0, 1)$.

Multi-Slice Clustering (MSC) for 3-order tensor

Note

1. The clustering selection of each mode is independent
2. The following presentation focuses on the mode-1

For $i \in [m_1]$,the i-th slice is:

$$T_i = \mathcal{T}(i,:,:) = X_i + Z_i \tag{2}$$

$$C_i = T_i^t T_i = \lambda_i \mathbf{v}_i \mathbf{v}_i^t + W_i$$

Proposition 1 [Y. Zhong et al. (2017)]

For $\lambda_i = \mathcal{O}(m_1)$ and $\alpha = \|\mathbf{v}_i\|_\infty$, we have

$$\|\hat{\mathbf{v}}_i - \mathbf{v}_i\|_\infty \leq \mathcal{O}\left(\frac{1}{\lambda_i}\alpha \log(m_1)\right) \tag{3}$$

with high probability as $m_1 \to \infty$, where λ_i is the top eigenvalue of $X_i^t X_i$ and $\hat{\mathbf{v}}_i$ is the lagest eigenvalue of C_i.

Noise slice

Noise slice [Gupta et al. (2018)] and [Johnstone et al. (2001)]

Let $Z = \begin{bmatrix} \mathbf{z}_1 \\ \vdots \\ \mathbf{z}_{m_3} \end{bmatrix} \in \mathbb{R}^{m_3 \times m_2}$, where $\mathbf{z}_i \in \mathbb{R}^{m_2}$ and $\mathbf{z}_i = (z_{i1}, z_{i2}, \cdots, z_{im_2}) \sim \mathcal{N}_{m_2}(0, I)$.

$$C = Z^t Z \sim \mathcal{W}_{m_2}(m_3, I) \tag{4}$$

- where $\mathcal{W}_{m_2}(m_3, I)$ is the Wishart distribution with parameter m_2, m_3 and I.
- The largest eigenvalue distribution λ of C

$$\frac{\lambda - \mu_{m_2 m_3, 1}}{\sigma_{m_2 m_3, 1}} \xrightarrow{\mathcal{D}} F_1 \tag{5}$$

where F_1 is the Tracy-Widom (TW) cumulative distribution function (cdf), and

$$\begin{aligned} \mu = \mu_{m_2 m_3, 1} &= (\sqrt{m_2 - 1} + \sqrt{m_3})^2 \\ \sigma = \sigma_{m_2 m_3, 1} &= \sqrt{\mu_{m_2 m_3, 1}} \left(\frac{1}{\sqrt{m_2 - 1}} + \frac{1}{\sqrt{m_3}} \right)^{\frac{1}{3}} \end{aligned} \tag{6}$$

Lemma

For each $i \in [n]$, let Z_i be a $M \times N$ noise matrix and $Y_i = Z_i^t Z_i$ that follows a white Wishart distribution $\mathcal{W}_N(M, I)$. Let λ_i be the largest eigenvalue of Y_i which obeys a TW cdf of parameters (μ, σ).

For $\mathbf{a} = (a_1, \cdots, a_n) \in \mathbb{R}^n$ such that $\forall i, a_i \geq 0$,

1. *If $\sum_i^n a_i \leq 1$, for $\lambda = \Omega(\mu)$, we have*

$$\frac{1}{\lambda}\sum_{i=1}^{n} a_i\lambda_i \leq \sqrt{\log(n)} \tag{7}$$

with probability at least $1 - en^{-c_1}$ for a constant $c_1 > 0$.

2. *If $1 < \sum_i^n a_i \leq n$, for $\lambda = \Omega(n\mu)$, we have*

$$\frac{1}{\lambda}\sum_{i=1}^{n} a_i\lambda_i \leq \sqrt{n} \tag{8}$$

with probability at least $1 - en^{-c_2}$ for a constant $c_2 > 0$.

- Principal component (PC) of the slice i:

$$\mathrm{PC}(T_i) \approx (\hat{\lambda}_i, \hat{\mathbf{v}}_i), \quad \forall i \in [m_1]$$

- Matrix representative of all slices in mode-1:

$$V = \begin{bmatrix} \tilde{\lambda}_1 \hat{\mathbf{v}}_1 & \cdots & \tilde{\lambda}_{m_1} \hat{\mathbf{v}}_{m_1} \end{bmatrix}, \text{ where } \tilde{\lambda}_i = \lambda_i / max_j(\lambda_j) \tag{9}$$

$\epsilon-$similarity

We call the $i-$th and $j-$th slices $\epsilon-$similar, if for a small $\epsilon \in]0, 1[$ we have,

$$c_{ij} = |\langle \tilde{\lambda}_i \hat{\mathbf{v}}_i, \tilde{\lambda}_j \hat{\mathbf{v}}_j \rangle| \geq 1 - \frac{\epsilon}{2}. \tag{10}$$

Let $\mathbf{d} = (d_1, \cdots, d_{m_1})$ a real vector defined by, $\forall i \in [m_1]$

$$d_i = \sum_{j \in [m_1]} c_{ij}. \quad (11)$$

Theorem 1

Let $|J_1| = l$, assume that $\sqrt{\epsilon} \leq \frac{1}{m_1 - l}$. $\forall i, n \in J_1$, for $\lambda = \mathcal{O}(\mu)$ we have

$$|d_i - d_n| \leq l\frac{\epsilon}{2} + \sqrt{\log(m_1 - l)} \quad (12)$$

with probability at least $1 - e(m_1 - l)^{-c_1}$ with a constant $c_1 > 0$.

Slice selection theorems

theorem 2

For $i \in \bar{J}_1$, if $\lambda = \mathcal{O}(\mu m_1)$,

$$d_i \leq \frac{l}{m_1} + \sqrt{\log(m_1 - l)} \tag{13}$$

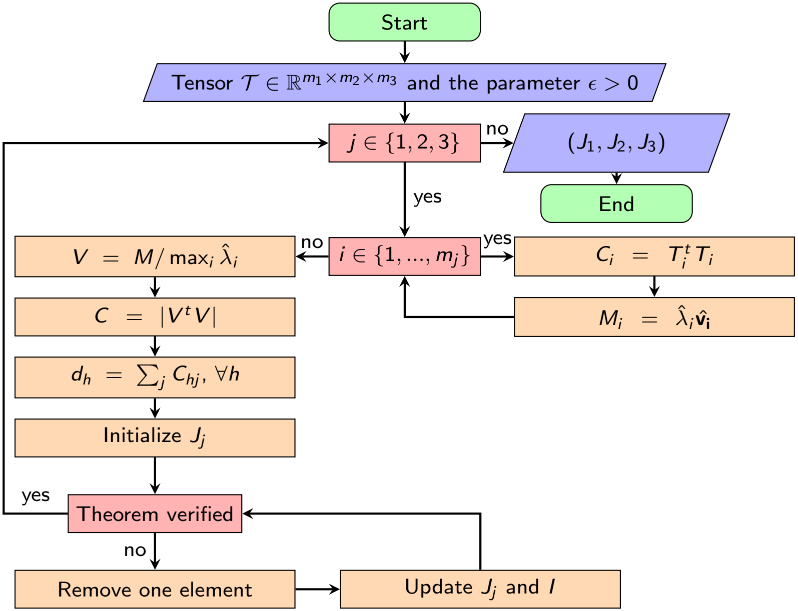
with probability at least $1 - e m_1^{-c_1}$ where c_1 is a constant positif.

Corollary

Let J_1 the set of all indices of the cluster in the first dimension $J_1 \subset [m_1]$. Then

$$\mathrm{dist}(d(J_1), d(\bar{J}_1)) \geq l(1 - \frac{\epsilon}{2} - \frac{1}{m_1}) - \sqrt{\log(m_1 - l)} \tag{14}$$

with probability at least $1 - e(m_1 - l)^{-c_1}$ where c_1 is a constant positif


Start
Tensor $\mathcal{T} \in \mathbb{R}^{m_1 \times m_2 \times m_3}$ and the parameter $\epsilon > 0$
$j \in \{1, 2, 3\}$
no
(J_1, J_2, J_3)
End
yes
$i \in \{1, ..., m_j\}$
yes
no
$C_i = T_i^t T_i$
$M_i = \hat{\lambda}_i \hat{\mathbf{v}}_\mathbf{i}$
$V = M / \max_i \hat{\lambda}_i$
$C = |V^t V|$
$d_h = \sum_j C_{hj}, \forall h$
Initialize J_j
yes
Theorem verified
no
Remove one element
Update J_j and I

Experiments: Synthetic data

We generate the data with:

- $\mathcal{T} = \mathcal{X} + \mathcal{Z} = \gamma \, \mathbf{w} \otimes \mathbf{u} \otimes \mathbf{v} + \mathcal{Z}$
- $m_1 = m_2 = m_3 = 50$
- $|J_1| = |J_2| = |J_3| = l = 10$
- $\mathbf{w}_i = \begin{cases} \frac{1}{|J_1|} \text{ if } i \in J_1 \\ 0 \text{ if not} \end{cases}, \mathbf{u}_i = \begin{cases} \frac{1}{|J_2|} \text{ if } i \in J_2 \\ 0 \text{ if not} \end{cases}, \mathbf{v}_i = \begin{cases} \frac{1}{|J_3|} \text{ if } i \in J_3 \\ 0 \text{ if not} \end{cases}$
- $\frac{1}{(m_1 - l)^2} = 0.00062$
- The similarity index

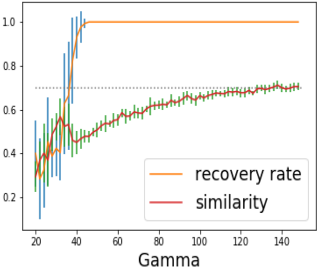
$$\mathrm{sim} = \frac{1}{3} \sum_{r=1}^{3} \frac{1}{\hat{J}_r^2} \sum_{i,j \in \hat{J}_r} c_{ij} \leq 1. \tag{15}$$

- Recovery rate (rec)

$$\mathrm{rec} = \frac{1}{3} \sum_{r=1}^{3} \frac{|J_r \cap \hat{J}_r|}{l} \tag{16}$$

Experiments: Synthetic data

For each γ, we do the computation 10 times.



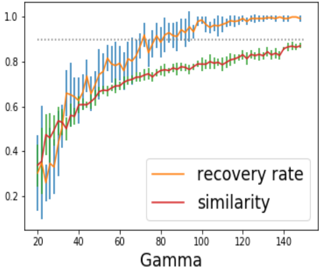
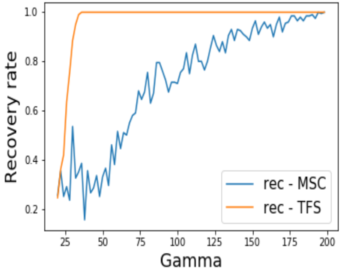


Figure: Recovery rate of the MSC method with $\gamma \in [20, 150]$: (left) with $\epsilon = 0.005$ and (right) with $\epsilon = 0.0006$.

- $m_1 = m_2 = 70$ and $m_3 = 50$
- $|J_1| = |J_2| = |J_3| = l = 10$



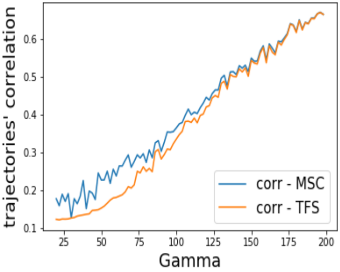


Figure: Recovery rates (rec) and correlation mean (corr) for MSC and TFS.

Experiments: MSC on real data

- Flow injection analysis (FIA) dataset [Norgaard and C(1994)]
- 12 (samples) $\times$100 (wavelengths) $\times$89 (times) and $\epsilon = 0.00013$

mode	**Indices of similar slices**	**frobenius norm between slices**	**fibers correlation**
mode-1	10, 11	3.08616	0.99915
mode-2	39, 40, 41, 42, 43	0.40307	0.97164
mode-3	45, 46, 47, 48, 49, 50, 51, 52	1.05795	0.99059

The RMSE of the triclustering (sub-cube) $= 0.48360$

Experiments: MSC on real data

- Flow injection analysis (FIA) dataset [Norgaard and C(1994)]
- 12 (samples) $\times$100 (wavelengths) $\times$89 (times) and $\epsilon = 0.00013$

mode	**Indices of similar slices**	**frobenius norm between slices**	**fibers correlation**
mode-1	10, 11	3.08616	0.99915
mode-2	39, 40, 41, 42, 43	0.40307	0.97164
mode-3	45, 46, 47, 48, 49, 50, 51, 52	1.05795	0.99059

The RMSE of the triclustering (sub-cube) = 0.48360

Evaluation: We add one index to the cluster (random index from outside the cluster)

E-mod	**frobenius norm between slices**	**fibers correlation**	**RMSE of triclustering**
E-1	3.17592	0.81419	0.49257
E-2	3.83777	0.74525	0.51133
E-3	4.86486	0.81829	1.19725

Table of Contents

1 Introduction

2 Reviewing the tensor biclustering

3 Multi-Slice Clustering (MSC)
- Signal and noise
- ϵ-similarity
- Slice selection theorems
- MSC algorithm
- Experiments

4 Conclusion and Perspectives

Conclusion

- The majority of the clustering algorithm depend on the cluster size or the number of the clusters.
- The MSC algorithm depend on ϵ to carry out the cluster selection and rafinement.
- The Experiments validates that the MSC performs well.

Perspective work

- Recursive form of the MSC
- Apply MSC to large tensor size (using parallel computing)
- Generalize MSC to multiple cluster detection (select the most dominant eigenvectors for each slice and work out a partition of the tensor of similar data)

Conclusion

- The majority of the clustering algorithm depend on the cluster size or the number of the clusters.
- The MSC algorithm depend on ϵ to carry out the cluster selection and rafinement.
- The Experiments validates that the MSC performs well.

Perspective work

- Recursive form of the MSC
- Apply MSC to large tensor size (using parallel computing)
- Generalize MSC to multiple cluster detection (select the most dominant eigenvectors for each slice and work out a partition of the tensor of similar data)

Thank you for your attention!

Bibliography

John A Hartigan and Manchek A Wong.
Algorithm as 136: A k-means clustering algorithm.
Journal of the royal statistical society. series c (applied statistics), 28(1):100–108, 1979.

Norgaard and Ridder C.
Rank annihilation factor analysis applied to flow injection analysis with photodiode-array detection, 1994.
Chemometrics and Intelligent Laboratory, Systems 23:107
`http://www.models.life.ku.dk/Flow_Injection`.

Joseph T. Chang Mark Gerstein Yuval Kluger, Ronen Basri.
Spectral biclustering of microarray cancer data: Co-clustering genes and conditions.
Genome Research, 13:703–716, 2003.