The Problem and Nature of Degenerate Solutions or Decompositions of 3-way Arrays

Richard A. Harshman
Psychology Department, University of Western Ontario

Workshop On Tensor Decompositions
Palo Alto, California, July 2004

Sponsored by: American Institute of Mathematics & Stanford University

Acknowledgements**

This work was done in collaboration with

Margaret E. Lundy

Dept. of Psychology, University of Western Ontario

And is supported in part by a grant from the

Natural Sciences and Engineering Research Council of Canada

Notes on slide changes: **

A few slides have been *reworded* for clarity since my talk.

These are indicated by a single asterisk (on the title).

In addition, some *new* slides have been inserted.

These are indicated by a double asterisk on the title (as above). These slides convey information that was only given orally or written on the blackboard when I gave the talk, but that now needs to be included visually to provide essential connections or explanation to those viewing this slide set for the first time.

There is also an Annotated Bibliography in a separate file

available on this site. It goes over the contributions mentioned on the "Short History" slide in much more detail, and has recently been updated to include several talks given at, or subsequent to, the 2004 Workshop.

An Example of the Degeneracy Phenomenon: The TV-Ratings Data**

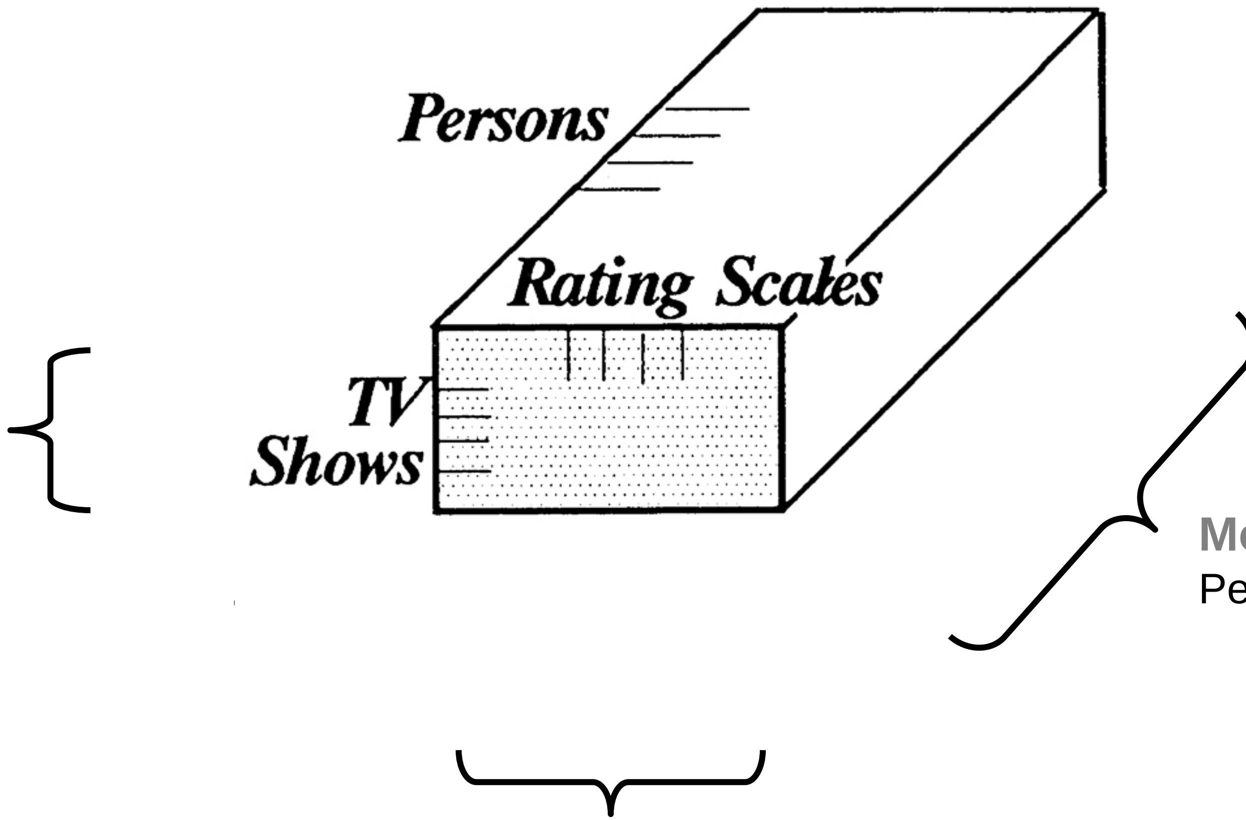
A research study collected ratings of 15 TV programs made by 40 people.

Each person rated each show on 16 rating scales, such as “1=Violent …7= Peaceful”.

The data formed a three-way array, size 15x16x40, as shown on the next slide.

The Three-way Array X

Mode A
TV-Show
Mode



Mode C
Person Mode

Mode B
Rating Scale Mode

TV Shows

1. Mash
2. Charlie's Angels
3. All in the Family
4. 60 Minutes
5. The Tonight Show
6. Let's Make a Deal
7. The Waltons
8. Saturday Night Live
9. News (any channel; national edition)
10. Kojak
11. Mork and Mindy
12. Jacques Cousteau
13. Football
14. Little House on the Prairie
15. Wild Kingdom

Rating Scales

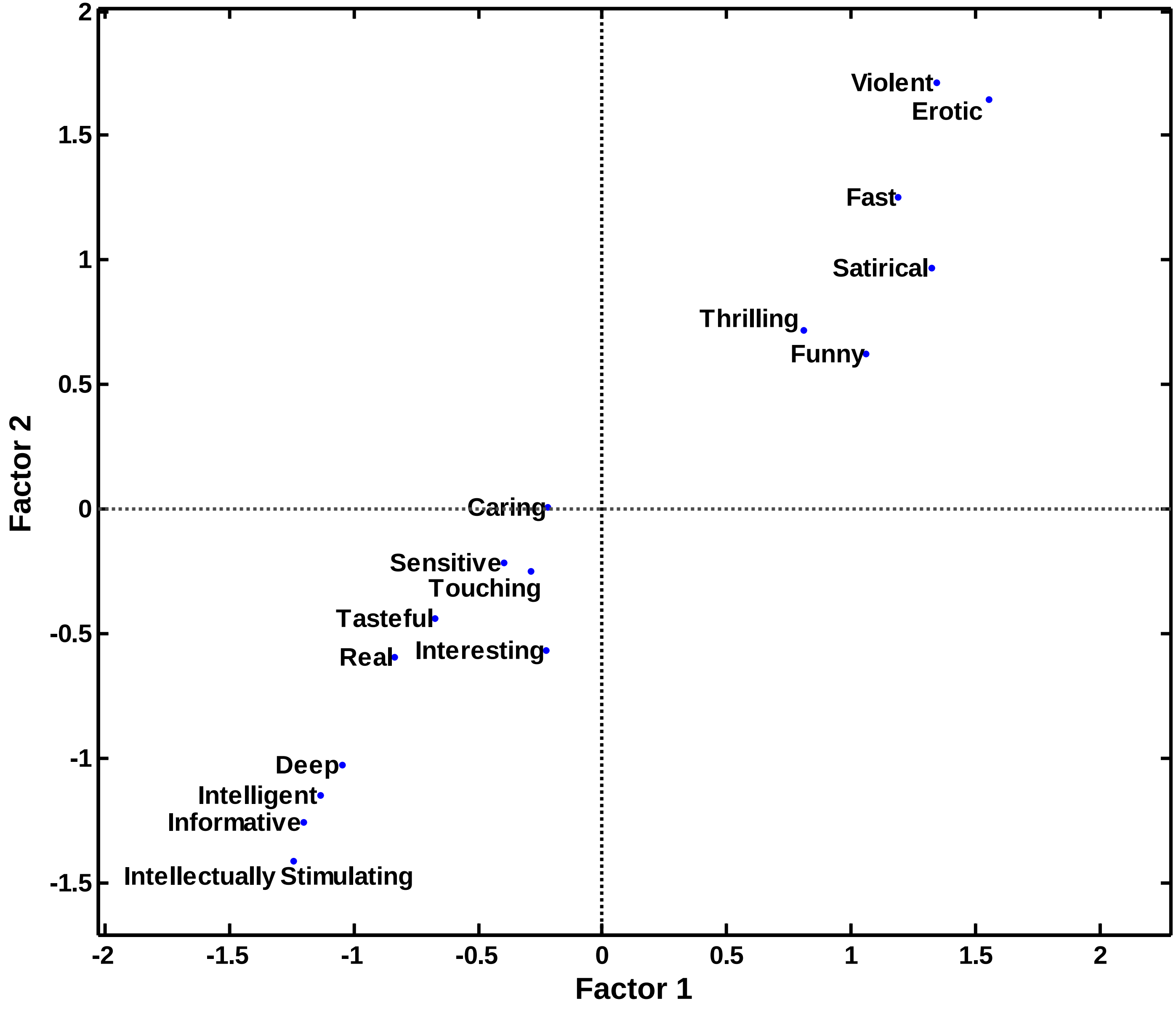
1. Thrilling . . . . Boring
2. Intelligent . . . Idiotic
3. Erotic . . . . . Not Erotic
4. Sensitive . . . . Insensitive
5. Interesting . . Uninteresting
6. Fast . . . . Slow
7. Intellectually Stimulating . . . Intellectually Dull
8. Violent . . . . Peaceful
9. Caring . . . . . Callous
10. Satirical . . . . Not Satirical
11. Informative . . Uninformative
12. Touching . . . "Leaves Me Cold"
13. Deep . . . . . . Shallow
14. Tasteful . . . . Crude
15. Real . . . . . . Fantasy
16. Funny . . . . . Not Funny

The Degenerate Parafac Solution**

In the standard Parafac analysis, the 3 factor solution contained two factors that had extremely high negative correlations in all modes. Contributions of the two factors got larger and larger (diverged) as the analysis progressed -- but since they almost cancelled out -- the sum of the two converged.

On the plot, the two factors that were “participating in the degeneracy” had collapsed into a ragged line of points.

Unconstrained Degenerate Solution - Adjectives Mode: Factor 1 vs. Factor 2



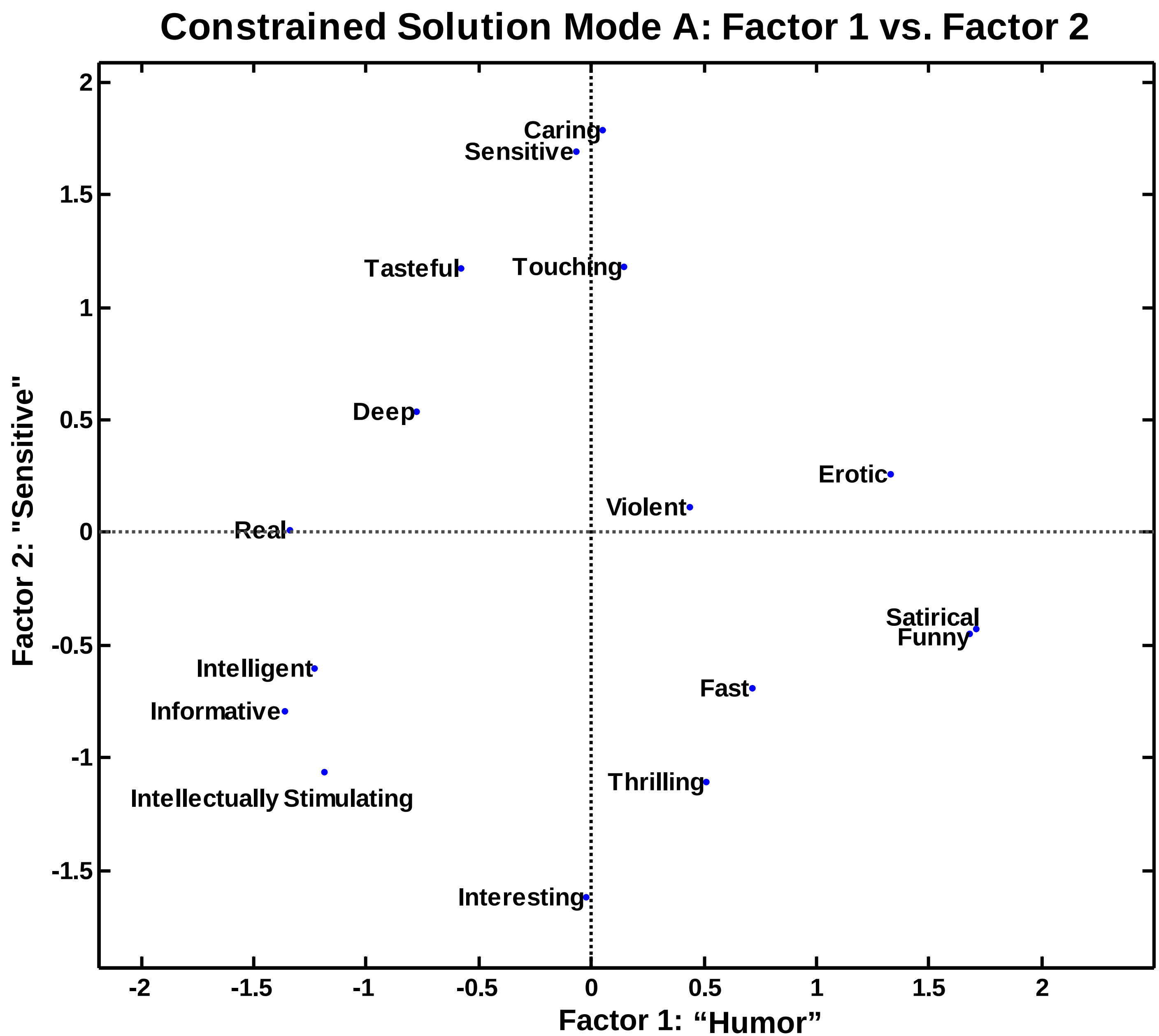
Blocking Degeneracy**

However, when the factors in any mode were constrained to have independent (uncorrelated) loadings, a new solution was obtained in which all 3 Modes had a meaningful 3-factor structure.

The TV-show Mode was chosen as the “most natural” one for the constraint.

In the constrained solution, the two previously degenerate (collinear) factors were found to be Program Humor and Program Sensitivity (later confirmed by “unshearing”).

The non-degenerate version of the Rating Scales plot is shown on the next slide.


Constrained Solution Mode A: Factor 1 vs. Factor 2
Factor 2: "Sensitive"
Factor 1: “Humor”
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
Caring
Sensitive
Tasteful
Touching
Deep
Erotic
Violent
Real
Satirical
Funny
Intelligent
Fast
Informative
Thrilling
Intellectually Stimulating
Interesting

Is it meaningful?**

We discovered that the degenerate solution actually recovers the full information, but it's hidden.

There exists a linear transformation that brings the degenerate solution into agreement with the nondegenerate one. After the transformation, both nondegenerate factors are recovered well (recovery correlation is over .99 for each).

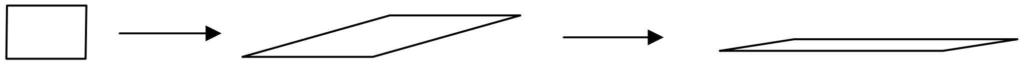
The transformation relating the two solutions was an extreme shear (to be explained) --sometimes in combination with rotation.

The degeneracy phenomenon:

(a) imposes an extreme shear on the subspace spanned by the factors involved

(b) It affects all three Modes, shearing some one way and some the reverse way

(c) often it keeps getting more and more extreme with each added iteration of the analysis...



What's Going On Here?

Short History of Study of Degeneracy*

- Harshman and Lundy (1984) first described phenomenon and showed, both by “filtering” real data and by constructing simulated data, that “Tucker-variation” (changes in axis angles) caused it.
- Kruskal et al. (1989) provided basic mathematical framework for relating properties of degeneracy to surprising array rank properties.
- Ten Berge (1988,1991) proved key part of main theorem stated by Kruskal and added insights about rank and “typical rank” of arrays.
- Mitchell and Burdick (1994) discovered novel degenerate behavior: *temporary* quasi-degeneracies, which they called “swamps”.
- Rayens & Mitchell (1997) introduced idea of “regularization” to stabilize algorithms, avoid swamps, and accelerate convergence.
- Paatero (2000) deepened structural understanding of degeneracy and showed how to create arrays yielding degeneracies and “swamps”.
- Zijlstra & Kiers (2002) identified universal properties of degenerate solutions that were present across wide variety of factor-like models.

“Swamps”

- Mitchell and Burdick (1994) identified temporary degenerate behavior of ALS on route leading to nondegenerate solution

- Progress slows down to almost standstill for thousands of iterations, with dimensions highly correlated, then speeds up and good and nondegenerate solution is found

Degeneracy is Linked to Uniqueness:

"Only decompositions with unique solutions show the problem of degeneracy" –
Zijlstra & Kiers 2002

Unique vs. Nonunique Decompositions

- Unique:

  Parafac, Parafac2, PARATUCK, Constrained Tucker, (and special 2-way models)

  – but only when factor independence conditions fulfilled

- Nonunique:

  Tucker T3, Tucker T2, PCA (!), etc.

Review: Why is Parafac Decomposition Unique?*

Geometrically, the space at each slice of the array differs from that at other slices only by reweighting of axes that are common to all slices. Direction of axes that give best reweighting is often unique.

In contrast, Tucker decomposition allows general linear transformation of axes from one slice to next – angles can change, but now the best-fitting combination of angles and weights is not unique.

Cautions about Uniqueness*

To be scientifically meaningful, uniqueness must be due to deep (low rank) latent structure rather than surface noise or disturbances.

Uniqueness requires that factors have adequate variation independence (e.g., meet Kruskal's k-rank condition) *before* noise is added.

For uniqueness to have scientific meaning, appropriateness of the model's structural form is important

Unique decompositions (e.g., Parafac1 or 2) are more "fragile" than nonunique ones (e.g. Tucker). Too much systematic variation inconsistent with a unique model can have surprising effects -- cause degenerate solutions.

Dealing with Degeneracy**

To prevent degeneracy, one can either

- *Block the correlations.* By imposing factor-independence constraints (i.e., requiring either orthogonality or zero correlation of loading vectors in any one mode) you can prevent an "image vs. anti-image" factor pair from forming since this would require a high negative correlation.

- *Or block the negative factors.* Constraining all factor loadings to be positive (reasonable in many physical applications) is an alternative way of preventing "anti-images", since factor cancellation requires negative loadings.

- *And avoid bad neighborhoods.* "Regularization" (a fit-penalty for diverging loadings) can be used to stabilize algorithms and avoid "swamps".

Puzzling Questions**

The strange phenomenon of degeneracy gives rise to several basic questions, including the following:

1. Why/how do degenerate solutions arise?

2. Why do they seem characteristic of unique-axis models like Parafac?

3. Why do these solutions take the strange form that they do: a severe shear of the underlying factor spaces?

4. Why do degenerate solutions sometimes diverge, with parameter values irresistibly growing toward plus and minus infinity?

5. What is the proper interpretation of ‘swamps’?

A Geometric Approach**

Algebraic analysis by Kruskal et al., ten Berge, Paatero and others has produced valuable progress in our understanding (e.g., see the Annotated Bibliography).

The goal here is to provide a *geometric* account that gives *direct insight* into the degeneracy phenomenon along with simple answers to these 5 questions, at least for the simplest cases.

Naturally, there are many refinements to be worked out for the more complex cases.

Nonetheless, it is hoped that this geometric approach (and its algebraic underpinnings) will provide a useful avenue for developing further understanding of this phenomenon.

An Explanation of Degeneracy in words and pictures **

It may seem surprising for a ‘geometric approach’, but we start with a verbal statement of our conclusion

A verbal explanation of degeneracy**

1. Degenerate solutions arise when more restricted models encounter (slightly) less restricted variation.

   – In our case, the model is Parafac, which has unique fixed axis orientations.

   – The less restricted variation is “Tucker-variation”, in which the orientation or skew of axes can vary across slices of the array.

A verbal explanation of degeneracy**

2. The model responds by 'transforming itself'. It alters its internal representation of the data in such a way that it can mimic the (slightly) more general variation.

   – In our case, the Parafac "base factor space" in one mode (e.g., A) becomes severely *sheared*, while the base space in another mode (e.g., B) becomes *inversely sheared*; this allows the third mode (e.g., C) to apply weights "in between" the shear and anti-shear

   – This lets its axis weights have the effect of varying angles as well as stretching axes.

A verbal explanation of degeneracy**

3. Sometimes, the degeneracy "is never enough", and parameter estimates go to infinity.

   – Some data arrays have a "difficult" combination of stretching and skewing for which the fit can always be improved by making the shear more severe.

   – So the fitting never stops: with each iteration, the degeneracy becomes more and more extreme, yet the improvements get smaller and smaller.

   – The loadings in each base space diverge to plus and minus infinity but the inverse relation between modes makes these effects "cancel" and the model's approximation to the data converges.

Mathematical Explanation of Degeneracy**

A simple mathematical analysis will show how degeneracy allows Parafac to (partly) fit “Tucker-variation”.

This increased flexibility of degenerate Parafac will then be demonstrated with a couple of numerical examples.

But first, we need to:

- define some notation, basic matrices, and terminology
- give a mathematical expression for Tucker-variation
- give a mathematical expression for the Parafac approximation to that variation, first without, and then with, a degenerate transformation of its representation of the factors

Review:**

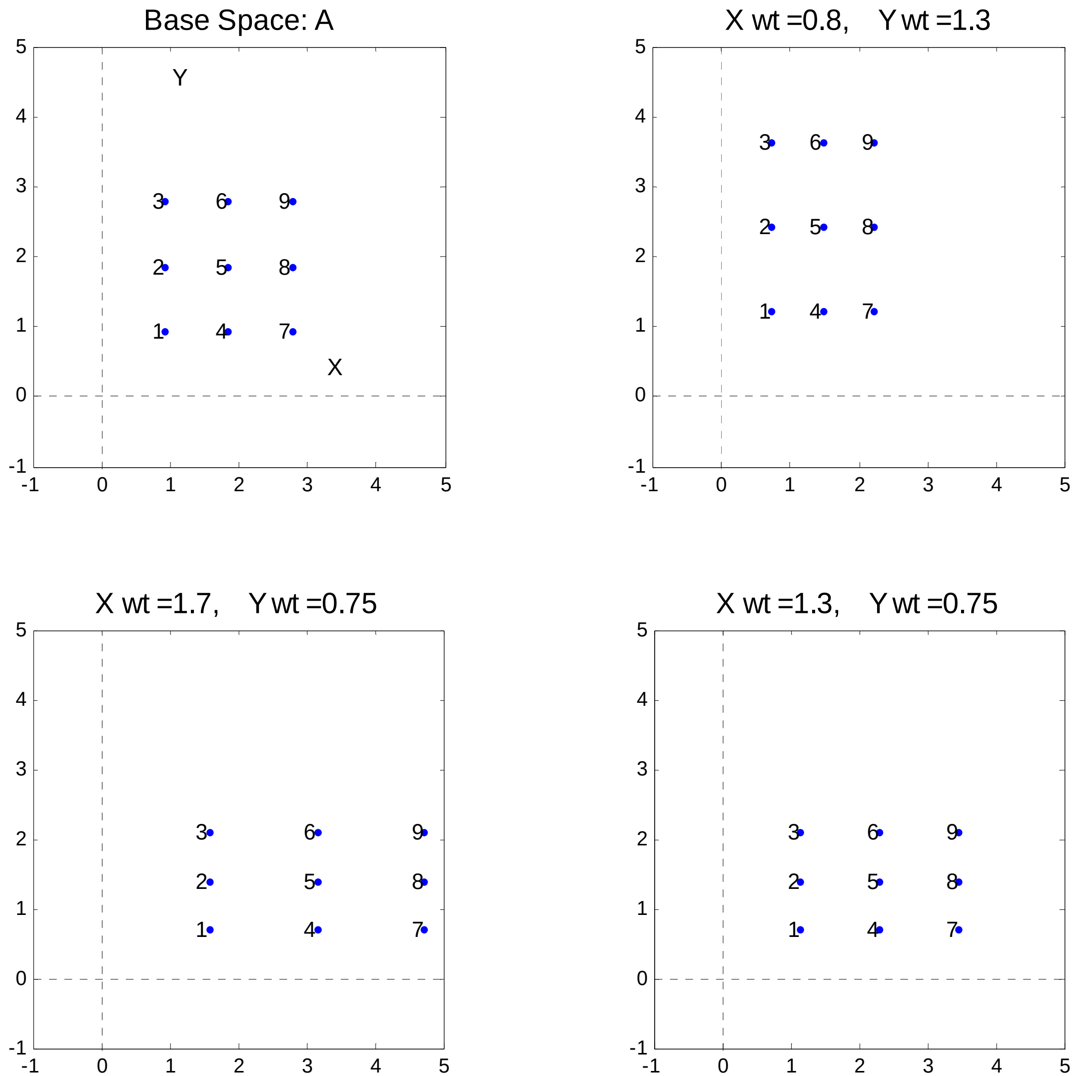
Variations across slices of an array

- Two kinds of variation can occur, alone or in combination: a) variations in the length of factor axes or basis vectors (which can be represented by Parafac and the Tucker models), and/or b) variations in their skew or orientation relative to the points (which can be represented by the Tucker models).

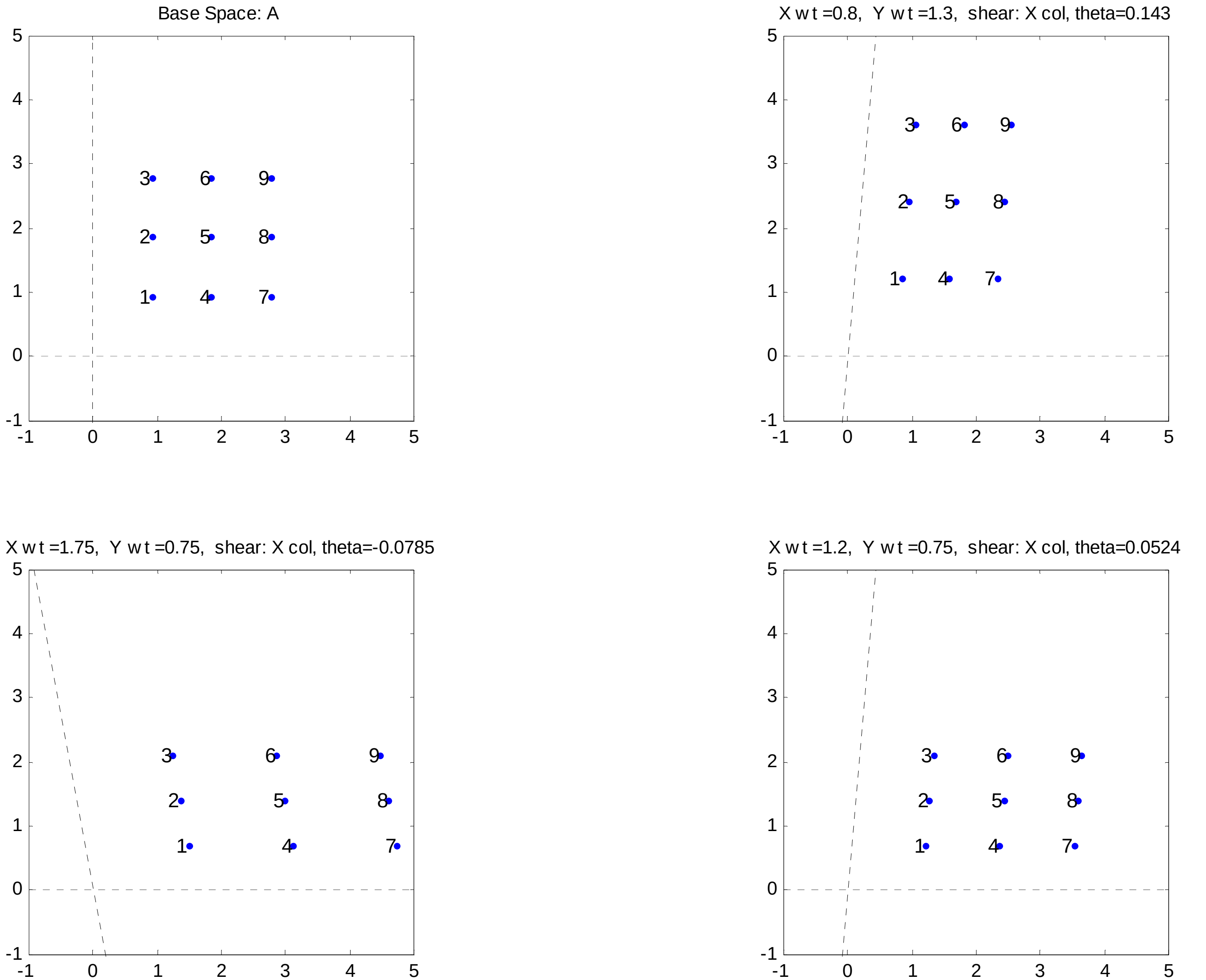
- In psychology, for example, changes in factor length would correspond to increases or decreases in psychological importance or impact of a given dimension, whereas changes in orientation would correspond to changes in “character” or “overtones of meaning” of a dimension.

- Next 2 slides illustrate how the models represent the two types of variation

Picture of Parafac variation: axis reweighting only for 2 factors X and Y**



Picture of “Tucker-variation”: axis weight *plus* skew variation for 2 factors**



Parafac model of cross-slice variation**

Consider an IxJxK array $\underline{\mathbf{X}}$ with elements $\mathbf{x}_{ijk}$ and "slices" $\mathbf{X}_k$. The Parafac model for variation across levels of Mode C can be written

$$\mathbf{X}_k = \mathbf{A}\,\mathbf{D}_k\,\mathbf{B}'$$

where **A** is an IxR matrix of factor loadings for Mode A, and **B** is a JxR matrix of factor loadings for Mode B. The matrix $\mathbf{D}_k$ is a diagonal matrix that contains the row of Mode C factor loadings or weights for the k^{th} slice of the array.

These weights stretch or contract the axes in the space of A and/or B by (typically) different amounts for each value of k.

The Fixed-Axis Model Parafac** Represents Variation Across Slices By $\mathbf{D}_k$:

$$\mathbf{D}_k = \begin{pmatrix} c_{k1} & 0 \\ 0 & c_{k2} \end{pmatrix} = \begin{pmatrix} Xweight_k & 0 \\ 0 & Yweight_k \end{pmatrix}$$

Its diagonal contains row *k* of **C**, the factor loading matrix for Mode C (the array slices).

Thus, the diagonal elements of $\mathbf{D}_k$ are weights that determine the effect of factor 1 and 2 in the *k*th slice. We sometimes refer to them as X and Y weights (as in the previous picture of Parafac axis reweighting).

Terminology Needed for the Tucker Case**

- “shear”
  = a *transformation* of a space in which points are shifted along one axis in progressively greater amounts; the amount is proportional to where they are on the other axis (see picture next slide)

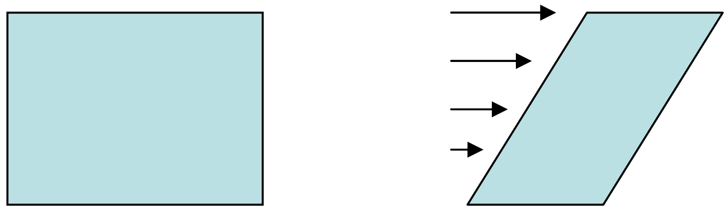
- “skew”
  = axis *slant* in the base space (when axes are plotted in Euclidian coordinates, see picture)

- “base space”
  = the factor space (e.g., for Mode A) and its point configuration. This is represented by the factor loading matrix (e.g., **A**)

Shear and Skew**

- Shear (along x-axis):

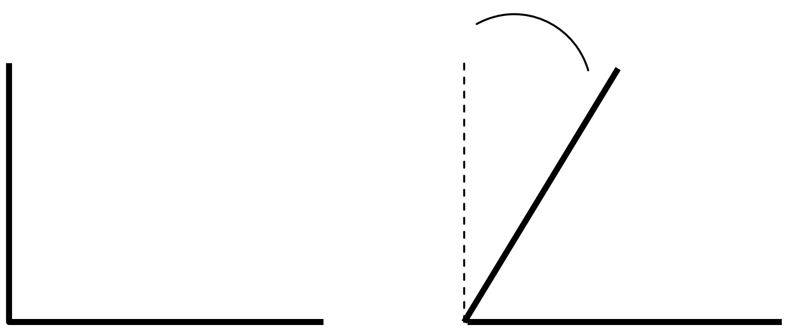
Shear



Matrix

$$\mathbf{S} = \begin{pmatrix} 1 & 0 \\ .7 & 1 \end{pmatrix}$$

- Skew (of y-axis that results):



Skew is changed by shearing the space:**

A shear matrix has the elementary form $\mathbf{S} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$.

The effect of a shear is to add a fixed amount of one dimension to another. This skews it in the direction of the other. For example, the matrix above replaces the original factor 1 with the sum (factor 1 + factor 2). This skews the new factor 1 axis 45 degrees toward factor 2.

To produce different amounts of shear, one simply changes the size of the lower left element. A parameterized version is shown on below, in which the amount of shear is controlled by the value of σ:

$$\mathbf{S}_k = \begin{pmatrix} 1 & 0 \\ \sigma_k & 1 \end{pmatrix}$$

Tucker model of cross-slice variation**

To fit Tucker-variation, we need to replace the "stretch only" Parafac model of cross-slice variation $\mathbf{X}_k = \mathbf{A}\mathbf{D}_k\mathbf{B}'$ with a more general one that allows for angle variation, i.e.,

$$\mathbf{X}_k = \mathbf{A}\mathbf{T}_k\mathbf{B}'$$

where $\mathbf{T}_k = \mathbf{S}_k\mathbf{D}_k$, and $\mathbf{S}_k$ is a shear that varies across *k.*

If, for example, only one axis varies in skew it becomes

$$\mathbf{T}_k = \mathbf{S}_k\mathbf{D}_k = \begin{pmatrix} 1 & 0 \\ \sigma_{k1} & 1 \end{pmatrix}\begin{pmatrix} c_{k1} & 0 \\ 0 & c_{k2} \end{pmatrix} = \begin{pmatrix} c_{k1} & 0 \\ c_{k1}\sigma_{k1} & c_{k2} \end{pmatrix}$$

With no restrictions on T_k, it is the T2 Model; otherwise, T3.

The Parafac 'degeneracy compromise'**

An imperfect compromise is possible. The method is to surround the $\mathbf{D_k}$ by a *fixed* shear and its inverse. Thus, we replace the three-parameter T with the two-parameter approximation

$$\hat{\mathbf{T}}_k = \overline{\mathbf{S}}\,\overline{\mathbf{D}}_k\,\overline{\mathbf{S}}^{-1}$$

For some fixed degree of shear σ , this becomes

$$\hat{\mathbf{T}}_k = \begin{pmatrix} 1 & 0 \\ \sigma & 1 \end{pmatrix} \begin{pmatrix} c_{k1} & 0 \\ 0 & c_{k2} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -\sigma & 1 \end{pmatrix}$$

Which gives a "reweighting" matrix that uses the same two parameters, but has the form

$$\hat{\mathbf{T}}_k = \begin{pmatrix} c_{k1} & 0 \\ \sigma\left(c_{k1} - c_{k2}\right) & c_{k2} \end{pmatrix}$$

Limitation of the compromise[**]

$$\hat{\mathbf{T}}_k = \begin{pmatrix} c_{k1} & 0 \\ \sigma\left(c_{k1} - c_{k2}\right) & c_{k2} \end{pmatrix}$$

This is an imperfect solution because the shear-strength parameter is no longer independent of the stretch parameters.

When there is substantial skew variation, this compromise usually improves fit and so the analysis will go degenerate. But how much of an improvement it provides in a given situation would seem to depend on the particular covariation of the skew and stretch parameters in the data.

The Degenerate Parafac Output**

- Parafac outputs only three factor loading matrices. Thus, the actual output for a degenerate solution is for the regrouped terms shown below as

$$\begin{pmatrix}\mathbf{A} & \bar{\mathbf{S}}\end{pmatrix} \quad \bar{\mathbf{D}}_k \quad \begin{pmatrix}\bar{\mathbf{S}}^{-1} & \mathbf{B}\,'\end{pmatrix}$$

$$\left(\mathbf{A} \begin{pmatrix} 1 & 0 \\ \sigma_A & 1 \end{pmatrix}\right) \begin{pmatrix} w_1 & 0 \\ 0 & w_2 \end{pmatrix} \left(\begin{pmatrix} 1 & 0 \\ -\sigma_A & 1 \end{pmatrix} \mathbf{B}\,'\right)$$

- Renaming the grouped terms as single matrices, we obtain

$$\bar{\mathbf{A}}\,\bar{\mathbf{D}}_k\,\bar{\mathbf{B}}'$$

- Where $\bar{\mathbf{A}} = \mathbf{A}\bar{\mathbf{S}}$ and $\bar{\mathbf{B}} = \mathbf{B}\bar{\mathbf{S}}'^{-1}$.

Summary of What Happens Inside Parafac**

$$\hat{\mathbf{X}}_k = \hat{\mathbf{A}}\,\hat{\mathbf{D}}_k\,\hat{\mathbf{B}}'$$

In standard Parafac, D_k stretches axes as needed to fit the space at level *k* of Mode C of array

$$\hat{\mathbf{X}}_k = \mathbf{A}\,(\bar{\mathbf{S}}\bar{\mathbf{D}}_k\bar{\mathbf{S}}^{-1})\quad\mathbf{B}'$$

For the ‘degeneracy compromise’, Parafac needs to embed the reweighting between shear transformations

let $\hat{\mathbf{T}}_k = (\bar{\mathbf{S}}\bar{\mathbf{D}}_k\bar{\mathbf{S}}^{-1})$ then $\hat{\mathbf{X}}_k = \mathbf{A}\,\hat{\mathbf{T}}_k\,\mathbf{B}'$

A better model, but it must be fit *indirectly*

$$\hat{\mathbf{X}}_k = (\mathbf{A}\bar{\mathbf{S}})\bar{\mathbf{D}}_k(\bar{\mathbf{S}}^{-1}\mathbf{B}')$$

$$\hat{\mathbf{X}}_k = (\mathbf{A}\bar{\mathbf{S}})\,\bar{\mathbf{D}}_k\,(\mathbf{B}\bar{\mathbf{S}}'^{-1})'$$

Since model only allows diagonal D, the shears must be absorbed into fixed part of the model, shearing them instead of D

let $\bar{\mathbf{A}} = (\mathbf{A}\bar{\mathbf{S}})$ $\qquad \bar{\mathbf{B}} = (\mathbf{B}\bar{\mathbf{S}}'^{-1})$

$$\hat{\mathbf{X}}_k = \bar{\mathbf{A}}\bar{\mathbf{D}}_k\bar{\mathbf{B}}'$$

The “degenerate” Parafac dimensions become sheared versions of nondegenerate ones

Note:**

All of the following MATLAB slides (except the simulated “degenerate” Mode A and B space shown in slides 47 and 48) were presented during the original 2004 Tensor Workshop talk. They show the results of numerical experiments carried out in Spring of 2004 in collaboration with Margaret Lundy.

However, here as elsewhere, the *explanation* of these results that was provided orally during the talk has now been provided in written form by inserting additional explanatory slides. Added slides are indicated by the double asterisk on the title.

Numerical Experiments that Test/Demonstrate the Explanation**

Degeneracy is simulated by hypothesizing a "true" structure with added non-Parafac variation and then applying to it the same distortions that occur during actual degenerate solutions.

We shear one "base space" (here taken to be A), reweight its axes, and then inversely shear the result. If **S** represents the shear, we apply

$$\mathbf{A}\,(\overline{\mathbf{S}}\overline{\mathbf{D}}_k\overline{\mathbf{S}}^{-1})$$

The shear reproduces the effect that degeneracy has on Mode A and the inverse shear reproduces the effect on Mode B.

The degree of shear, and the sizes of axis weights, are both manipulated to see their effects. We are particularly interested in the non-stretch changes they produce.

Recall: Parafac approximates Tucker-variation by reweighting sheared axes**

More generally, we simulate skew variation in **A** factor 2 (Y) due to shears of **A** factor 1 (X). The simulated degenerate transformation has the form

$$\mathbf{A} \quad \overline{\mathbf{S}} \quad \overline{\mathbf{D}}_k \quad \overline{\mathbf{S}}^{-1}$$

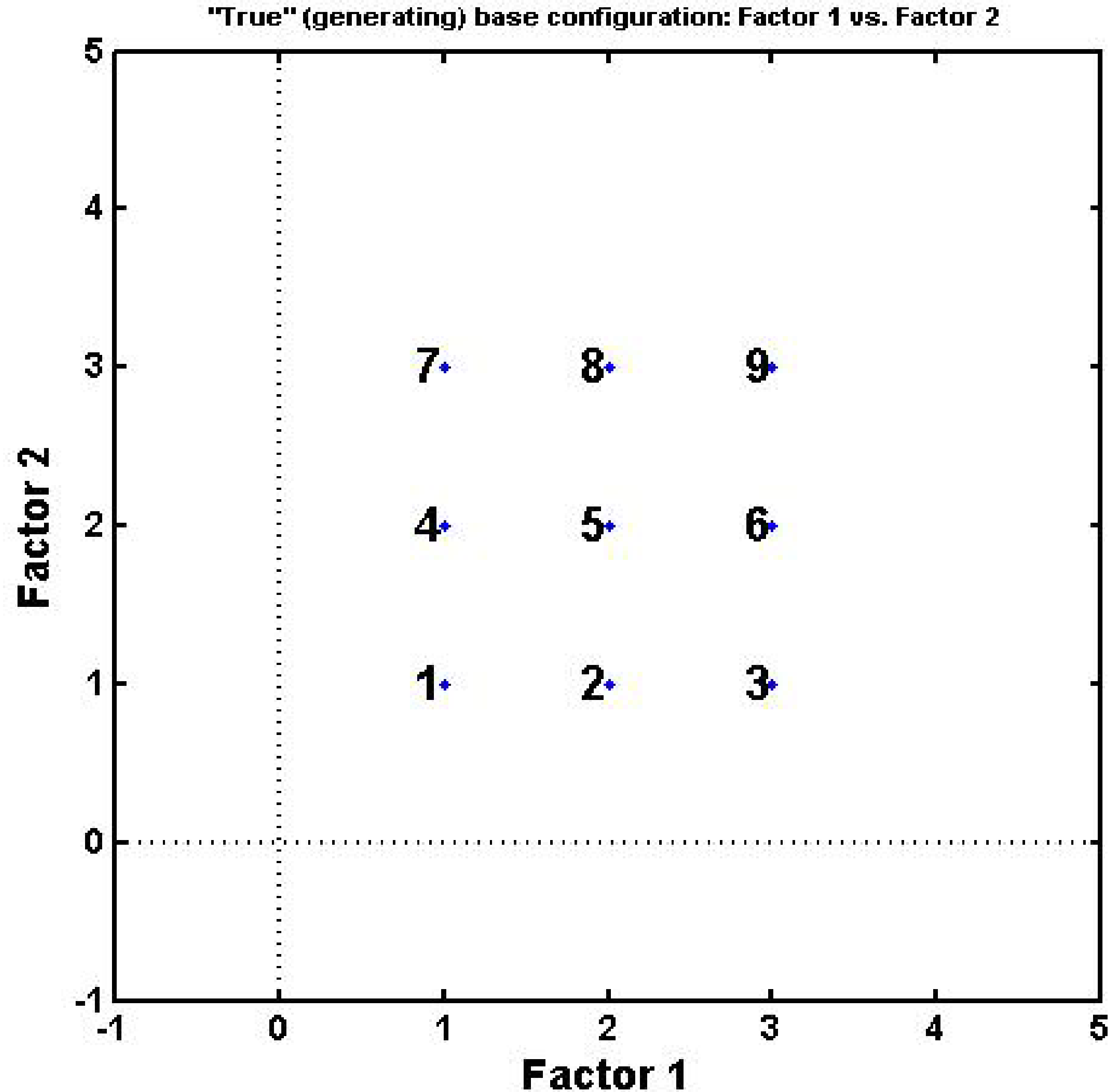
shear * stretch * (un)shear

$$\mathbf{A} \begin{pmatrix} 1 & 0 \\ \sigma_A & 1 \end{pmatrix} \begin{pmatrix} w_1 & 0 \\ 0 & w_2 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -\sigma_A & 1 \end{pmatrix}$$

where w_1 and w_2 are the axis weights for slice k, and σ_A determines the degree of shear.

Some arbitrary choices**

- We start each experiment with the rectangular array of points shown in the next slide for **A** and **B**. This is our “base space”.
  - (Actually, any scatter of points would do; this orderly grid is chosen to make it easier to see subsequent distortions)
- For convenience, we take the configuration to be factor 1 and 2 in Mode A.
  - but it could be any two factors in any mode.



"True" (generating) base configuration: Factor 1 vs. Factor 2
Factor 2
Factor 1
5
4
3
2
1
0
-1
-1
0
1
2
3
4
5
7
8
9
4
5
6
1
2
3

Hypothesized Tucker-variation of the data structure**

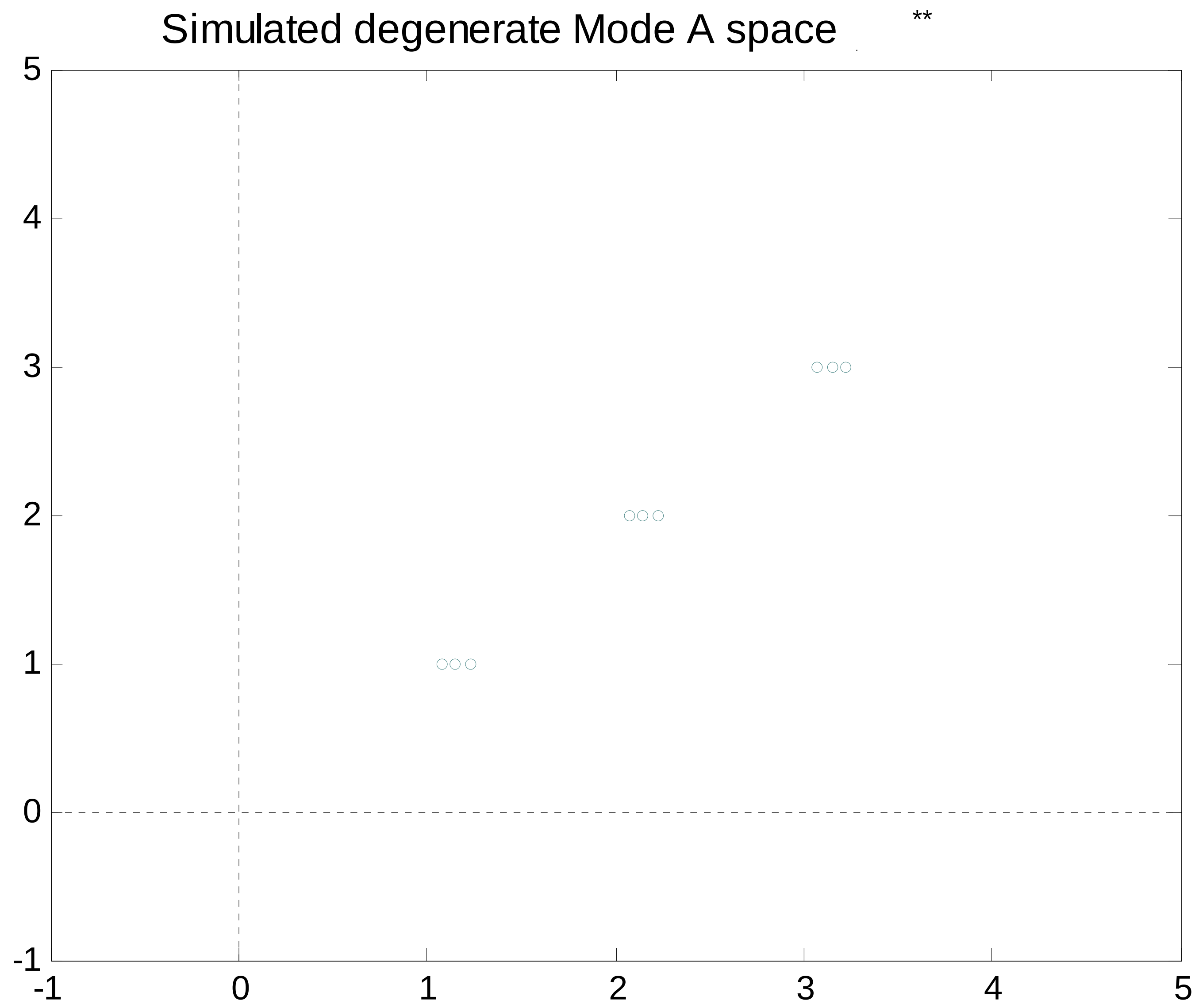
- For simplicity, we hypothesize just one extra source of variation beyond the Parafac model:
  - One axis of one ‘base space’ changes orientation or “skew” across array slices
  - For convenience, we arbitrarily take the mode involved to be Mode A, the axis that changes skew to be factor 2 or “y”, which means the space is sheared along factor 1 or “x”; the factor 2 skew changes across levels (slices) of Mode C.
  - Plots of sheared *common* base spaces will show factors to be highly correlated in opposite directions (this would be result of Parafac analysis of data containing the hypothesized Tucker variation)

Parameters for Demonstration 1:
Slight reweighting of a ‘strong degeneracy’**

Shear imposed on **A**: $\bar{\mathbf{S}} = \begin{pmatrix} 1 & 0 \\ 4.68 & 1 \end{pmatrix}$

Reweighting: $\bar{\mathbf{D}}_1 = \begin{pmatrix} 1.01 & 0 \\ 0 & 0.99 \end{pmatrix}$ and $\bar{\mathbf{D}}_2 = \begin{pmatrix} .99 & 0 \\ 0 & 1.01 \end{pmatrix}$

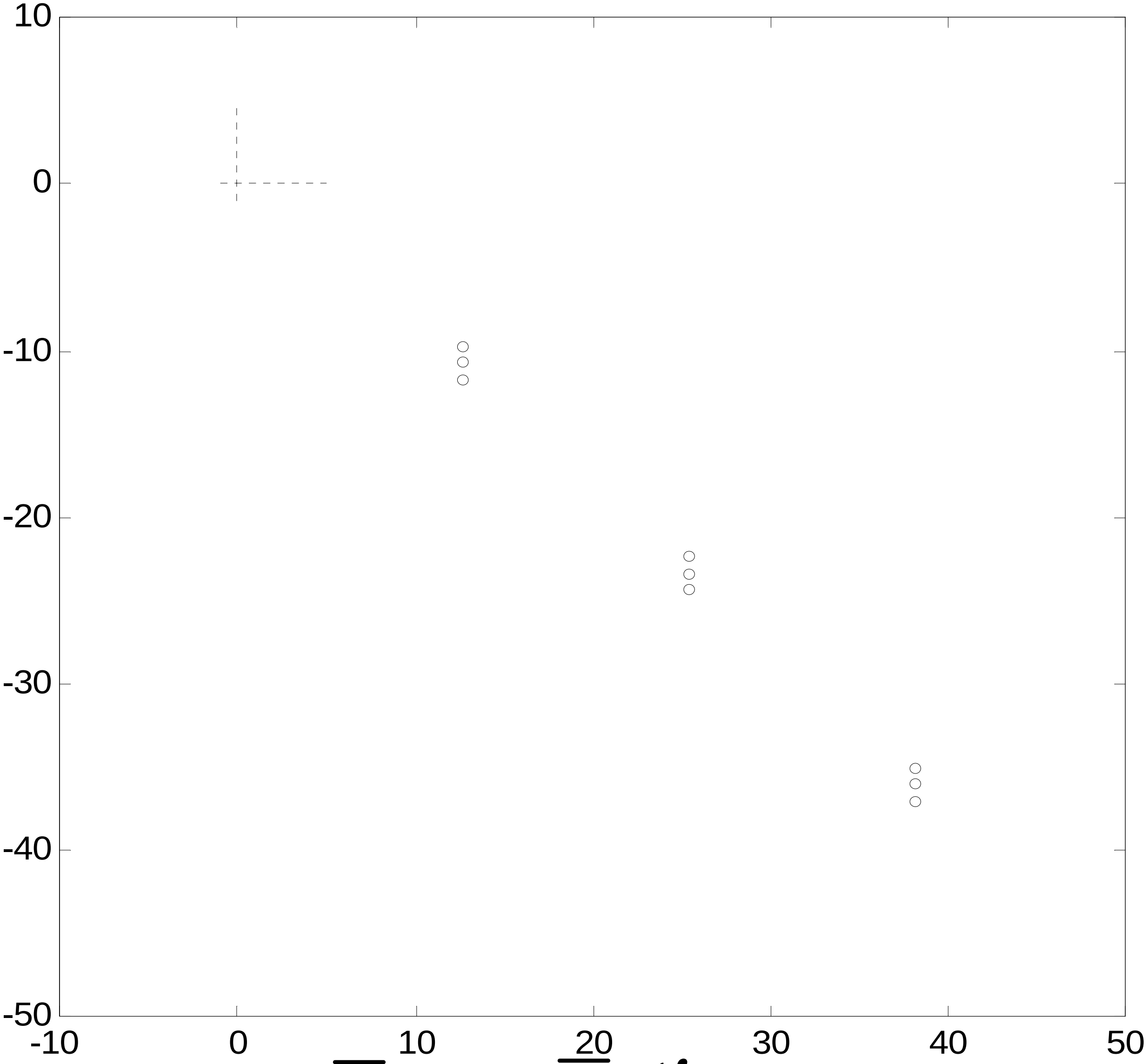
Shear imposed on **B**: $\left(\bar{\mathbf{S}}^{-1}\right)' = \begin{pmatrix} 1 & -4.68 \\ 0 & 1 \end{pmatrix}$



$$\bar{\mathbf{A}} = \mathbf{A}\bar{\mathbf{S}}$$

**

Simulated degenerate Mode B space.



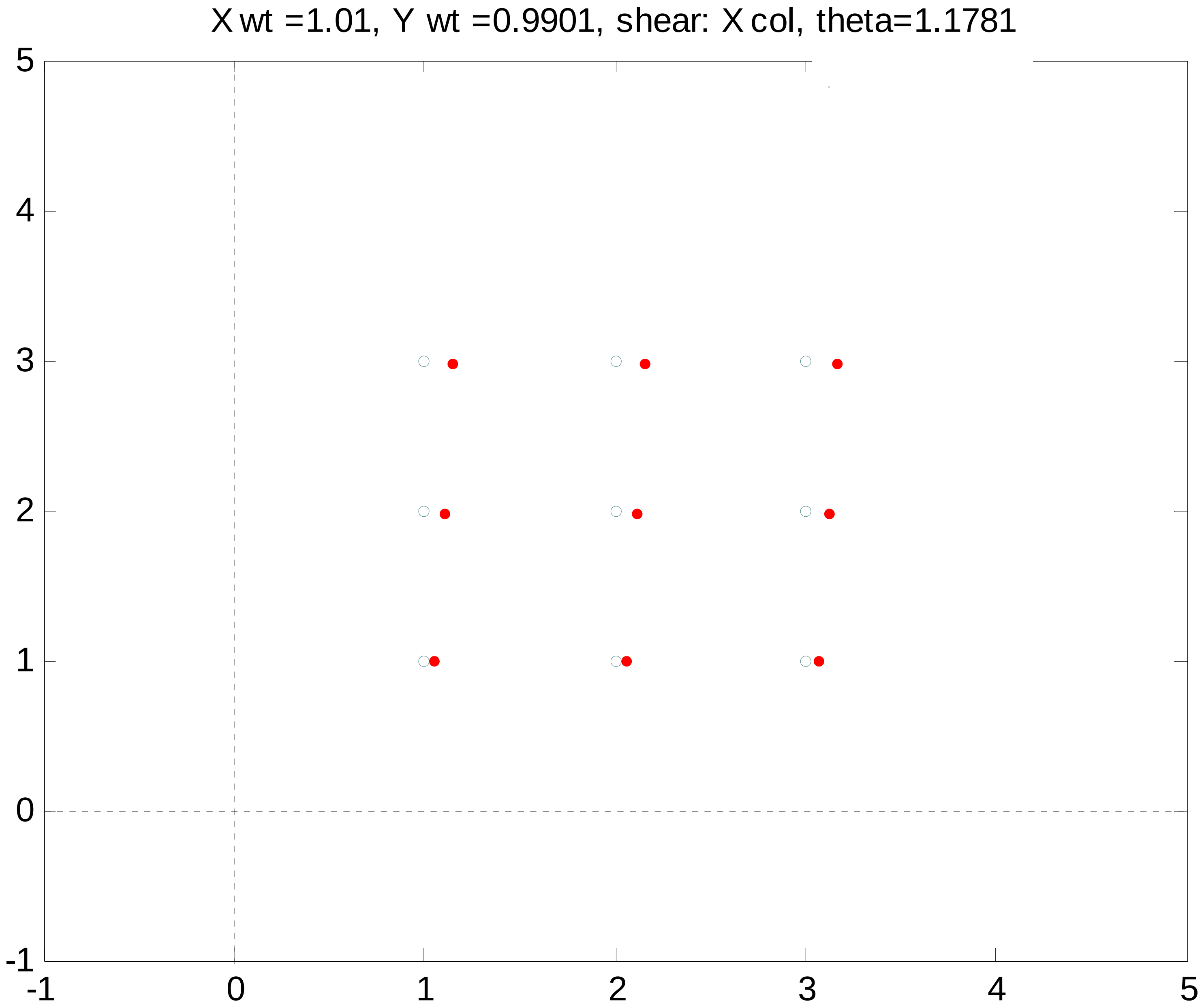
$$\bar{\mathbf{B}} = \mathbf{B}\bar{\mathbf{S}}^{-1\prime}$$

Results:**

The next two plots show the sheared A space weighted 2 different ways. They indicate different Mode A axis orientations across 2 slices of Mode C (Tucker variation which violates the fixed-axis Parafac model), which Parafac can represent by reweighting the degenerate common A and B spaces shown previously.

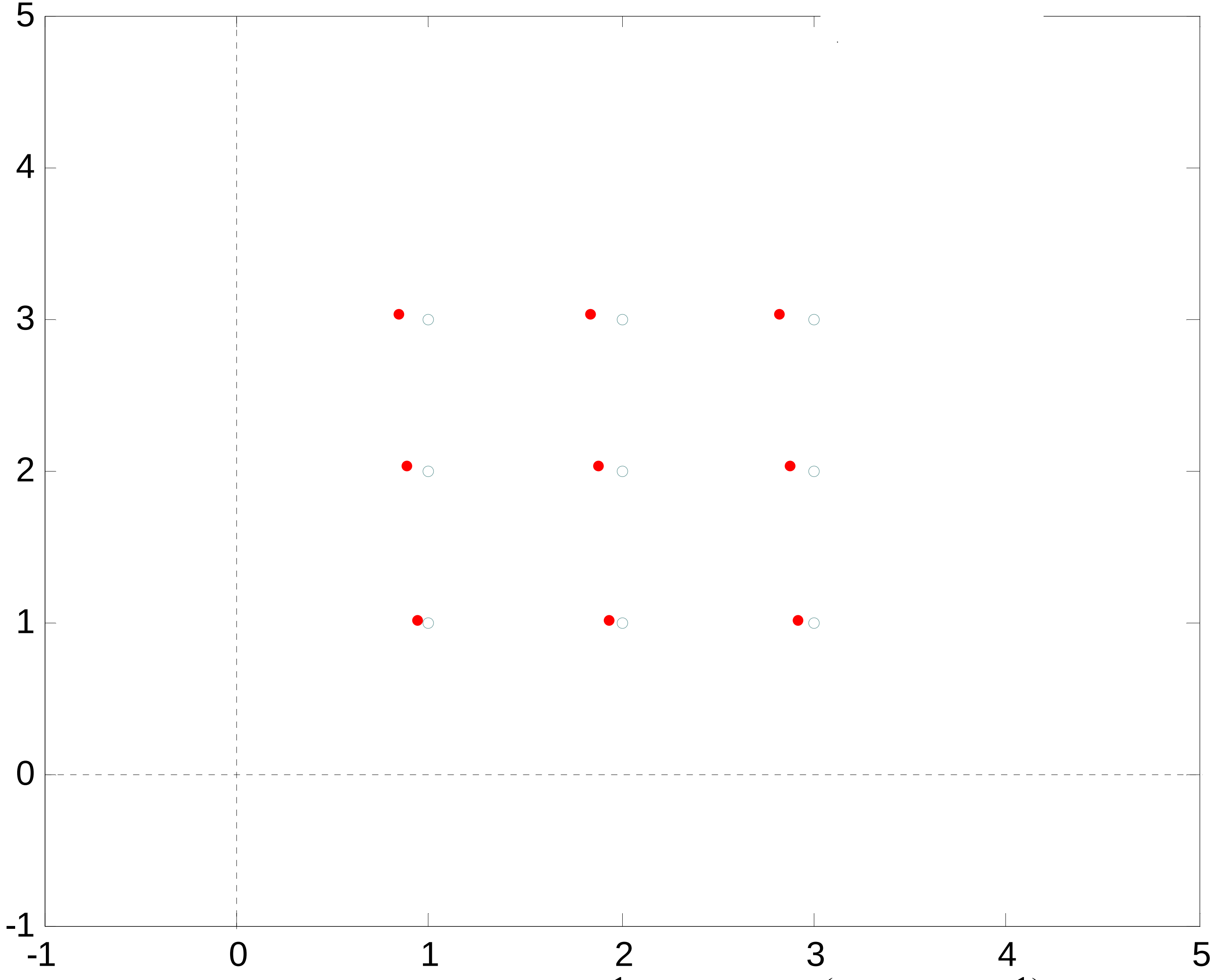
Compare the small circles (both weights equal) to the red points (weights slightly different) to see how small inequalities affect Y axis size and skew.

The first slide shows a rightward skew produced by weights 1.01 and .99. The second shows the opposite skew produced when the weights are reversed (.99 and 1.01).



$$\dot{\mathbf{A}}_1 = \bar{\mathbf{A}}\bar{\mathbf{D}}_1\bar{\mathbf{S}}^{-1} = \mathbf{A}\left(\bar{\mathbf{S}}\bar{\mathbf{D}}_1\bar{\mathbf{S}}^{-1}\right)$$

X wt =0.99, Y wt =1.0101, shear: X col, theta=1.1781



$$\dot{\mathbf{A}}_1 = \bar{\mathbf{A}}\bar{\mathbf{D}}_2\bar{\mathbf{S}}^{-1} = \mathbf{A}\left(\bar{\mathbf{S}}\bar{\mathbf{D}}_2\bar{\mathbf{S}}^{-1}\right)$$

Interpretation**

The synthetic degeneracy shows how weight variation of fixed but highly sheared axes in Modes A and B can imitate what is really angle variation of non-fixed unsheared axes in Mode A of the data.

In this way, degenerate Parafac can fit some Tucker variation.

Demonstration 2:**

Weak vs. Strong Shearing of the Base Space

This next simulation compares effects of a *weak* degeneracy to those of a *strong* degeneracy.
--What kind of data requires strong shear?

The stronger the degeneracy, the more stretch variation is translated into skew variation.
--Strong shear is needed when skew variation is large and stretch variation is small.

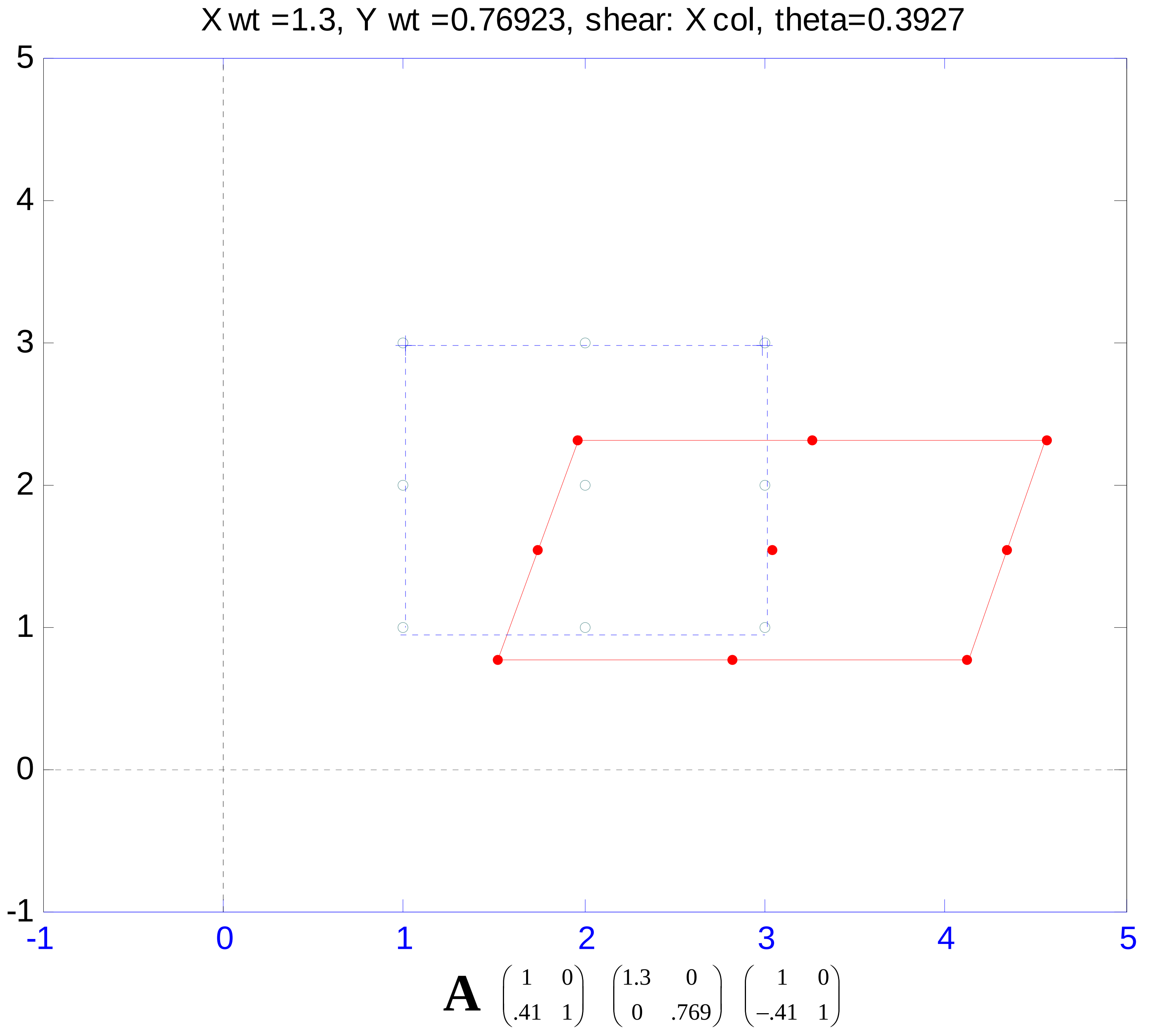
In this first case, we simulate a *weak* degeneracy:**

(a) A mild shear ($\sigma = .41$) is imposed on the A space and a mild inverse-shear ($\sigma = -.41$) is imposed on the B space.

(b) Weights for the axes are chosen to produce a Y skew of approximately 45 degrees. This requires a 1.3 weight on X and a .769 weight on Y.

When the shear, weights, and anti-shear are all absorbed into A, the 45 degree skew appears. In addition, there is substantial stretching of X and compression of Y. The resulting effect is a combination of stretch/compression and skew.

With weak degeneracy, stretches are only *partly* translated into skews.


X wt =1.3, Y wt =0.76923, shear: X col, theta=0.3927
5
4
3
2
1
0
-1
-1
0
1
2
3
4
5
A
1 0
.41 1
1.3 0
0 .769
1 0
−.41 1

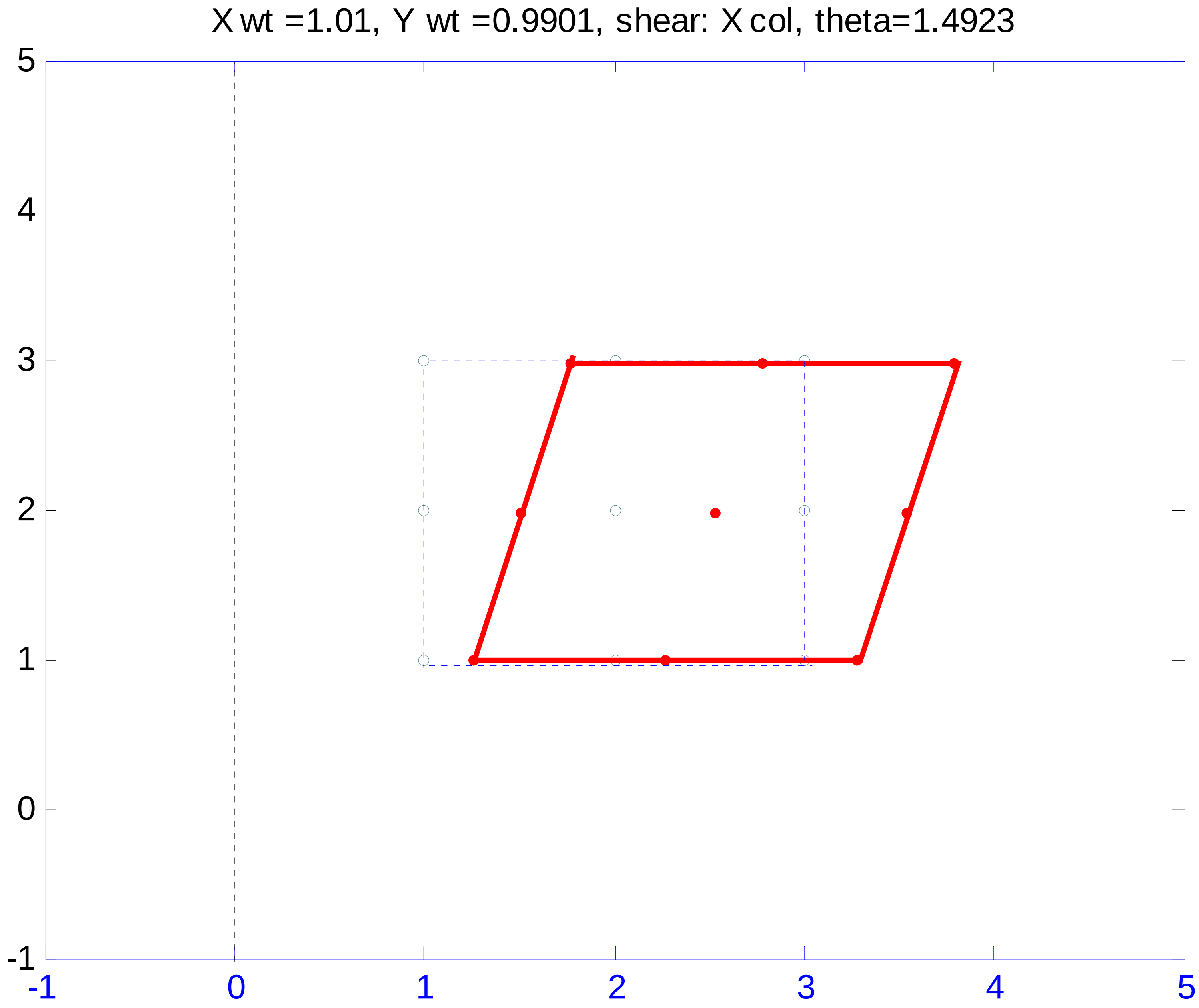
In the second case, we simulate a _**strong**_ degeneracy:**

(a) A very strong shear is imposed on the A space and very strong inverse-shear on B (σ = 12.7).

(b) Again, axis weights were found that produced an approximate 45 degree shear. This turned out to be a very small stretch and compression (X*1.01, Y*.99).

When all the transformations are absorbed into A, only the skew is apparent.

With strong degeneracy, the stretches are almost completely translated into skew.



$$\mathbf{A} \begin{pmatrix} 1 & 0 \\ 12.7 & 1 \end{pmatrix} \begin{pmatrix} 1.01 & 0 \\ 0 & .99 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -12.7 & 1 \end{pmatrix}$$

Interpretation:**

The degree of degeneracy determines what proportion of the Mode C weight-ratio variation is transformed into skew variation, and what remains as axis stretch variation.

Possible Implications**

This suggests that:

In datasets where little variation is due to changes in axis skew, only mild degeneracy will occur. Too much degeneracy would reduce the model's ability to fit the axis-stretch part of the cross-slice variation.

On the other hand, in datasets where most of the variation is due to changes in axis skew, strong degeneracy will be needed to optimize fit.

In the latter case, a well defined optimal solution may exist but be embedded inside a 'swamp'. For such a solution, the recovered Mode A, B, and/or C base space configurations will be moderately to severely sheared, even if convergence is attained (but even this may not occur -- see next slide).

However, these spaces could be linearly transformed into the "true" configurations if only the correct transformation could be determined. Lacking this, transformations that maximize X-Y independence might be used as a "first guess".

Non-converging degenerate solutions**

There are circumstances where no amount of shearing is enough to attain optimum fit, and instead the loadings and shear parameters diverge to infinity.

Even in the very simplified case that we have been considering (where skew varies on only one axis in one mode of a two-factor solution), there is at least one circumstance that will obviously lead to diverging parameter values. This is the case where *all* the cross-slice variation consists of skew variation (i.e., there is no differential axis stretching or contraction). This corresponds to the cross-slice T_k matrices shown below on the left.

$$\mathbf{T}_k = \begin{pmatrix} 1 & 0 \\ \sigma_k & 1 \end{pmatrix} \qquad \mathbf{T}_k = \begin{pmatrix} 1 & 0 \\ \sigma_k(\delta) & 1+\delta \end{pmatrix}$$

In this case, Parafac would want to fit a family of T_k approximations having the form on the right, where delta is the difference between the two c weights.

Fit of the right hand approximation to the left hand pattern is improved whenever the right hand sigma is made larger and delta smaller. Since any given solution can be further improved, the shear diverges "forever".

In Sum:**

- Spaces to be fit can vary by stretching, skewing, or both.
- The more it's due to skewing, the harder it is for fixed-axis models to fit -- so the stronger their shearing must be.
- Some combinations of stretch and skew cannot be well approximated by *any* shearing and/or rotation of the base configurations followed by stretches of the transformed dimensions.
- Trying to fit such patterns produces shearing and anti-shearing that diverges to infinity. This is where the worst degeneracy problems arise.
- One can try increasing the dimensionality of the spatial representation and stretching in higher dimensions…
- **Or:** restate the fitting problem in a suitably restricted form.

Further comments:**
Some Complications and Caveats

Stretch-Skew Covariation**

- The ‘degeneracy compromise’ gives up some accuracy in fitting stretch-variation to improve accuracy in fitting skew-variation. Some benefit of this kind should usually be possible, though not necessarily enough to get full fit.

- However, we have ignored up to now an important consequence of this kind of compromise. The two kinds of variation cannot be modeled independently. The modeled skew is determined by the ratio and size of the modeled stretches.

- Therefore, the benefit of the ‘degeneracy compromise’ depends on how axis skew and stretch happen to actually covary in the data--whether the empirical relationship is consistent with, or at least not severely at odds with, the kind of covariation that the compromise requires.

Multiple axis skews**

- The initial case we have explored here is the simplest possible. When extending it to more realistic situations and higher numbers of factors, we need to consider the interaction of variation in one axis skew with that of another axis in the same mode.

- To do so is beyond the scope of this presentation, but our current analysis suggests that the multi-axis case may have relatively straightforward representation in terms of (a) added shear matrices and/or (b) added linearly dependent dimensions in the base space. The second kind is equivalent to added nonzero cells of the Tucker T3 core.

Cross Mode interactions**

- We have also ignored the question of how representation and approximation of skews in one mode interact with those in another.
- There seems no reason that the effects and equations modeling them cannot be applied to two modes simultaneously. They should be able to be “multiplied together”, since they are all represented by linear and multilinear transformations (as applied, perhaps, to an order three tensor).

PFCORE

[In my talk, this title on an otherwise blank slide was put here as a reminder to say something at the end about PFCORE. However, the reminder didn’t help because I ran out of time(!)

Now, I have tried to make up for that: I have appended a few follow-up slides that tell a bit of the story --and hopefully do it better than I would have then, anyway.]

Analyzing what *causes* a degeneracy -- for even greater insight into your data**

The PFCORE method combines Parafac and Tucker T3 models. It retains the uniqueness of fixed-axis Parafac yet sheds light on the skews that complicate the picture.

A Tucker T3 “core matrix” estimated for the dimensions obtained via constrained Parafac reveals the specific skew-like angle variations of the Parafac factors that led to the degeneracy. These can then be interpreted.

PFCORE turned the degeneracy in the TV data into added insight about individual differences in “sense of humor” .

Applying PFCORE to the TV Data**

- The Parafac solution had 3 factors:

  Humor -- how funny the show was
  Sensitivity -- how gentle vs. crude
  Violence -- how much violence it had

- Degeneracy arose (a) because these dimensions *interacted:* the perceived level of one affected the perceived level of another

- And (b) because the amount of the interaction *changed* across raters. Thus, axis angles would change and the space would skew differently from one rater to the next

- PFCORE uses the Parafac factors to compute a "core array" that reveals these patterns of factor interaction.

Core Array**

The "core array" shows three patterns of *factor interaction*.

These reflect differences in degree of functional relatedness of dimensions.

Geometrically, these become changes of inter-factor angles and produce skews of the factor spaces.

Table 1: Core Matrix for PARAFAC Factors **

C_{F1}		B_{F1} *Funny Shows*	B_{F2} *Sensitive Shows*	B_{F3} *Violent Shows*
A_{F1}	*Humor Scales*	1.058	-.257	.301
A_{F2}	*Sensitivity Scales*	-.084	.102	-.216
A_{F3}	*Violence Scales*	-.095	.093	-.057

First Idealized Individual (Core Slice One)

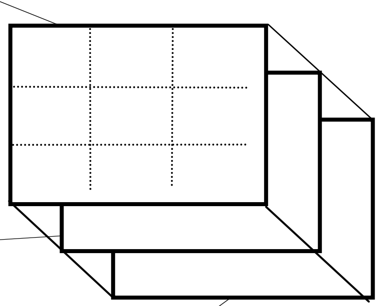
C_{F2}		B_{F1} *Funny Shows*	B_{F2} *Sensitive Shows*	B_{F3} *Violent Shows*
A_{F1}	*Humor Scales*	.014	-.013	.089
A_{F2}	*Sensitivity Scales*	-.071	.951	.113
A_{F3}	*Violence Scales*	.016	-.049	-.025

Second Idealized Individual (Core Slice Two)

C_{F3}		B_{F1} *Funny Shows*	B_{F2} *Sensitive Shows*	B_{F3} *Violent Shows*
A_{F1}	*Humor Scales*	-.125	.485	-.667
A_{F2}	*Sensitivity Scales*	.244	-.099	.194
A_{F3}	*Violence Scales*	.135	.087	1.061

Third Idealized Individual (Core Slice Three)

Idealized person 1 rates violent shows as more humorous, whereas idealized person 3 rates sensitive shows as more humorous, but violent shows as less humorous.



From: Lundy et al. 1989

First Slice of Core from PFCORE analysis of TV data**

Basic effect of humor

Interaction:
Violent shows more funny

C_{F1}		B_{F1} *Funny Shows*	B_{F2} *Sensitive Shows*	B_{F3} *Violent Shows*
A_{F1}	*Humor Scales*	1.058	-.257	.301
A_{F2}	*Sensitivity Scales*	-.084	.102	-.216
A_{F3}	*Violence Scales*	-.095	.093	-.057

First Idealized Individual (Core Slice One)

Second Slice of Core Array **

Basic effect of sensitivity

	C_{F2}	B_{F1} *Funny Shows*	B_{F2} *Sensitive Shows*	B_{F3} *Violent Shows*
A_{F1}	*Humor Scales*	.014	-.013	.089
A_{F2}	*Sensitivity Scales*	-.071	.951	.113
A_{F3}	*Violence Scales*	.016	-.049	-.025

Second Idealized Individual (Core Slice Two)

Third Slice of TV Data Core Array**

Interaction:
Violent shows *less* funny

	C_{F3}	B_{F1} *Funny Shows*	B_{F2} *Sensitive Shows*	B_{F3} *Violent Shows*
A_{F1}	*Humor Scales*	-.125	**.485**	**-.667**
A_{F2}	*Sensitivity Scales*	.244	-.099	.194
A_{F3}	*Violence Scales*	.135	.087	1.061

Third Idealized Individual (Core Slice Three)

Interaction:
Sensitive shows *more* funny

Basic effect of violence

Interpretation of PFCORE results**

Idealized person 1 rates violent shows as more humorous, whereas idealized person 3 rates sensitive shows as more humorous, but violent shows as less humorous.

Conclusion**

Some people think violence can be funny. They probably like "The Three Stooges".

Others think violence is *no laughing matter.* To them, "The Three Stooges" are politically incorrect.

Overcoming degeneracy can be a wonderful thing.

Thank You

Postscript:**

There is a different mathematical perspective (first developed by Kruskal) that has played an important role in the study of degeneracy…

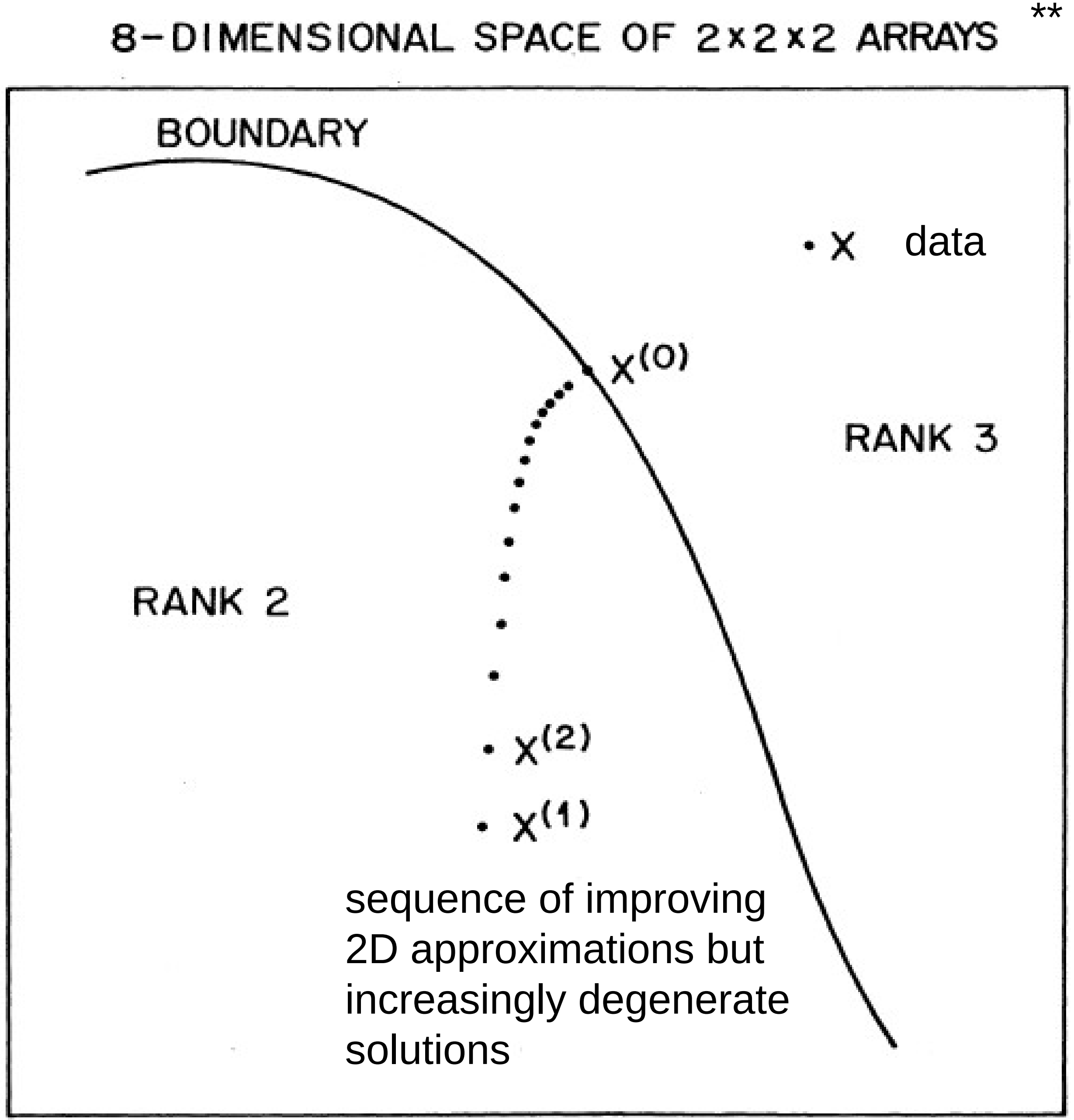
A Solution "Going Degenerate"**

Kruskal et al. (1989) examined the process of a two-factor Parafac solution becoming progressively more degenerate in terms of its location in, and progress through, the rank-2 region of the 8-dimensional space of possible 2x2x2 arrays.

As the Parafac rank-2 approximation gets better, it gets closer to the data in the 8-D space. But (by hypothesis) the data array is embedded in the rank-3 region, and hence is unreachable.

As Parafac progresses closer to the rank boundary, the solution must take on an increasingly degenerate form.

**



$x^{(0)}$ is the boundary point closest to x

Kruskal et al., 1989

Degeneracy appears when Parafac's sequence of improving approximations approaches a higher-rank region in array-space**

- Subsequent work by ten Berge, Paatero and others have proven this correct, and extended it in interesting ways.
- This also accounts for "swamps"– as was conjectured by Mitchell, Burdick, and Ryan, and subsequently demonstrated by Paatero.
- For more information on this "array-space" perspective, see articles listed in the annotated bibliography on degenerate Parafac solutions that is provided as a supplement to these slides.