Multilinear modeling of brain imaging data

Sungjin Hong

University of Illinois, Urbana-Champaign

TRICAP 2009, Vall de Núria, Spain, June 14 – 19

Thanks to

- Richard Harshman
- Michelle Voss
- Alwin Stegeman
- Lieven de Lathauwer

Overview

- PCA vs. ICA
- Algorithms for ICA
- Parafac vs. three-mode ICA (tPICA)
- Three-mode ICA using Comon4 vs. Fixed point
- Typical brain imaging data
- Two-mode simulation
- Three-mode simulation
- Volumic images too big to fit by Parafac (and Parafac2)
- Conclusions

Principal component analysis

- R principal components maximally account for sum of squares of a data matrix $\mathbf{X}_{I\times J}$ such that:

$$\mathbf{X} = \mathbf{A}_{I\times R}\mathbf{B}'_{R\times J} + \mathbf{E}_{I\times J}, \quad R \le \min\left(I, J\right)$$

$$\mathbf{A}'\mathbf{A} = I\mathbf{I}, \quad \mathbf{B}'\mathbf{B} = I^{-1}\operatorname{diag}\left(\lambda_1, ..., \lambda_R\right)$$

$\lambda_1, \ldots, \lambda_R$ are successively maximum

$$\sum_{r=1}^{R}\lambda_r = \operatorname{tr}\left(\mathbf{X}'\mathbf{X}\right), \text{ if } R = \operatorname{rank}\left(\mathbf{X}\right)$$

- The rank-reduced principal component model is equivalent to truncated SVD, and it is unique --- Eckart-Young theorem:

$$\mathbf{X} = \underset{I\times R}{\mathbf{U}_1}\ \underset{R\times R}{\mathbf{S}_1}\ \underset{R\times J}{\mathbf{V}'_1} + \underset{I\times R_2}{\mathbf{U}_2}\ \underset{R_2\times R_2}{\mathbf{S}_2}\ \underset{R_2\times J}{\mathbf{V}'_2}, \quad R_2 = J - R$$

$$\mathbf{A} = \sqrt{I}\mathbf{U}_1, \quad \mathbf{B} = \sqrt{I}^{-1}\mathbf{V}_1\mathbf{S}_1, \quad \operatorname{tr}\left(\mathbf{S}_1^2\right) = \sum_{r=1}^{R}\lambda_r$$

- Principal components can be rotated without loss of fit:

$$\mathbf{X} - \mathbf{E} = \mathbf{A}_0\mathbf{B}_0' = \mathbf{A}_0\mathbf{T}\left(\mathbf{B}_0\mathbf{T}\right)' = \mathbf{AB}', \quad \mathbf{TT}' = \mathbf{I}$$

  where $\mathbf{A}_0$ and $\mathbf{B}_0$ indicate the unrotated principal components and loading matrices

- The rotation matrix $\mathbf{T}$ is sought to optimize a criterion (e.g., Varimax)
- $\mathbf{A}$ is a new basis for the space spanned by the columns of $\mathbf{A}_0$, approximating so called “simple structure” (Thurstone, 1947)
- Columns of $\mathbf{T}$ can be non-orthogonal, producing a set of oblique basis axes

- The independent component model takes the same bilinear form as the PC model, with an additional condition on $\mathbf{A}$ --- its columns are mutually independent

- If $\mathbf{A}$ is Gaussian, PCA gives independent components and so does any orthogonal rotation (ICs not unique) --- so, sources are assumed to be non-Gaussian in ICA

- Central limit theorem --- a linear combination of independent non-Gaussian random variables is more Gaussian

- Thus, $\mathbf{T}$ is sought to maximize a criterion of non-Gaussianity as a means of maximizing the independence

- Since the non-Gaussianity is scale free, data are projected on a pre-determined R-D space --- called whitening

Comon (1994)

- Fourth-order cumulants of (realizations of) $\mathbf{A}$ is defined as:

$$\mathrm{cum}\left(a_i, a_j, a_k, a_l\right) = E\left(a_i a_j a_k a_l\right) - E\left(a_i a_j\right) E\left(a_k a_l\right)$$
$$- E\left(a_i a_k\right) E\left(a_j a_l\right) - E\left(a_i a_l\right) E\left(a_j a_k\right)$$

yielding an $R \times R \times R \times R$ array of cumulants $\mathrm{cum}(\mathbf{A})$, given an $\mathbf{A}$

- If $\mathbf{A}$ is non-Gaussian and independent, $\mathrm{cum}(\mathbf{A})$ becomes diagonal, while if $\mathbf{A}$ is Gaussian, $\mathrm{cum}(\mathbf{A})$ becomes all zeros
- Non-Gaussianity is maximized by minimizing sum of squares of off-diagonal elements of $\mathrm{cum}(\mathbf{A})$ (will be called Comon4) --- equivalent to maximizing SS of the diagonal elements (univariate kurtosis)

Fast ICA (Hyvärinen, 1999)

- To find the rotation matrix, fast ICA uses a non-Gaussian function:

$$J(a_r) = \left[E\left(G(a_r)\right) - E\left(G(z)\right) \right]^2$$

where a_r is the r-th rotated component (i.e., IC) and z is a standard Gaussian variable, and G is some non-quadratic function (typically, kurtosis or hyperbolic tangent)

- $J(a_r)$ is maximized over $\mathbf{T}$ for one component at a time; or can be done simultaneously for all components
- Convergence is in theory guaranteed to be at least quadratic; also a stabilized version can be used when convergence is uncertain
- This ICA estimation will be called “fixed point”

Probabilistic ICA (PICA, Penny et al., 2001, Beckmann & Smith, 2004)

- Like PPCA (Tipping & Bishop, 1999), PICA parametrically defines the error distribution as Gaussian and spherical:

$$\mathbf{X} = \mathbf{A}\mathbf{B}' + \mathbf{E}, \quad \mathbf{E} \sim N\left(\mathbf{0}, \sigma^2 \mathbf{I}\right)$$

- Consequently, PICA solution is adjusted for the error variance as:

$$\mathbf{B} = \mathbf{V}_1 \left(I^{-1}\mathbf{S}_1^2 - \sigma^2 \mathbf{I}\right)^{1/2} \mathbf{T}$$

$$\mathbf{A} = \mathbf{U}_1 \mathbf{S}_1 \left(I^{-1}\mathbf{S}_1^2 - \sigma^2 \mathbf{I}\right)^{-1/2} \mathbf{T}$$

- To make the error structure follow the sphericity assumption, $\mathbf{x}_j$ ($j = 1,\ldots,J$) may be transformed according to some auto-regressive model --- Beckmann et al. called “temporal” whitening to distinguish from the PCA step as spatial whitening

Three-mode Parafac

- Parafac describes $I \times JK$ matricized data as:

$$\mathbf{X}_{I \times JK} = \mathbf{A}\left(\mathbf{C} \odot \mathbf{B}\right)' + \mathbf{E}$$

- With typical functional imaging data, indices I, J, K may indicate vectorized voxels, scans responding to systematically designed stimuli, and different brains, respectively

- ALS used to find $\mathbf{A}, \mathbf{B}$ and $\mathbf{C}$ --- with compression in the voxel mode since $I \gg JK$

Three-mode ICA

- Beckmann & Smith (2005) proposed a three-mode extension of PICA and called it tensor PICA
- Like the two-mode case, tensor PICA takes the same model form as Parafac with the independence condition imposed in the voxel mode
- An alternating estimation procedure (tPICA_fixedp):

a) fast ICA fit to find $\mathbf{A}_{I\times R} = \mathbf{A}_{0(I\times R)}\mathbf{T}_{R\times R}, \quad \mathbf{X}_{I\times JK} = \mathbf{A}_0\mathbf{M}_0' + \tilde{\mathbf{E}}$

b) (1) JK levels are restored as $\mathbf{M} = \mathbf{M}_0\mathbf{T}$; (2) each column of $\mathbf{M}$ is matricized into $J \times K$; and then (3) rank-1 approximated to yield columns of $\mathbf{B}$ and $\mathbf{C}$

Steps (a) and (b) iterate until convergence

- Same two-step procedure also used with Comon4 for step (a) (tPICA_Comon4)

Typical brain imaging data

- fMRI images (46 x 55 x 46 voxels of 4mm resolution) obtained from one brain in response to visual stimuli (block design of rest, face, rest, house, rest, …)

- After taking intra-cranial voxels and vectorizing them, 55544 voxels x 180 scans obtained --- a matrix fittable by PCA or ICA

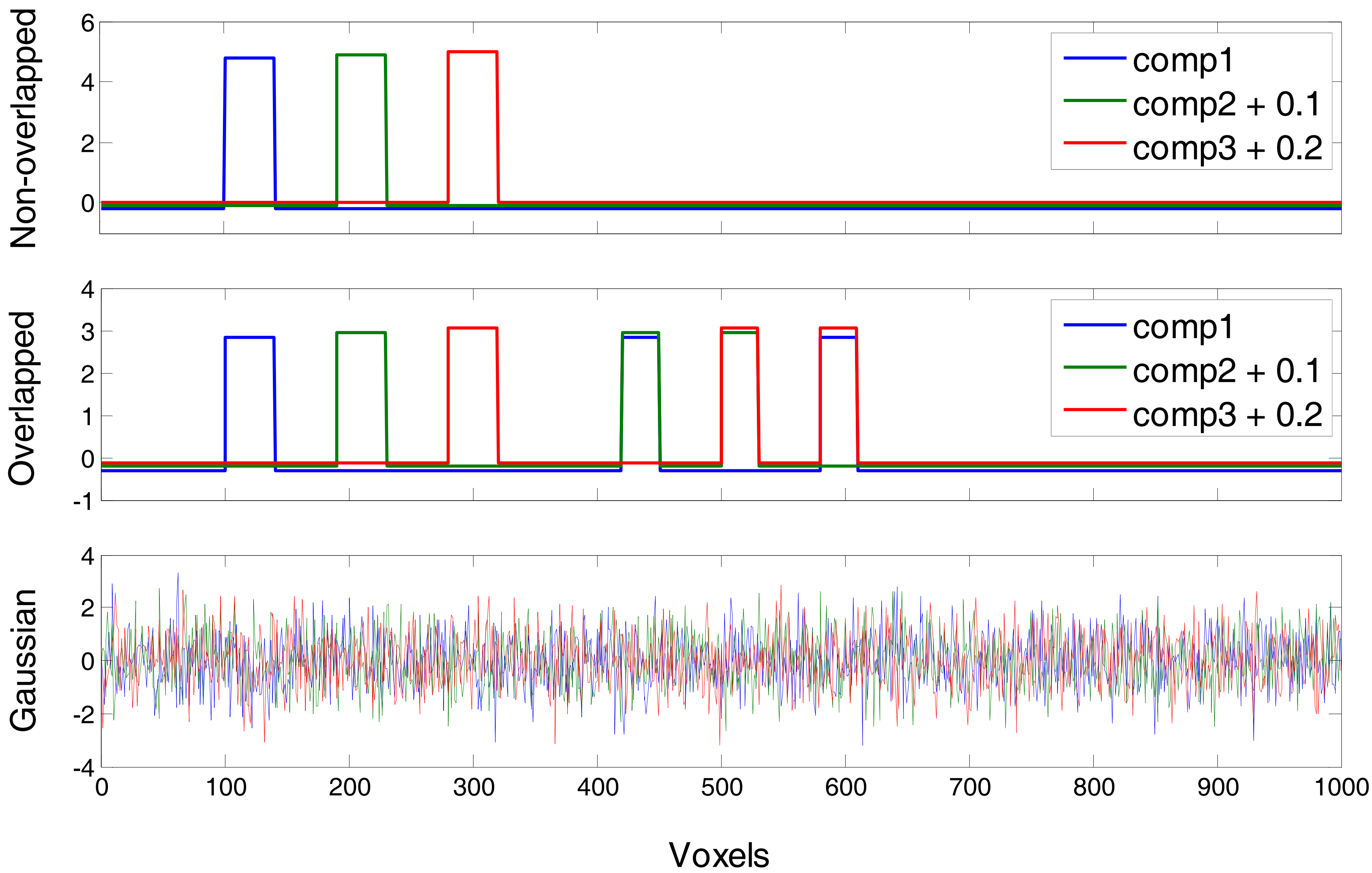
- If we obtain the same data from many brains (e.g., 20 young and old men and women), the data become an array of 55544 voxels x 180 scans x 20 subjects --- maybe fit by three-mode Parafac or ICA (extended to trilinear structure)

- Independence is typically imposed in the voxel mode (spatial ICA) or sometimes in the scan mode (temporal ICA)

Two-mode simulation

- ICA often produces non-overlapped activation pattern --- similar to "interesting structure" produced by projection pursuit
- Intuitively, ICA would successfully recover non-overlapped sources, but not overlapped ones
- The component analysis of brain imaging data does not use information on the voxel location, and so three kinds of source components $\mathbf{A}_{1000\times3}$ were generated as vectors:
  1. non-overlapped voxel activation (12%)
  2. partly overlapped (21%)
  3. Gaussian

Three kinds of components



Two-mode simulation

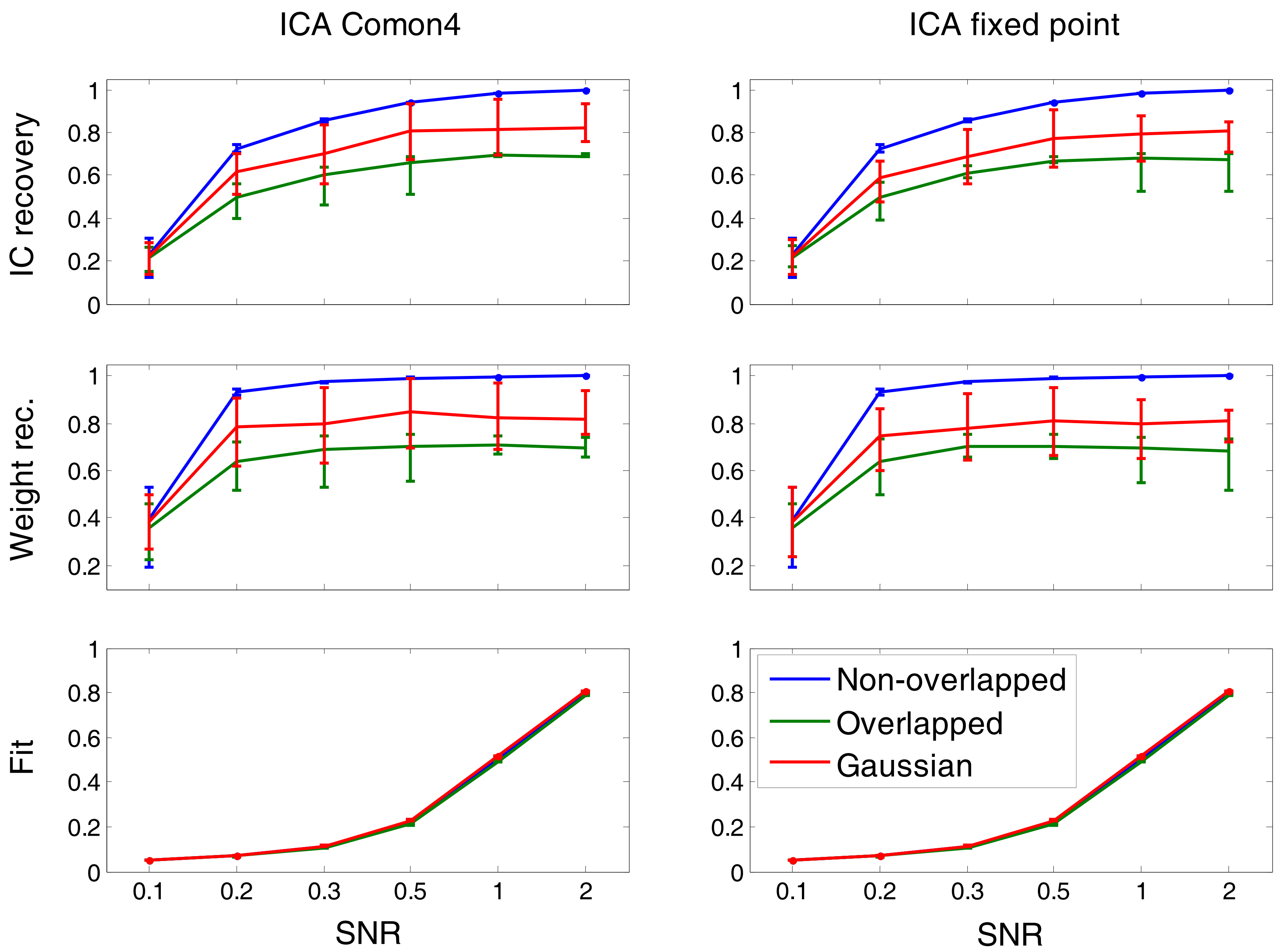
- Scan-mode loading matrix $\mathbf{B}_{100\times 3}$ randomly sampled from $\mathrm{U}(0,1)$ and then column mean centered --- "demeaning"
- Gaussian random errors added so that SNR became 0.1, 0.2, 0.3, 0.5, 1 and 2, which is defined as:

$$\mathrm{SNR} = \|\mathbf{X} - \mathbf{E}\| / \|\mathbf{E}\|$$

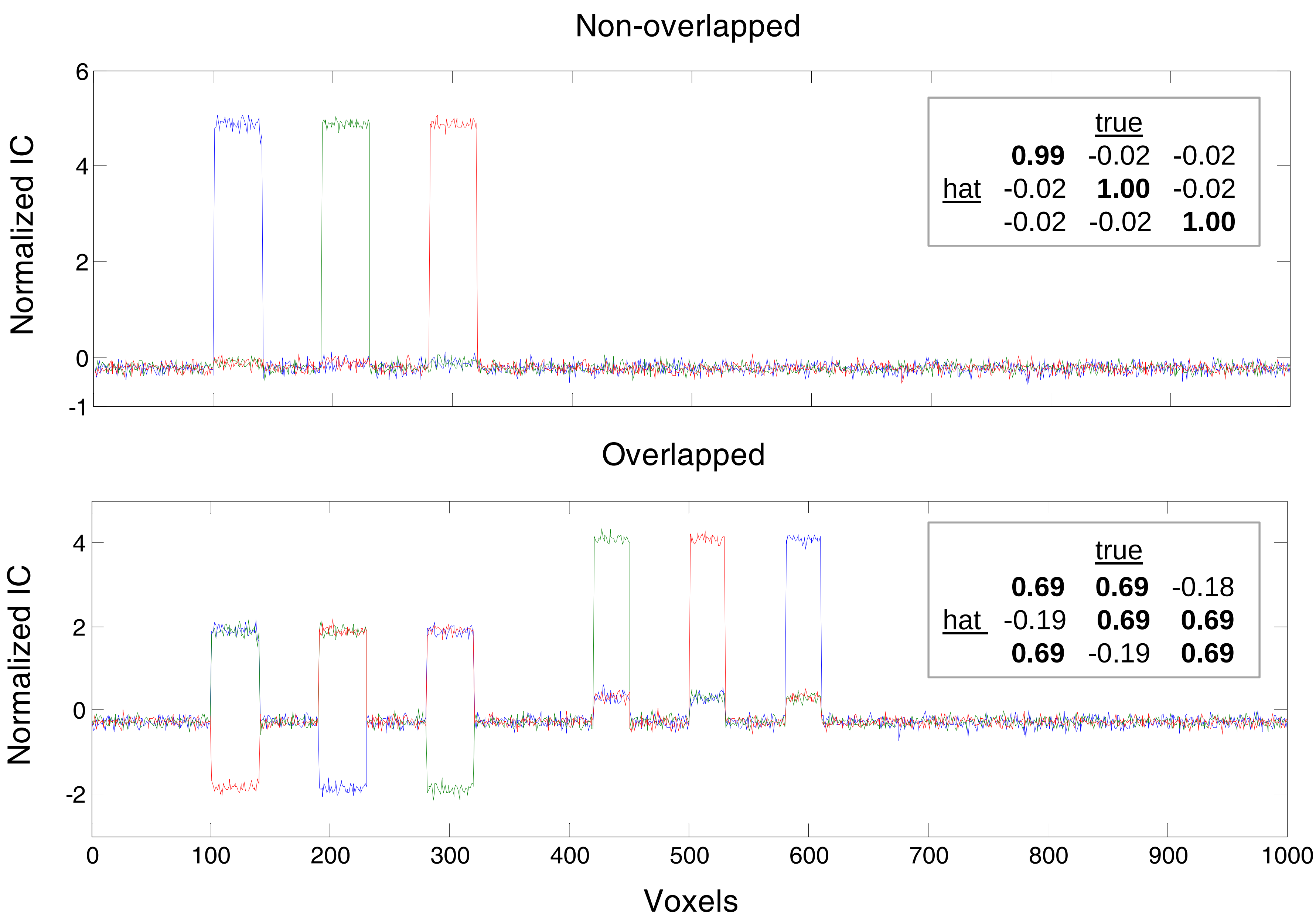
These SNRs are similar to those used in Stegeman (2007) for comparison of three-mode ICA and Parafac

- 30 datasets were replicated per data condition, producing 540 datasets (= 3 types of sources x 6 SNRs x 30 replications)
- Each dataset was fit once by ICA_Comon4 and 10 times by ICA_Fixedp (random starts)

Two-mode ICA recovery and fit



Typical recovered ICs from two-mode data (SNR = 2)



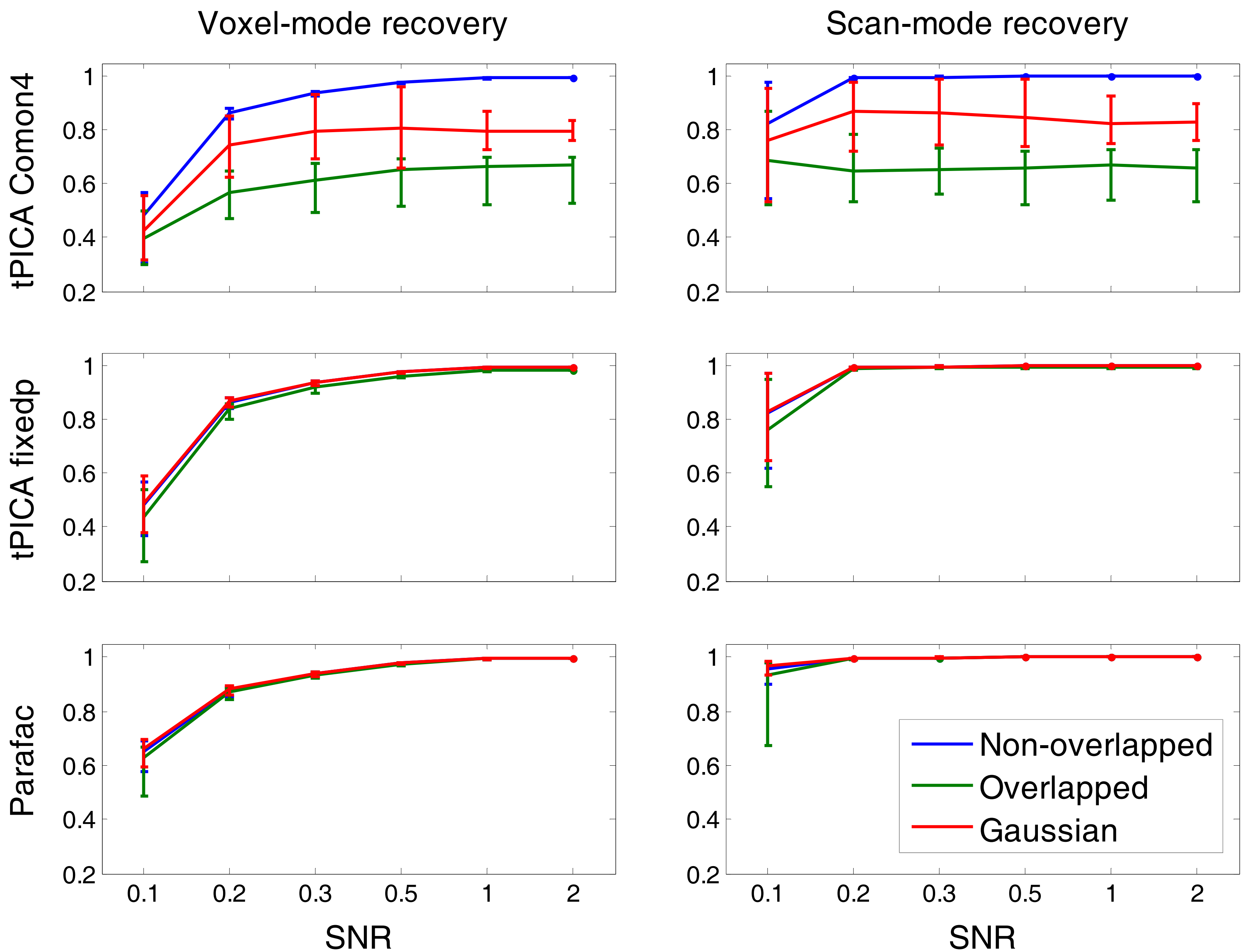
Two-mode results

- As expected, neither algorithm accurately recovered the overlapped and the Gaussian sources at all SNR levels
- Both Comon4 and Fixed point accurately recovered the non-overlapped sources for SNR as little as 0.3 (= expected variance of the model ~8%; $r = 0.85$ in the voxel mode)
- Both Comon4 and Fixed point consistently produced biased components such that highest activations were estimated as non-overlapped on one of the 2 overlapped parts, while the non-overlapped parts in the true sources were estimated as overlapped with moderate sized activations
- Fit didn't vary much by the types of sources

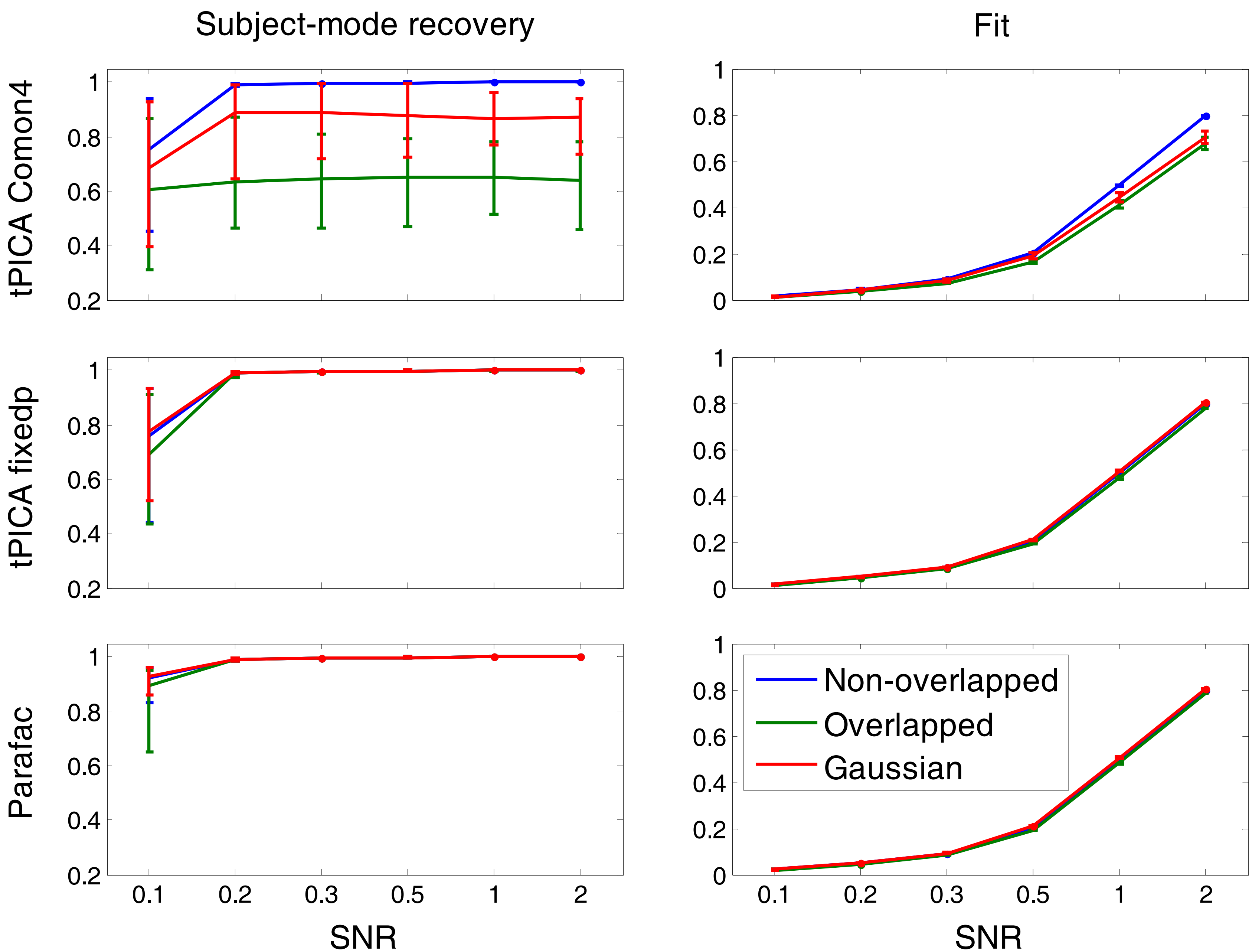
Three-mode simulation

- Scan-mode weights $\mathbf{B}_{20\times3}$ sampled from $\mathbf{U}(0,1)$ and column-mean centered
- Subject-mode weights $\mathbf{C}_{15\times3}$ sampled from $\mathbf{U}(0,1)$
- All other data conditions were the same as the two-mode simulation
- Each of 540 datasets were fit by three-mode ICA_Comon4, ICA_fixedp and Parafac --- best fitting solutions out of 10 random starts were chosen for ICA_fixedp and Parafac

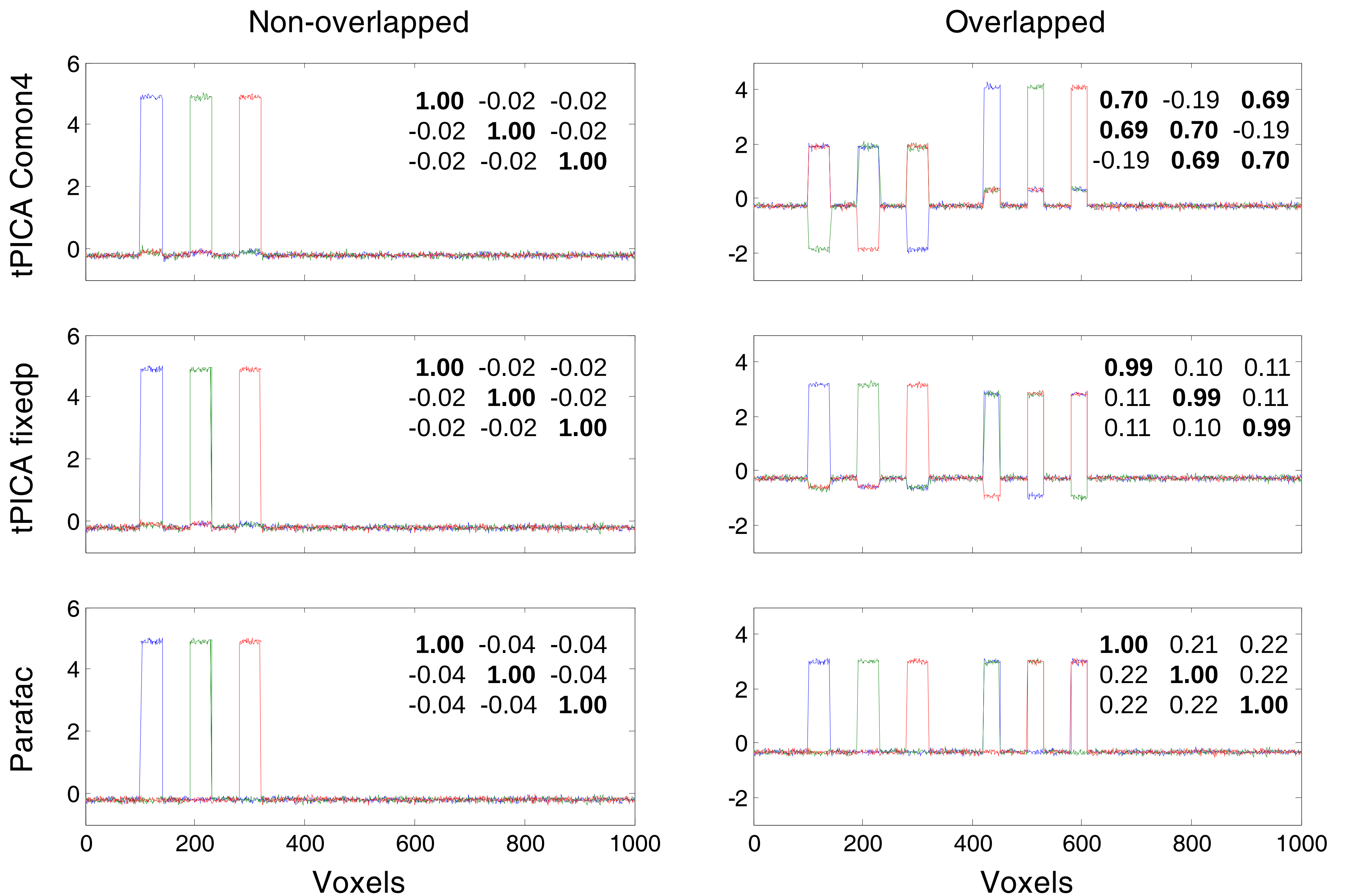
Three-mode recovery and fit



Three-mode recovery and fit



Typical recovered ICs from three-mode data (SNR = 2)



Three-mode results

- Three-mode ICA_Comon4, as in the two-mode case, successfully recovered only the non-overlapped sources

  For the overlapped sources, it persistently produced the same kind of biased components

- Parafac quite successfully recovered all types of sources if SNR is at least 0.3 (= expected fit ~8%; r = 0.94 in the voxel mode)

- Surprisingly, three-mode ICA_Fixedp recoverd all types of sources almost as successfully as Parafac

- Unlike the two-mode results, the fit by ICA_Comon4 was a little less for the overlapped and the Gaussian sources, particularly so with larger SNR

Volumic images too big to fit by Parafac

- When 3D images are obtained for whole brain in a typical resolution (2 or 4mm), a single image becomes too large for a Parafac fitting on any typical desktop computer (~222K or 56K intracranial voxels, respectively)

- Since the Parafac fitting optimizes a 2nd-order function (maximizing sum of squares or variance), the problem can be easily reduced to a much smaller size
  - 62 or 15 times reduction with no loss if $\mathbf{X}_{I \times JK}$ reduced to $JK \times JK$
  - and even more if willing to pay

Compressed Parafac and Parafac2

- Only the voxel mode is compressed
- Parafac: $\mathbf{X}_{I\times JK} = \mathbf{U}_{I\times D}\mathbf{S}\mathbf{V}' + \mathbf{E}, \quad D \le JK$

$$\mathbf{U}'\mathbf{X} = \tilde{\mathbf{A}}\left(\mathbf{C} \odot \mathbf{B}\right) + \tilde{\mathbf{E}}$$

- Parafac2 --- experimental conditions (e.g., visual stimuli) are often presented in different orders to cancel carry-over effects, and so the scan mode is not comparable across brains

$$\mathbf{X} = \left[\mathbf{X}_1, \ldots, \mathbf{X}_K\right] = \mathbf{U}_{I\times D}\mathbf{S}\mathbf{V}' + \mathbf{E}, \quad D \le \sum_{k=1}^{K} J_k$$

$$\mathbf{U}'\mathbf{X} = \tilde{\mathbf{A}}\left[\mathbf{D}_1\mathbf{B}_1', \ldots, \mathbf{D}_K\mathbf{B}_K'\right] + \tilde{\mathbf{E}}$$

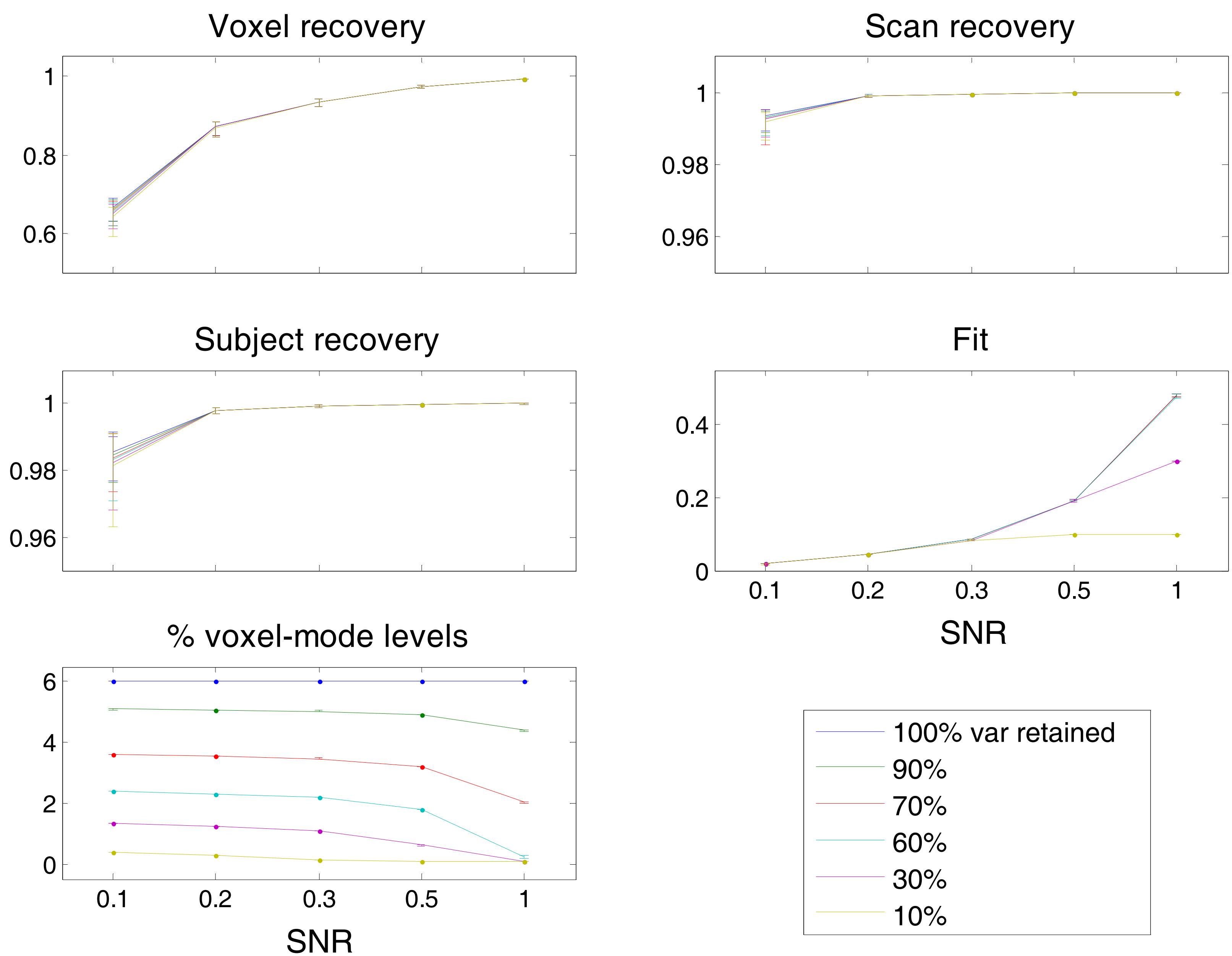
$$= \tilde{\mathbf{A}}\left[\mathbf{D}_1\mathbf{F}_{R\times R}'\mathbf{P}_1', \ldots, \mathbf{D}_K\mathbf{F}'\mathbf{P}_K'\right] + \tilde{\mathbf{E}}$$

$$\mathbf{D}_k = \mathrm{diag}\left(\mathbf{c}_k\right), \quad \mathbf{P}_k'\mathbf{P}_k = \mathbf{I}, \quad k = 1, \ldots, K$$

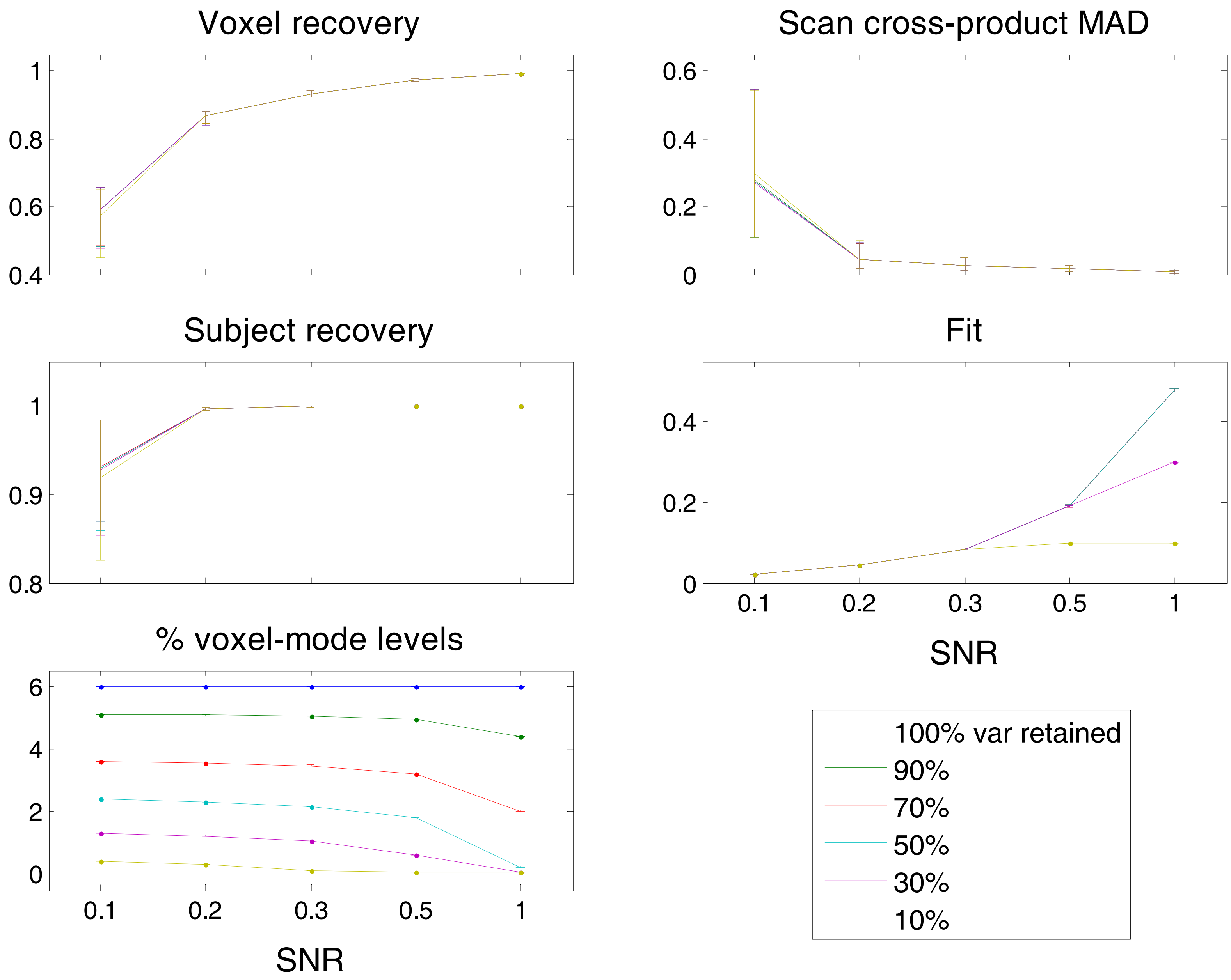
Compressed Parafac and Parafac2 analysis

- # voxels increased to 5000 so as to make I/JK similar to the 3-mode imaging data example
- SNR = 2 were dropped
- For Parafac2, $\mathrm{cor}(\mathbf{B}_k)$ varied $-0.58 \sim 0.51$
- Voxel mode compressed to retain data variance of 100, 90, 70, 50, 30 and 10%, with condition $D \geq R$

Compressed Parafac results



Compressed Parafac2 results



Conclusions and future work

- As expected, two-mode ICA successfully recovered the non-overlapped sources, but not the overlapped and the Gaussian sources
- Tensor PICA with fast ICA did recover all three kinds of sources while tensor PICA with Comon4, like the two-mode case, didn't recover the overlapped and the Gaussian sources --- the Newton update used in fast ICA did find the LS solution given the conditionally optimal rank-1 update for $\mathbf{B}$ and $\mathbf{C}$
- ICA persistently produced biased sources --- the true overlapped part was estimated as non-overlapped whereas the true non-overlapped part was estimated as overlapped
- When recovery was successful, SNR = 0.3 was sufficient
- Given the data size used (5000 x 20 x 15), compressing the voxel mode beyond JK didn't affect the accuracy

Conclusions and future work

- Parafac (and Parafac2) performed quite well even with ~92% errors (SNR = 0.3), but it won't be as good in practice since some irrelevant components are not as nice as Gaussian
- Though not included in the simulation, (two-mode) ICA is very robust against over-fitting, while ALS-Parafac is greatly affected
- Beckmann et al.'s voxel-wise normalization didn't work with the simulated data
- To better understand the performance of tPICA_Comon4 and tPICA_fixedp, further simulation needed with truly independent and dependent components
- Ready to sell Parafac and Parafac2 to brain imaging folks --- Richard's long-time wish about to be realized