Low-Rank Joint Subspace Construction for Multimodal Data Clustering

Aparajita Khan

March 14, 2019

Multimodal Data Clustering

Common set of n samples
M sets of observations

Multimodal Data Clustering

Common set of n samples
M sets of observations

$\mathbf{X_1} \in \mathbb{R}^{n \times d_1}$ $\mathbf{X_2} \in \mathbb{R}^{n \times d_2}$ ... $\mathbf{X_M} \in \mathbb{R}^{n \times d_M}$

$$n << d_m, \text{ for } m = 1, ..., M$$

Multimodal Data Clustering

Common set of n samples
M sets of observations

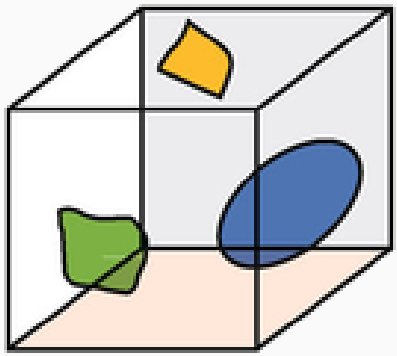
$\mathbf{X_1} \in \mathbb{R}^{n \times d_1}$ $\mathbf{X_2} \in \mathbb{R}^{n \times d_2}$ ... $\mathbf{X_M} \in \mathbb{R}^{n \times d_M}$

$$n << d_m, \text{ for } m = 1, ..., M$$

Cluster the n samples considering information from M different modalities/views

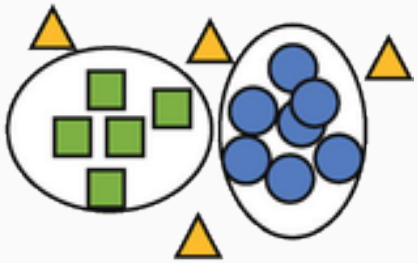
Multimodal Data Clustering

$\mathbf{X}_1 \in \mathbb{R}^{n \times d_1}$ $\mathbf{X}_2 \in \mathbb{R}^{n \times d_2}$ $\cdots$ $\mathbf{X}_M \in \mathbb{R}^{n \times d_M}$



Lower dimensional subspace

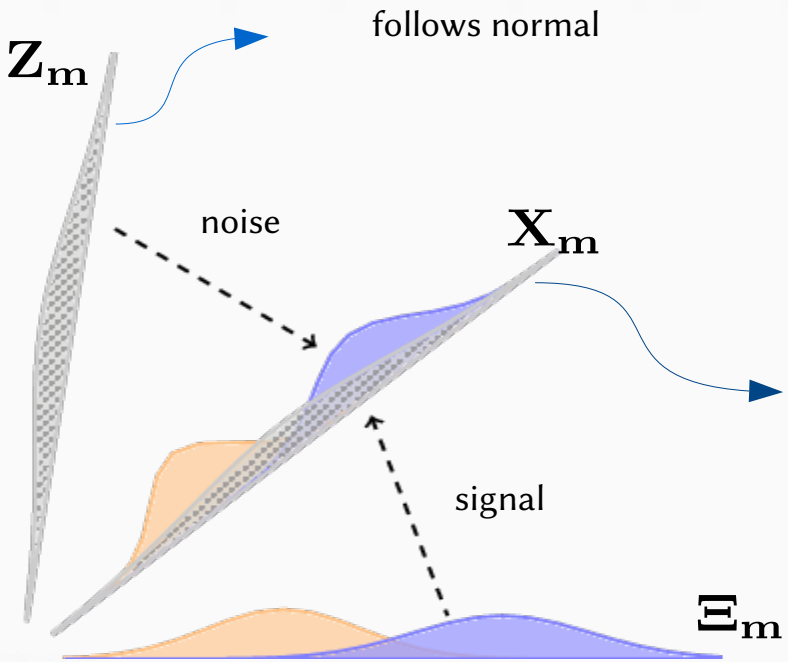
Clustering



Joint Subspace construction

Statistical hypothesis testing based Rank Estimation

$$\mathbf{X_m} = \mathbf{\Xi_m} + \mathbf{Z_m}, \;\; \text{signal } \mathbf{\Xi_m} \sim \sum_{\mathbf{j=1}}^{\mathbf{k}} \pi_\mathbf{j} \mathcal{N}(\mu_\mathbf{j}, \mathbf{\Sigma_j}), \;\; \text{noise } \mathbf{Z_m} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Lambda})$$



deviates from normality

$$\mathbf{X_m} = \mathbf{U_m \Sigma_m V_m^T}$$

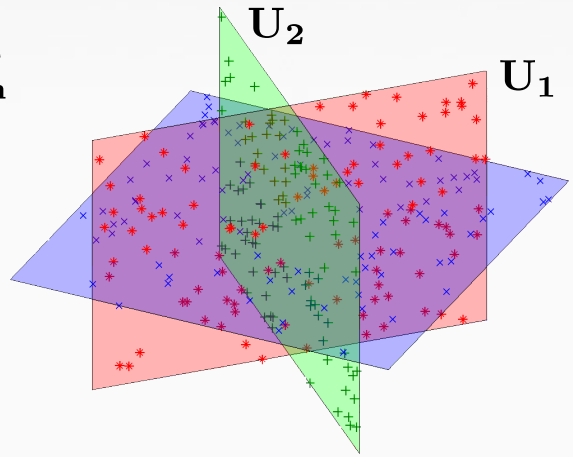
Principal Subspace of X_m

Joint Subspace construction

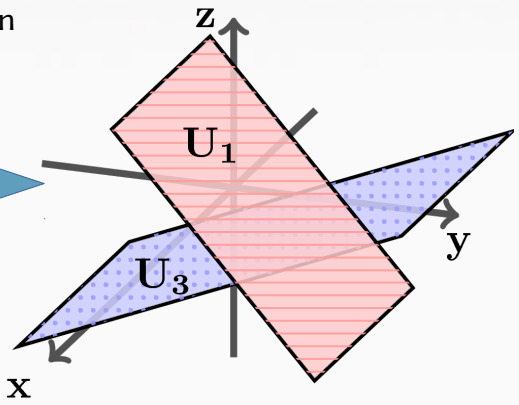
$$\text{subspaces} \quad U_1^r, U_2^r, \ldots, U_M^r$$

$$\mathbf{X_m = U_m \Sigma_m V_m^T}$$

Principal Subspace of X_m

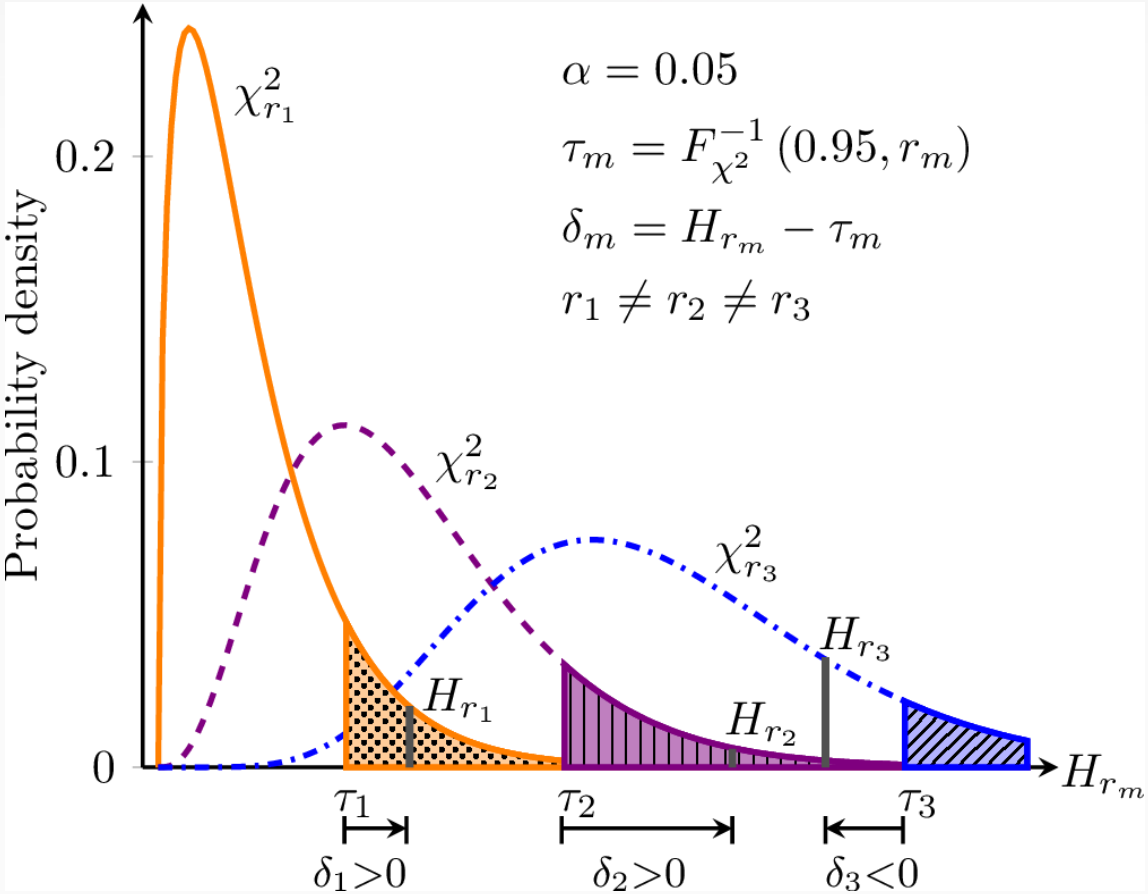


Select Only
Relevant Cluster Information
from Each Subspaces



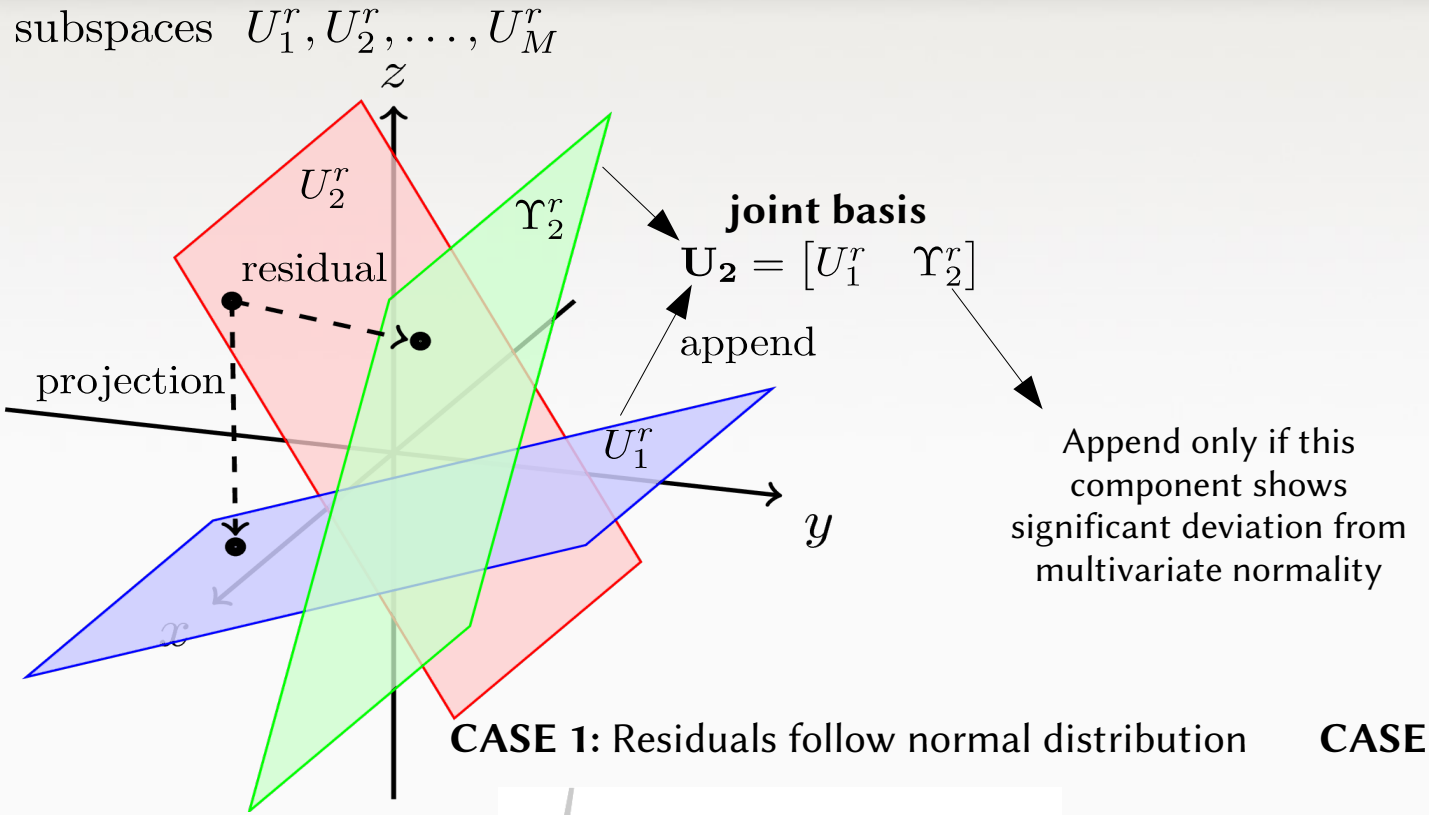
Multivariate Normality based Relevance Estimation

Relevance $$\mathcal{R}_l(\mathbf{X_m}) = \frac{1}{2}\left[1 + \frac{H_{r_m} - F_{\chi^2}^{-1}((1-\alpha), r_m)}{\max\{H_{r_m}, F_{\chi^2}^{-1}((1-\alpha), r_m)\}}\right]$$



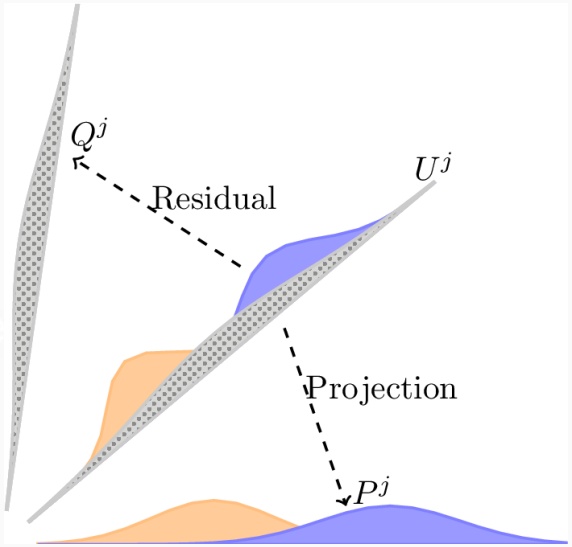
H_{r_m} Royston's H-statistic for r_m dimensional principal subspace of modality X_m

Joint Subspace construction

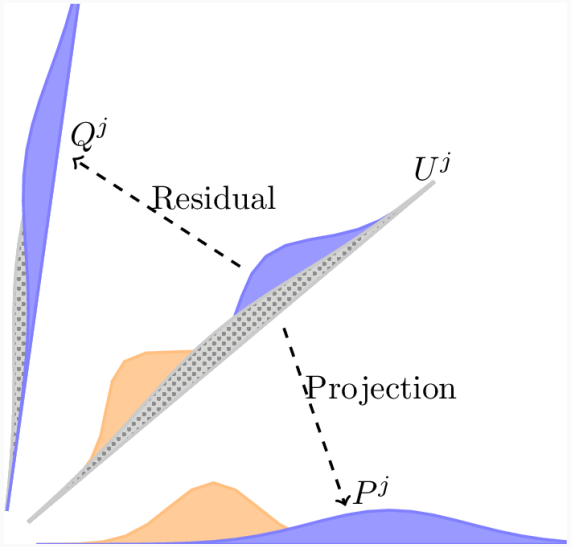


Append only if this component shows significant deviation from multivariate normality

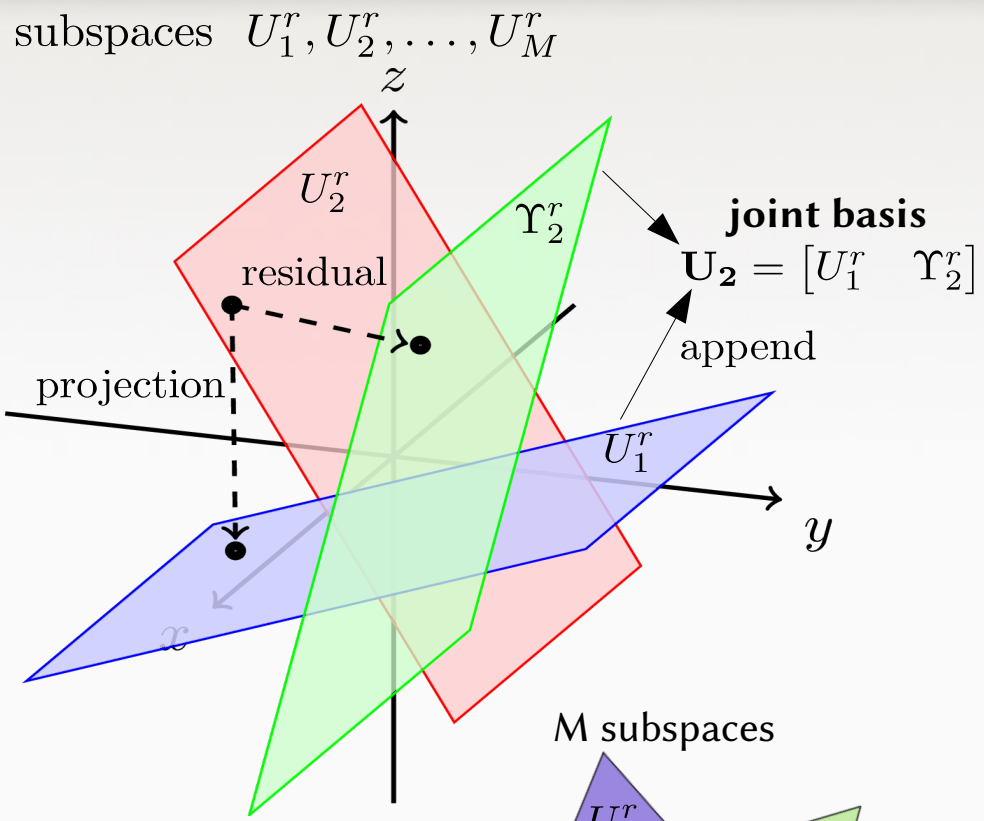
CASE 1: Residuals follow normal distribution

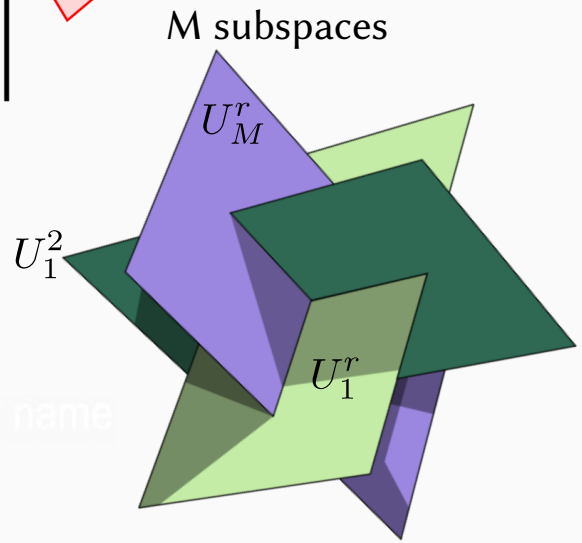


CASE 2: Residuals deviate from normality



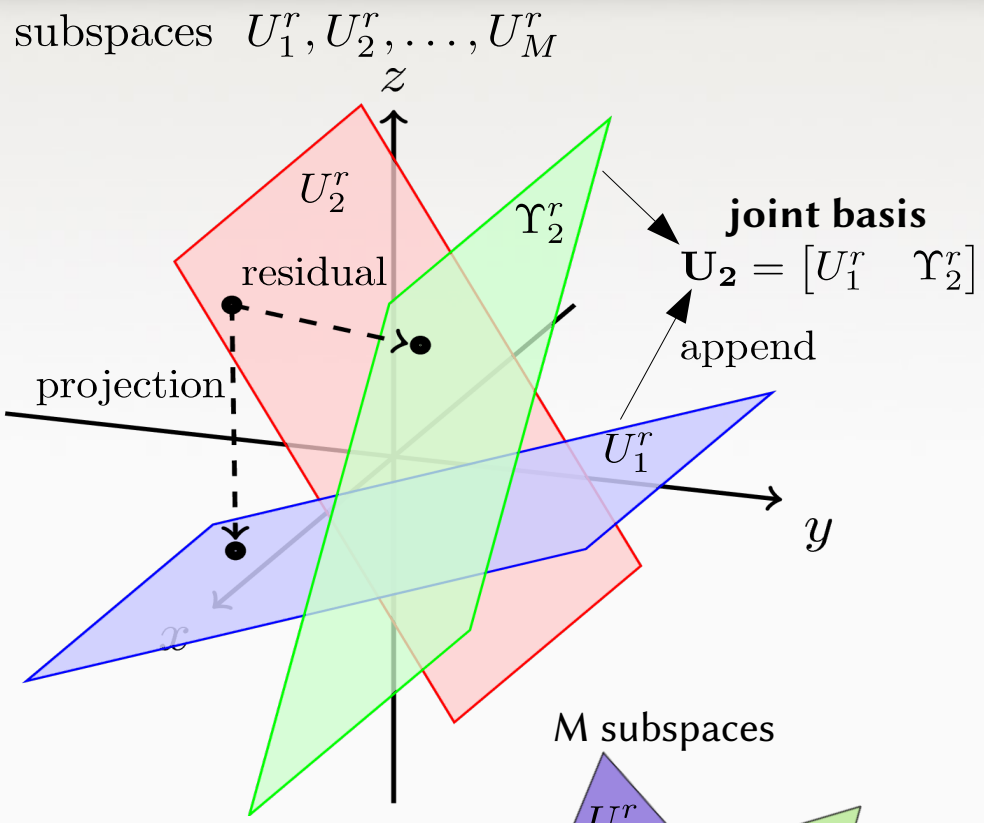
Joint Subspace construction

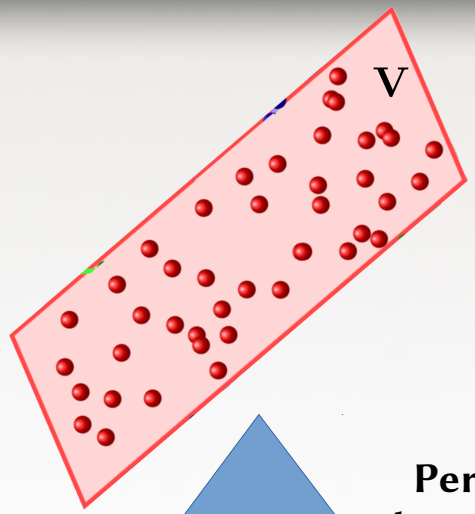




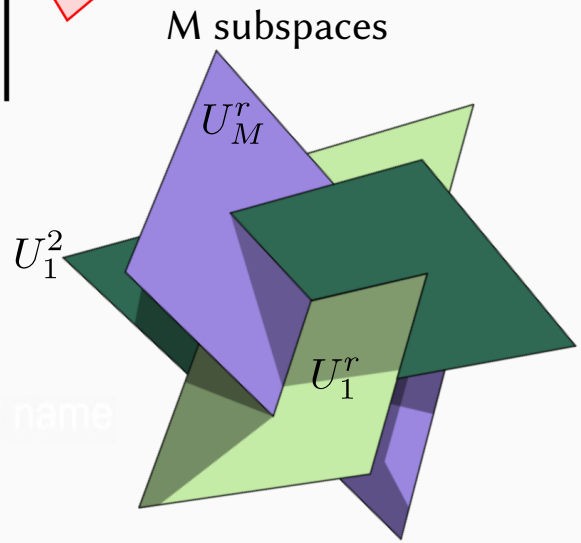
compute joint basis $\mathbf{U_M}$
from M subspaces

Joint Subspace construction





Perform clustering in Joint Subspace



compute joint basis $\mathbf{U_M}$ from M subspaces

Results

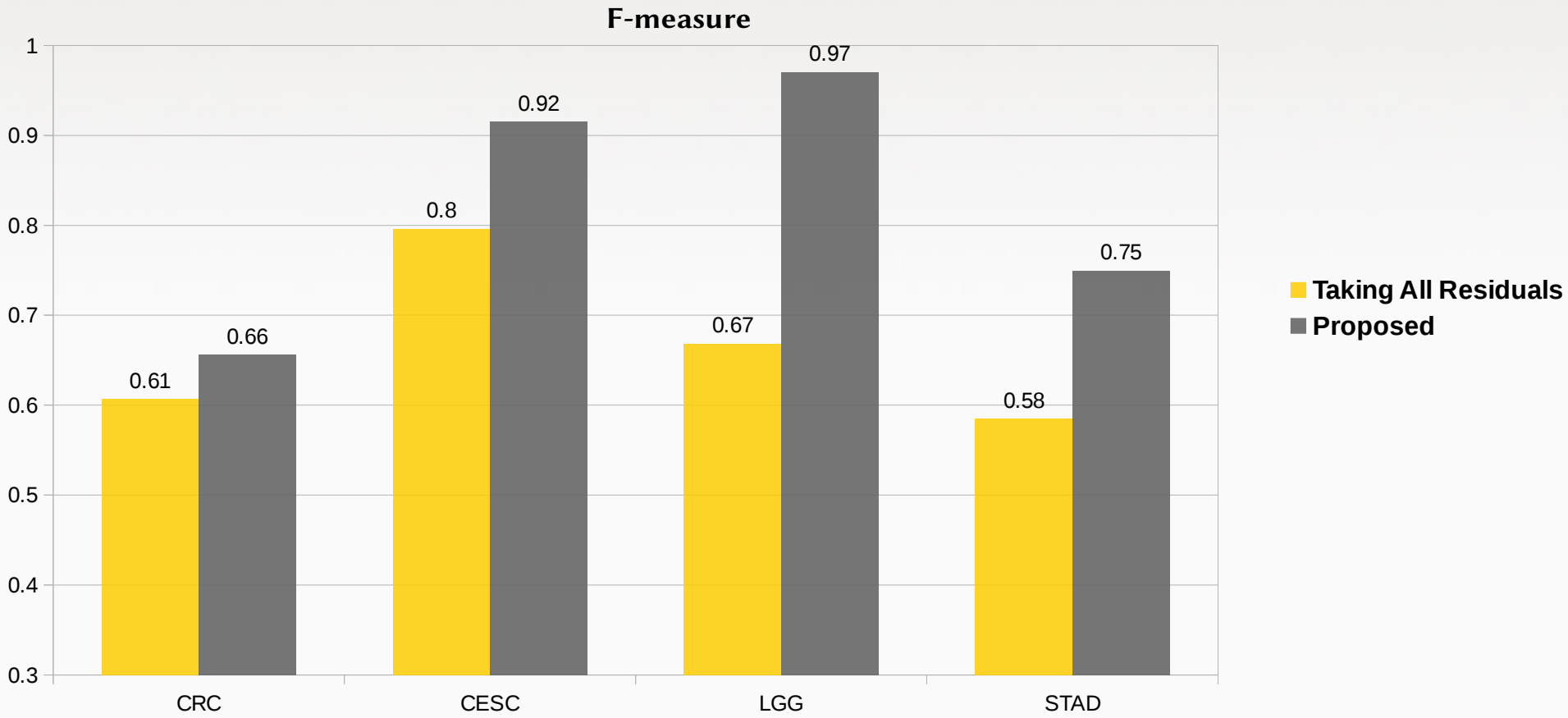
Data sets:

Colorectal carcinoma (CRC)	[307 samples, 2 clusters]
Cervical carcinoma (CESC)	[124 samples, 3 clusters]
Lower grade glioma (LGG)	[267 samples, 3 clusters]
Stomach adenocarcinoma (STAD)	[199 samples, 4 clusters]

Modalities: DNA Methylation
Gene Expression
Micro-RNA Expression
Protein Expression

Results

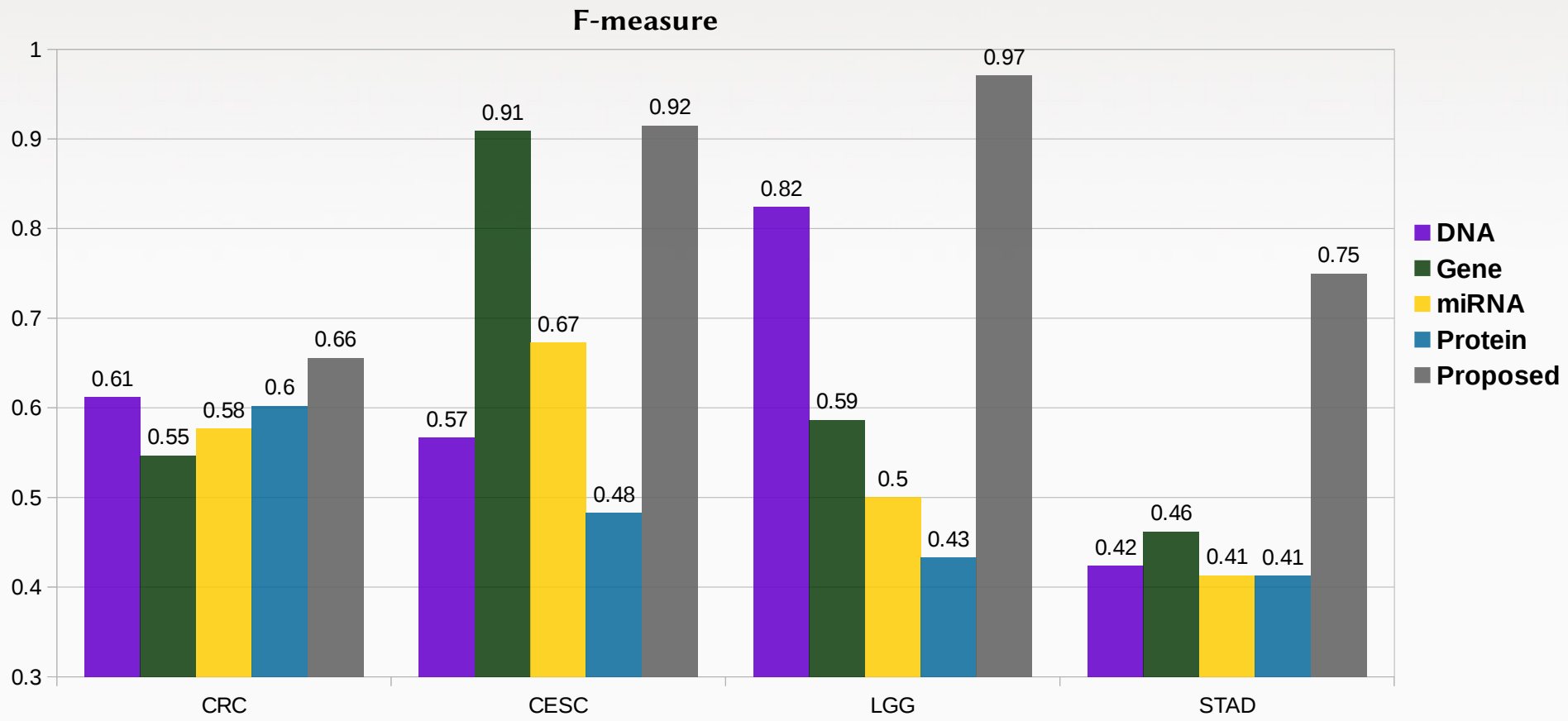
Importance of Noise-Free Approximation



- **Performance of selected proposed joint subspace is better than taking full-rank subspace from each modality**

Results

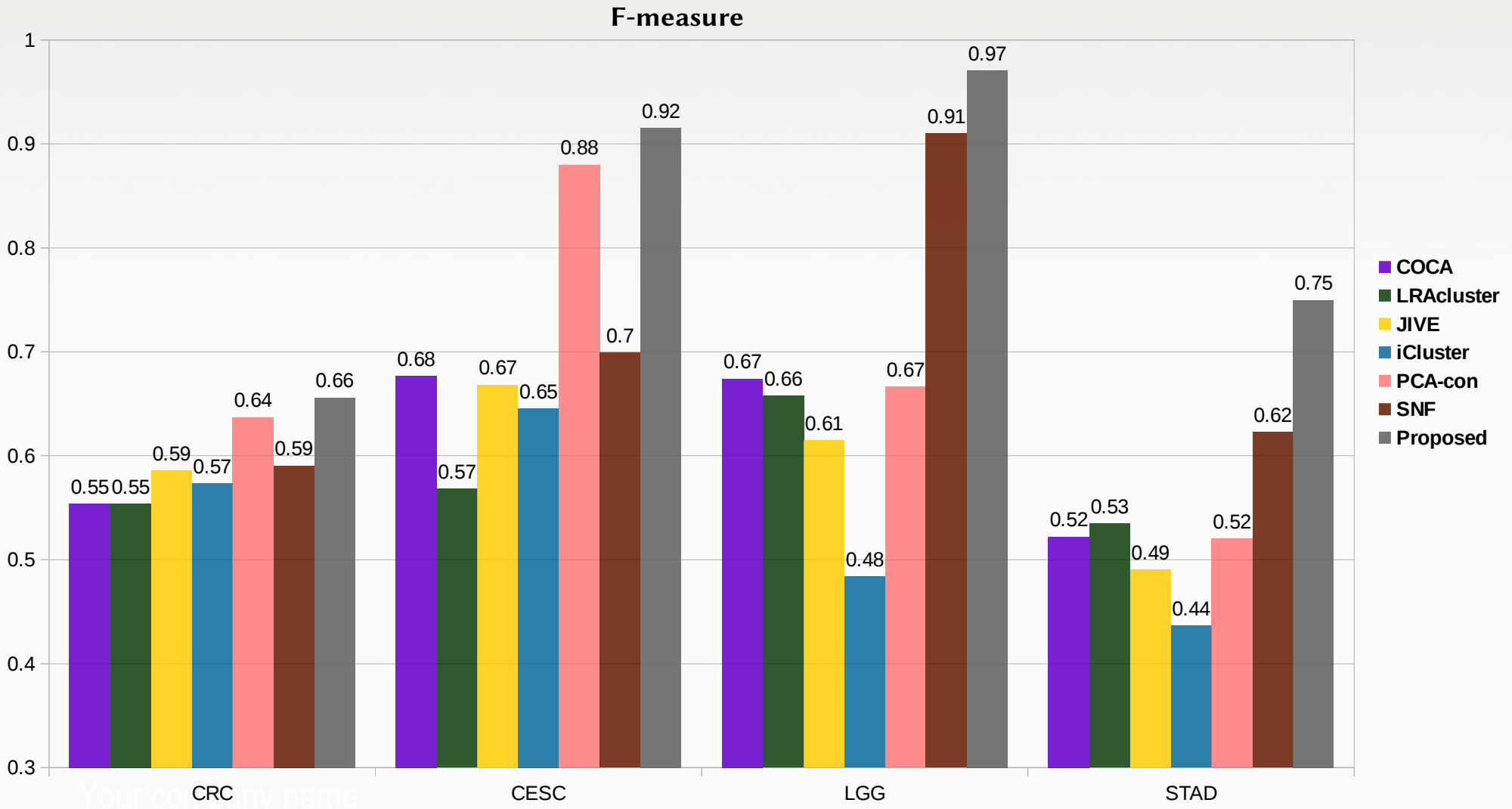
Importance of Integration



- **Proposed method performs better than all the individual modalities in all 4 data sets**
- **Very high performance for LGG and STAD data sets**

Results

Comparison with Existing Approaches



- **Performance of proposed better than existing in all data sets, especially for STAD data set**
- **PCA and SNF have next best performance**

Conclusion

- **Statistical hypothesis testing to estimate rank of subspaces**
- **Multivariate Normality based relevance of individual modalities**
- **Select only relevant modalities**
- **Select only complementary information between two modalities**

A. Khan and P. Maji, "Low-Rank Joint Subspace Construction for Cancer Subtype Discovery," IEEE/ACM Transactions on Computational Biology and Bioinformatics. (Accepted)

References

[1] F. R. Chung, Spectral graph theory. American Mathematical Society, 1997, no. 92.

[2] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in Advances in neural information processing systems, 2002, pp. 849–856.

[3] J. Shi and J. Malik, "Normalized cuts and image segmentation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 8, pp. 888–905, Aug 2000.

[4] G. W. Stewart and J.-g. Sun, Matrix perturbation theory. Academic press New York, 1990.

[5] S. Ji-guang, "Perturbation of angles between linear subspaces," Journal of Computational Mathematics, vol. 5, no. 1, pp. 58–61, 1987.

[6] B. Wang et al., "Similarity network fusion for aggregating data types on a genomic scale," Nature methods, vol. 11, no. 3, pp. 333–337, 2014.

[7]T. Xia, D. Tao, T. Mei, and Y. Zhang, "Multiview spectral embedding," IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol. 40, no. 6, pp. 1438–1446, Dec 2010.

[8] R. Shen, A. B. Olshen, and M. Ladanyi, "Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis," Bioinformatics, vol. 25, no. 22, pp. 2906–2912, 2009.

[9]D. Wu, D. Wang, M. Q. Zhang, and J. Gu, "Fast dimension reduction and integrative clustering of multi-omics data using low-rank approximation: application to cancer molecular classification," BMC genomics, vol. 16, no. 1, p. 1022, 2015.

[10] A. V. Knyazev and P. Zhu, "Principal angles between subspaces and their tangents," Mitsubishi Electric Research Laboratories, Tech. Rep. TR2012-058, 2012.

THANK YOU!