

Unsupervised Learning and Dimension Reduction

Part I: Nonnegative Matrix Factorization

Karthik Devarajan, Ph.D.

Department of Biostatistics & Bioinformatics

Fox Chase Cancer Center, Temple University Health System

Center for High-Dimensional Statistics, Department of Statistics

Lewis Katz School of Medicine, Temple University, Philadelphia

Data Science Institute, University of Delaware

June 28, 2025

Background

Nonnegative Matrix Factorization

Modeling complex noise and data structures

Application to Biomedical Signal Processing

Application to Large Scale Genomic Data Analytics

Some variations and generalizations of NMF

Introduction

- ▶ Modern scientific fields of study routinely generate large volumes of data requiring analysis and interpretation
- ▶ Examples include genomics, bioinformatics, chemometrics, imaging, neuroscience and natural language processing
- ▶ Massive data sets offer tremendous potential for growth in our understanding of the pathophysiology of many diseases

Examples of large scale biological data

- ▶ Neuroscience and biomedical imaging (EMG, EEG, fMRI, CT)
- ▶ High-throughput genomics
- ▶ Patient outcomes
- ▶ Natural language processing and information retrieval
- ▶ Fundamental to molecular pattern recognition, biomarker discovery, disease detection & treatment and survival prediction

We are drowning in information and starving for knowledge
Rutherford D. Roger

Key questions of interest include

- ▶ Can we isolate a sub-group of genomic features that are strongly associated with a specific phenotype?
- ▶ Are there discernible patterns in the data that can inform us about the structure of genetic networks?
- ▶ What genomic features have a prognostic impact on patient outcomes?
- ▶ Can we develop a genomic profile predictive of survival or recurrence?
- ▶ How do we better understand the activations of hundreds of muscles?
- ▶ How do we extract literature profiles using common semantic features to group documents relevant to a large set of genes (such as publications)?

Potential challenges

- ▶ High dimensionality
- ▶ Data measured on a variety of scales
- ▶ Data measured in time-frequency and spatial domains
- ▶ Signal-dependent and/or correlated noise structures
- ▶ Inherent sparsity in the data or the lack of it (depending on application)
- ▶ Unstructured data, as in natural language text
- ▶ Combination of two or more of the above

To call in a statistician after the experiment is done may be no more than asking him to perform a post-mortem examination; he may be able to say what the experiment died of.

Ronald A. Fisher (1890-1962)

Indian Statistical Congress, Sankhya, ca. 1938.

Methods for unsupervised learning include

- ▶ Principal Component Analysis
- ▶ Independent Component Analysis
- ▶ Nonnegative Matrix Factorization
- ▶ Vector Quantization
- ▶ Factor Analysis
- ▶ Exploratory Projection Pursuit
- ▶ Hierarchical Clustering
- ▶ K -means Clustering
- ▶ Consensus Clustering, Self-Organizing Maps etc.

Nonnegative Matrix Factorization

- ▶ Nonnegative Matrix Factorization (NMF) is an unsupervised, parts-based learning paradigm introduced by Lee & Seung (Nature, 1999)
- ▶ $V_{p \times n} \sim W_{p \times k} H_{k \times n}$, i.e., $V = WH + \epsilon$ where V, W, H are nonnegative
- ▶ Origins in image reconstruction and text mining
- ▶ Decomposition via a multiplicative updates algorithm

Original NMF formulation

- ▶ For two random variables with dfs F, G and densities f, g Kullback-Leibler (KL) divergence between f and g is

$$D(f||g) = \int_{\mathcal{R}} \left(\log \frac{f(x)}{g(x)} \right) dF(x)$$

- ▶ For two normal random variables with means μ_1 and μ_2 and (equal) variance σ

$$D(f||g) = (\mu_1 - \mu_2)^2$$

- ▶ For two Poisson random variables with means μ_1 and μ_2

$$D(f||g) = \mu_1 \log \left(\frac{\mu_1}{\mu_2} \right) - \mu_1 + \mu_2$$

Original NMF formulation

KL divergence between V and WH is given by

- ▶ Gaussian model (Euclidean distance)

$$D(V||WH) = \sum_{i,j} [V_{ij} - (WH)_{ij}]^2$$

- ▶ Poisson model

$$D(V||WH) = \sum_{i,j} \left\{ V_{ij} \log \left(\frac{V_{ij}}{(WH)_{ij}} \right) - V_{ij} + (WH)_{ij} \right\}$$

Applications of NMF

- ▶ Large-scale biological data analytics
- ▶ Computer vision (facial recognition, video, color research)
- ▶ Information retrieval
- ▶ Speech recognition

Some biomedical applications include

- ▶ Biomedical signal processing (EMG, EEG, fMRI, CT etc.)
- ▶ Biomedical Informatics (text mining, document clustering)
- ▶ Molecular pattern discovery
- ▶ Functional characterization of genes
- ▶ Structure-based drug design
- ▶ For a review, see Devarajan (*PLoS Computational Biology*, 2008; *Matrix and Tensor Decompositions*, Springer, 2011)

Interpretation of decomposition

p genes/faces/muscles, n biometric signals, $k=2$

V matrix of gene expression/pixel intensity/EMG measurements

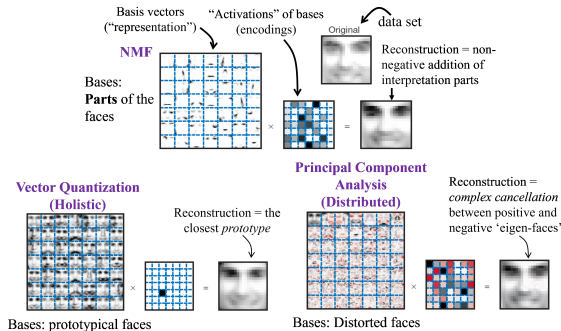
$$\begin{array}{cccc}
 & \mathbf{V} & & \mathbf{W} & & \mathbf{H} \\
 \mathbf{v(1,1)} & \mathbf{v(1,2)} & \dots & \mathbf{v(1,n)} & = & \mathbf{w(1,1)} & \mathbf{w(1,2)} & * & \mathbf{h(1,1)} & \mathbf{h(1,2)} & \dots & \mathbf{h(1,n)} \\
 \mathbf{v(2,1)} & \mathbf{v(2,2)} & \dots & \mathbf{v(2,n)} & & \mathbf{w(2,1)} & \mathbf{w(2,2)} & & \mathbf{h(2,1)} & \mathbf{h(2,2)} & \dots & \mathbf{h(2,n)} \\
 \cdot & & & & & \cdot & \cdot & & & & & \\
 \cdot & & & & & \cdot & \cdot & & & & & \\
 \cdot & & & & & \cdot & \cdot & & & & & \\
 \mathbf{v(p,1)} & \mathbf{v(p,2)} & \dots & \mathbf{v(p,n)} & & \mathbf{w(p,1)} & \mathbf{w(p,2)} & & & & &
 \end{array}$$

Interpretation of decomposition

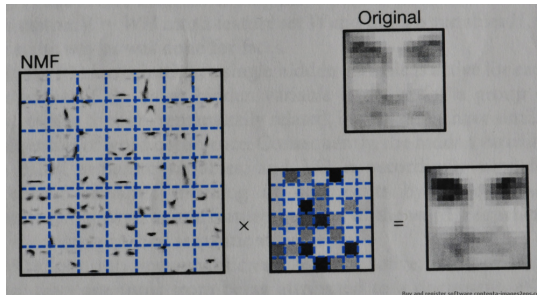
- ▶ Columns of W are basis vectors (metagenes, metafaces, muscle synergies)
- ▶ Rows of H are encodings (metagene/metaface/synergy expression patterns)
- ▶ Perception of whole as a linear combination of parts (represented by basis vectors and encodings)
- ▶ Basis vectors are overlapping and non-orthogonal

Interpretation of decomposition

NMF is a dimensionality reduction technique that learns **parts-based**, **local representation** of the data (Lee & Seung, *Nature*, 1999)



Parts-based, local representation



Parts-based, local representation

- The parts-based representation (“sparse” basis vectors and encodings) due to NMF **is a direct consequence of its non-negativity constraints.**
- Complex cancellation between mixed signs is avoided.

$$V_{i\mu} \approx (WH)_{i\mu} = \sum_{a=1}^r W_{ia} H_{a\mu}$$

$r < \text{number of rows in } V \text{ (data dimensionality)}$

data $\rightarrow V_{i\mu}$
 Bases as column vectors $\rightarrow W_{ia}$
 Activations of each basis as row vectors $\rightarrow H_{a\mu}$

Non-negativity constraints

$$V, W, H \geq 0$$

imposed on the bi-linear model

$$V = WH + \epsilon$$

Why are nonnegativity constraints significant?

- ▶ Data tend to be measured naturally on the nonnegative scale
- ▶ Non-negativity desirable and/or necessary in most applications
- ▶ Parts-based, local representation in NMF vs. holistic in VQ and distributed in PCA
- ▶ Nonnegative (additive) linear combination contrasts with superposition of orthogonal or statistically independent components (with mixed signs) in other methods

Why are nonnegativity constraints significant?

- ▶ Basis vectors tend to be inherently sparse
- ▶ Dependence among genomic/facial/muscle activation profiles can be captured
- ▶ Identifies hierarchy in data structure - does not force one like hierarchical clustering when none exists
- ▶ Facilitates context dependent clustering

NMF based on Euclidean Distance

The sum of squared errors between the original and reconstructed data, or the Euclidean distance, is **non-increasing across iterations** under the NMF update rules given below

- The algorithm is **guaranteed to converge**, if V , W , and $H \geq 0$
- Update rules result in optimal W and H that minimize the squared error

Problem Minimize $\|V - WH\|^2$ with respect to W and H , subject to the constraints $W, H \geq 0$.

Theorem The Euclidean distance $\|V - WH\|^2$ is nonincreasing under the update rules

$$H_{a\mu} \leftarrow H_{a\mu} \frac{(W^T V)_{a\mu}}{(W^T W H)_{a\mu}} \quad W_{ia} \leftarrow W_{ia} \frac{(V H^T)_{ia}}{(W H H^T)_{ia}}$$

The Euclidean distance is invariant under these updates if and only if W and H are at a stationary point of the distance.

Lee and Seung, *NIPS*, 2001

Expectation-Maximization (EM) Algorithm for NMF

- Proof of convergence based on the EM algorithm
- Requires finding an **auxiliary function** for the NMF cost function.

$G(h, h')$ is an auxiliary function for $F(h)$ if the conditions

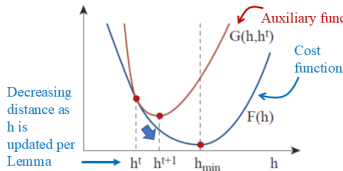
$$G(h, h') \geq F(h), \quad G(h, h) = F(h)$$

are satisfied.

Lemma If G is an auxiliary function, then F is nonincreasing under the update

$$h^{t+1} = \arg \min_h G(h, h^t)$$

Proof: $F(h^{t+1}) \leq G(h^{t+1}, h^t) \leq G(h^t, h^t) = F(h^t)$ ■



- NMF update rules can be derived when an appropriate auxiliary function, G , for the NMF cost function is identified, and when G is formulated according to the above lemma

The NMF Algorithm

- ▶ For a given run, start with random initial guesses for W and H
- ▶ Simultaneously update W and H via multiplicative rules, with most recent estimates of the matrices, until convergence
- ▶ Algorithm may not converge to the same solution for different runs (local optima)
- ▶ Decomposition (W and H) corresponding to minimum reconstruction error (or other similar criterion) across N runs can be used
- ▶ Usually $N = 50 - 100$ runs of the algorithm are sufficient
- ▶ Information across N runs can be exploited for unsupervised clustering (more on this later)

Why do we need complex models for biomedical data?

- ▶ Data measured on a variety of scales (continuous, count, ordered, binary)
- ▶ Data measured in time-frequency and spatial domains
- ▶ Signal-dependent or correlated noise structures (skewed, heavy-tailed)
- ▶ Inherently sparse data structure (structural zeros)
- ▶ Need for imposition of sparsity to aid interpretation
- ▶ Unstructured data, as in natural language text
- ▶ Combination of two or more of the above

All models are wrong, but some are useful
George Box

NMF algorithms for signal-dependent noise

- ▶ Heuristic algorithm for the exponential family (Cheung & Tresch, 2005)
- ▶ Heuristic algorithm based on β -divergence (Cichocki et al., 2006)
- ▶ MM and ME algorithms based on β -divergence (Févotte & Idier, 2011)
- ▶ β -divergence between V and WH is given by (for $\beta \in \mathbb{R} \setminus \{0, 1\}$)

$$D_{\beta}(V||WH) = \frac{1}{\beta(\beta-1)} \sum_{i,j} \left\{ V_{ij}^{\beta} - \beta V_{ij}(WH)_{ij}^{\beta-1} + (\beta-1)(WH)_{ij}^{\beta} \right\}.$$

- ▶ Includes the Gaussian ($\beta = 2$), Poisson ($\beta \rightarrow 1$), gamma ($\beta \rightarrow 0$) and inverse Gaussian ($\beta = -1$) models as special cases.

NMF algorithms for signal-dependent noise

- ▶ Information-theoretic approach based on alternate likelihoods
- ▶ Reinforcement learning objective
- ▶ Based on dual KL divergence for gamma and inverse Gaussian models
- ▶ Based on symmetric J divergence for the gamma model
- ▶ Devarajan & Cheung, *Neural Computation*, 2014

NMF algorithms for heteroscedastic and correlated data

- ▶ Families of algorithms based on generalized {linear models, dual divergence and intrinsic information}
- ▶ Extends to exponential family of models
- ▶ Useful for modeling a variety of noise structures (including temporal)
- ▶ Incorporates non-linear effects via link functions
- ▶ Applicable to biomedical signal processing, genomics and text mining
- ▶ *Neural Networks*, 2021; *Neural Computation*, 2014, 2016; *Machine Learning*, 2015; *IEEE BIBM*, 2015; *IEEE EMBS*, 2015; *AJMMS*, 2008

Algorithms for signal-dependent noise

Algorithm	Divergence	Reference
EM	KL [†]	Lee & Seung (2001)
Heuristic	Likelihood*	Cheung & Tresch (2005)
Heuristic	β^*	Cichocki et al. (2006)
MM, ME	β^*	Févotte & Idier (2011)
EM	Renyi [†]	Cichocki et al. (2008)
		Devarajan & Ebrahimi (2015)
EM	Quasi-likelihood ^{*,a}	Devarajan & Cheung (2016)

† Poisson model * includes exponential family models

a incorporates link functions

Alternative approaches

Algorithm	Divergence	Reference
EM	Bregman [*]	Dhillon & Sra (2005)
EM	Dual KL ^{†,a}	Devarajan & Cheung (2014)
EM	$J^{\dagger}$	Devarajan & Cheung (2014)
EM	Dual KL ^b	Devarajan et al. (2015, 2021)
EM	Intrinsic Info. [*]	Devarajan et al. (2015)
EM	Dual KL [*]	Devarajan (2021)

* includes exponential family models † Gamma model
 a IG model b Poisson model

Generalized dual divergence of order β

- ▶ Generalized dual divergence between V and WH is given by (for $\beta \in \mathbb{R} \setminus \{0, 1\}$)

$$D_{\beta}(WH||V) = \frac{1}{\beta(\beta-1)} \sum_{i,j} \left\{ (WH)_{ij}^{\beta} - \beta(WH)_{ij} V_{ij}^{\beta-1} + (\beta-1)V_{ij}^{\beta} \right\}.$$

- ▶ Choice of β generalizes dual KL divergence for exponential family models.
- ▶ For example, $D_{\beta}(WH||V)$ includes Gaussian ($\beta = 2$), Poisson ($\beta \rightarrow 1$), gamma ($\beta \rightarrow 0$), IG ($\beta = -1$) etc. as special cases.
- ▶ Based on a reinforcement learning objective.

Maximum entropy reinforcement learning

- ▶ Dual KL divergence between F and G_θ can be expressed as

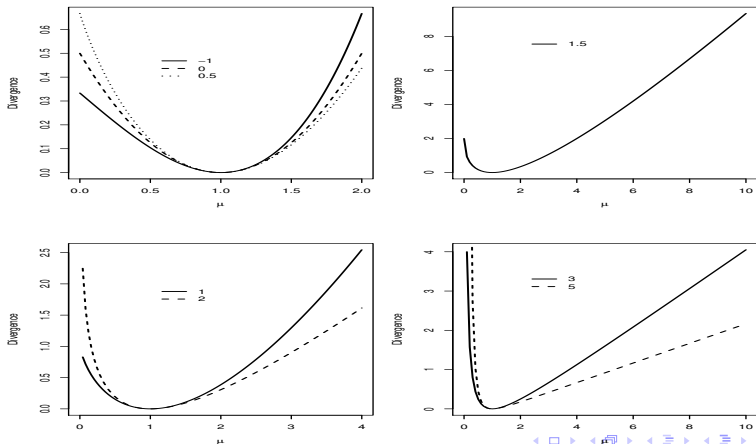
$$KL(g_\theta || f) = \int \left(\log \frac{g_\theta(x)}{f(x)} \right) dG_\theta(x) = -\mathbf{E}_{g_\theta}[\log f(x)] - H[g_\theta(x)].$$
- ▶ $H[g_\theta(x)]$ is the entropy corresponding to G_θ .
- ▶ Minimizing $K(g_\theta || f)$ with respect to the parameter θ gives

$$\begin{aligned} \arg \min_{\theta} KL(g_\theta || f) &= \arg \min_{\theta} \{-\mathbf{E}_{g_\theta}[\log f(x)] - H[g_\theta(x)]\} \\ &= \arg \max_{\theta} \{\mathbf{E}_{g_\theta}[\log f(x)] + H[g_\theta(x)]\}. \end{aligned}$$

- ▶ Also used in variational inference.

Generalized dual divergence of order β

Figure 1: Generalized Dual KL Divergence



Algorithm based on generalized dual divergence of order β

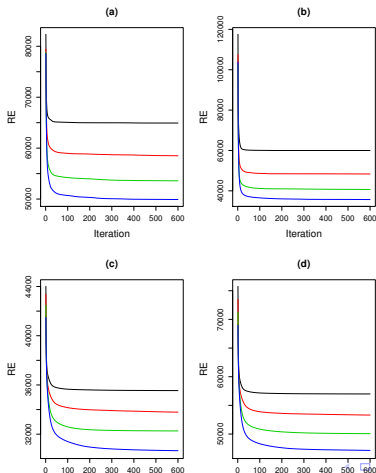
- ▶ Separate auxiliary functions $G(H_{aj}, H_{aj}^t) \geq F(H_{aj})$ exist for $1 < \beta < 2$, $\beta \in (1, 2)$ and $\beta \rightarrow 1$.
- ▶ $D_\beta(WH||V)$ is non-increasing under the update rules

$$H_{aj}^{t+1} = H_{aj}^t \left(\frac{\left(\sum_i \left(\frac{1}{\sum_b W_{ib} H_{bj}^t} \right)^{1-\beta} W_{ia} \right)^{1/(1-\beta)}}{\sum_i W_{ia} V_{ij}^{\beta-1}} \right),$$

$$W_{ia}^{t+1} = W_{ia}^t \left(\frac{\left(\sum_j \left(\frac{1}{\sum_b W_{ib}^t H_{bj}} \right)^{1-\beta} H_{aj} \right)^{1/(1-\beta)}}{\sum_j H_{aj} V_{ij}^{\beta-1}} \right).$$

- ▶ It is invariant under these updates if and only if W and H are at a stationary point of the divergence.

Convergence of algorithm



Gamma model: Algorithm based on dual KL divergence

- Dual KL divergence for the gamma model is given by

$$KL_G^d(V, WH) = \sum_{i,j} \left\{ \log \left(\frac{V_{ij}}{(WH)_{ij}} \right) + \frac{(WH)_{ij}}{V_{ij}} - 1 \right\}$$

- $KL_G^d(V, WH)$ is non-increasing under the multiplicative update rules for H (W)

$$H_{aj}^{t+1} = H_{aj}^t \left(\frac{\sum_i \left(\frac{1}{\sum_b W_{ib} H_{bj}^t} \right) W_{ia}}{\sum_i \left(\frac{W_{ia}}{V_{ij}} \right)} \right)$$

Gamma model: Algorithm based on J divergence

- ▶ J divergence for the gamma model is given by

$$J_G(V, WH) = \sum_{i,j} \left\{ \frac{(V_{ij} - (WH)_{ij})^2}{V_{ij}(WH)_{ij}} \right\}$$

- ▶ $J_G(V, WH)$ is non-increasing under the multiplicative update rules for H (W)

$$H_{aj}^{t+1} = H_{aj}^t \left(\frac{\sum_i \left(\frac{V_{ij}}{\sum_b W_{ib} H_{bj}^t} \right)^2 \left(\frac{W_{ia}}{V_{ij}} \right)}{\sum_i \left(\frac{W_{ia}}{V_{ij}} \right)} \right)^{1/2}$$

IG model: Algorithm based on dual KL divergence

- Dual KL divergence for the IG model is given by

$$KL_{IG}^d(V, WH) = \sum_{i,j} \frac{\{V_{ij} - (WH)_{ij}\}^2}{V_{ij}^2 (WH)_{ij}}$$

- $KL_{IG}^d(V, WH)$ is non-increasing under the multiplicative update rules for H (W)

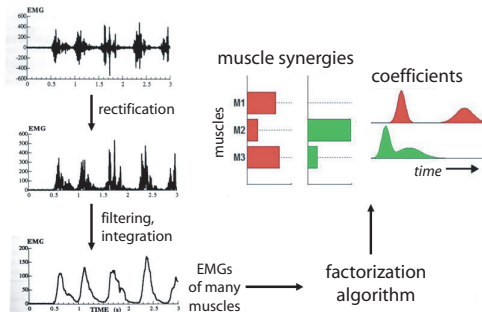
$$H_{aj}^{t+1} = H_{aj}^t \left(\frac{\sum_i \left(\frac{V_{ij}}{\sum_b W_{ib} H_{bj}^t} \right)^2 \left(\frac{W_{ia}}{V_{ij}^2} \right)}{\sum_i \left(\frac{W_{ia}}{V_{ij}^2} \right)} \right)^{1/2}$$

Application to Electromyography (EMG) Data

- ▶ Intramuscular EMG signals collected from all major hindlimb muscles (13) of the American bullfrog during natural behaviors (jump and swim)
- ▶ Natural locomotive behaviors constructed by motor primitives (muscle synergies) organized within the spinal cord
- ▶ Structures of these primitives centrally organized (i.e., independent of sensory feedback)
- ▶ EMGs recorded before and after the surgery of deafferentation (which removed sensory feedback)
- ▶ Four frogs each in jump or swim group before and after deafferentation
- ▶ Total of 16 data sets containing temporal EMGs for 13 muscles

EMG data decomposition

Identifying muscle synergies from EMGs using a factorization algorithm



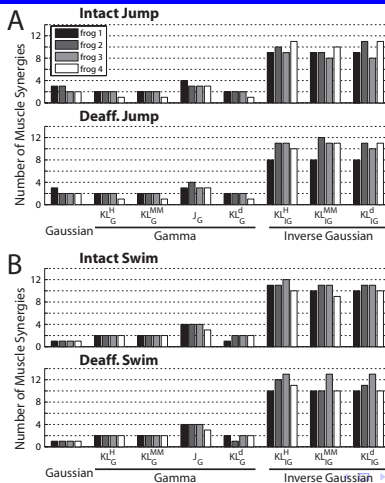
How do we isolate centrally organized muscle synergies?

- ▶ Goal - Discover structures that persist after deafferentation
- ▶ Signal-dependent nature of EMG data established in the literature (Harris & Wolpert, 1998, Cheung et al., 2005)
- ▶ Considered a variety of NMF algorithms based on Gaussian, gamma and inverse Gaussian models (for ranks $k = 1, \dots, 13$)
- ▶ Approach useful for identifying muscle synergies in stroke survivors
- ▶ How do we select the appropriate number of muscle synergies?

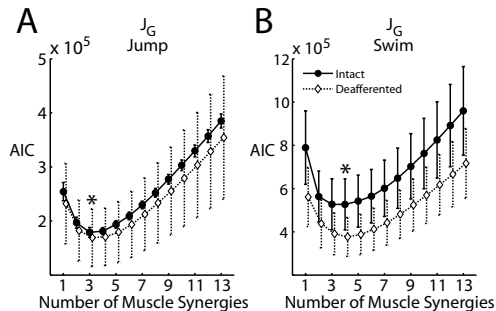
Model selection

- ▶ Information criteria (*IC*) such as Akaike (*AIC*), Bayesian (*BIC*), Deviance (*DIC*) etc. and proportion of explained variation R^2 .
- ▶ For a rank k factorization of $V_{p \times n}$, $AIC = -2 \log L + 2n_p$ where $L = \log\text{-likelihood}$ and $n_p = (p + n)k$
- ▶ We use
 - ▶ an *AIC*-type criterion based on deviance D_k (minimized value of cost function for rank k); in essence similar to *DIC*
 - ▶ the proportion of variation explained by the model for rank k , $R^2 = 1 - \frac{D_k}{D_0}$ where D_0 is the null deviance.
- ▶ Maisog et al. (2021) is a useful reference

Model selection for EMG data



Model selection for EMG data



Algorithm-specific proportion of explained variation R^2

Algorithm	R^2	Algorithm	R^2
Gaussian	$1 - \left\{ \frac{\sum_{i,j} (V_{ij} - (WH)_{ij})^2}{\sum_{i,j} (V_{ij} - \bar{V})^2} \right\}$	KL_G^d	$1 - \left\{ \frac{\widehat{KL}_G^d(V, WH)}{\sum_{i,j} \left\{ \log \left(\frac{V_{ij}}{\bar{V}} \right) + \frac{\bar{V}}{V_{ij}} - 1 \right\}} \right\}$
KL_G^H, KL_G^{MM}	$1 - \left\{ \frac{\widehat{KL}_G(V, WH)}{\sum_{i,j} \left\{ -\log \left(\frac{V_{ij}}{\bar{V}} \right) + \frac{V_{ij}}{\bar{V}} - 1 \right\}} \right\}$	J_G	$1 - \left\{ \frac{\widehat{J}_G(V, WH)}{\sum_{i,j} \left\{ \frac{(V_{ij} - \bar{V})^2}{V_{ij} \bar{V}} \right\}} \right\}$
KL_{IG}^H, KL_{IG}^{MM}	$1 - \left\{ \frac{\widehat{KL}_{IG}(V, WH)}{\sum_{i,j} \frac{\{V_{ij} - \bar{V}\}^2}{V_{ij} \bar{V}^2}} \right\}$	KL_{IG}^d	$1 - \left\{ \frac{\widehat{KL}_{IG}^d(V, WH)}{\sum_{i,j} \frac{\{V_{ij} - \bar{V}\}^2}{V_{ij}^2 \bar{V}}} \right\}$

R^2 achieved at ranks with minimum AIC

Algorithm	Behavior	Rank ($N = 4$; median $\pm$ SD)	R^2 ($N = 4$; mean $\pm$ SD)
Gaussian	Intact Swim	1 ± 0.00	56.78 ± 4.89
	Deaff. Swim	1 ± 0.00	56.58 ± 12.33
KL_G^H	Intact Swim	2 ± 0.00	76.66 ± 3.68
	Deaff. Swim	2 ± 0.00	80.00 ± 5.07
KL_G^{MM}	Intact Swim	2 ± 0.00	76.66 ± 3.69
	Deaff. Swim	2 ± 0.00	80.00 ± 5.07
KL_G^d	Intact Swim	2 ± 0.50	92.81 ± 2.80
	Deaff. Swim	2 ± 0.50	94.78 ± 2.70
J_G	Intact Swim	4 ± 0.50	93.03 ± 1.28
	Deaff. Swim	4 ± 0.50	95.18 ± 1.05
KL_{IG}^H	Intact Swim	11 ± 0.82	99.25 ± 0.34
	Deaff. Swim	12 ± 1.29	99.61 ± 0.30
KL_{IG}^{MM}	Intact Swim	11 ± 0.96	98.50 ± 0.46
	Deaff. Swim	10 ± 1.50	98.95 ± 0.43
KL_{IG}^d	Intact Swim	11 ± 0.58	99.92 ± 0.03
	Deaff. Swim	11 ± 1.41	99.97 ± 0.02

Fraction of variation explained†, R^2

Algorithm	Intact Jump $r = 3$	Deaff. Jump $r = 3$	Intact Swim $r = 4$	Deaff. Swim $r = 4$
Gaussian	87.69 $\pm$ 1.21	87.38 $\pm$ 1.04	85.43 $\pm$ 1.83	87.89 $\pm$ 4.51
KL_G^H, KL_G^{MM}	91.97 $\pm$ 1.48	90.58 $\pm$ 1.17	87.17 $\pm$ 1.13	90.11 $\pm$ 2.47
KL_G^d	98.93 $\pm$ 0.35	98.74 $\pm$ 0.11	96.46 $\pm$ 0.41	97.62 $\pm$ 0.77
J_G	97.85 $\pm$ 0.64	97.41 $\pm$ 0.29	93.56 $\pm$ 0.55	95.50 $\pm$ 1.38
KL_{IG}^H	90.97 $\pm$ 2.06	88.72 $\pm$ 0.58	86.67 $\pm$ 0.87	89.49 $\pm$ 2.30
KL_{IG}^{MM}	90.89 $\pm$ 2.07	88.68 $\pm$ 0.52	86.50 $\pm$ 0.83	89.43 $\pm$ 2.25
KL_{IG}^d	99.79 $\pm$ 0.11	99.77 $\pm$ 0.05	98.94 $\pm$ 0.30	99.37 $\pm$ 0.22

†physiologically interpretable # of muscle synergies (r) determined using kinematic analysis (d'Avella et al., 2003; d'Avella & Bizzi, 2005)

Comparison of algorithms

- ▶ Remarkable consistency in the number of synergies determined (k) across the 4 animals
- ▶ J_G showed best overall performance based on AIC in choosing the number of synergies k
- ▶ J_G outperformed the gamma, IG and Gaussian models for both jump and swim
- ▶ Gamma-based algorithms outperformed the Gaussian model
- ▶ $KL(WH||V)$ and $J(V||WH)$ for gamma showed the best overall performance

Comparison of algorithms

- ▶ Synergies from J_G could be functionally interpreted
- ▶ Synergies were activated during pre-landing hip flexion, post-landing-knee flexion and extension phases of the jump
- ▶ J_G required 3 & 4 synergies, respectively, (for both intact and deaff.) jumps and swim
- ▶ Number of synergies found using J_G validated by numbers determined from prior studies

Looking beyond neuroscientific applications

The best thing about being a statistician is that you get to play in everyone's backyard

John W. Tukey

Application to gene expression: Leukemia data

- ▶ Tissue samples from acute myelogenous leukemia (AML) and acute lymphoblastic leukemia (ALL) (Golub et al., Science, 1999)
- ▶ Most widely used publically available benchmark data set in genomics
- ▶ $n = 38$ bone marrow samples hybridized on Affy Hu6800 chips;
 $p = 5000$ gene expression profiles
- ▶ 27 ALL and 11 AML samples
- ▶ 27 ALL: 19 ALL-B type and 8 ALL-T type
- ▶ Investigated clustering of samples from ALL and AML as well as the division of ALL into T and B subtypes

Interpretation of decomposition

p genes/faces/muscles, n biometric signals, $k=2$

V matrix of gene expression/pixel intensity/EMG measurements

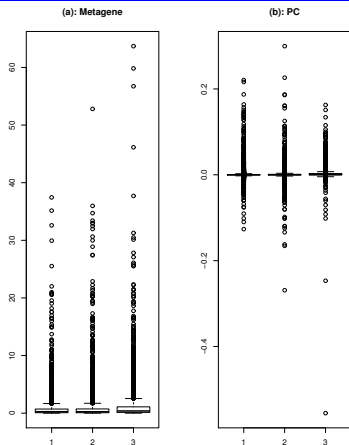
$$\begin{array}{cccc}
 & \mathbf{V} & & \mathbf{W} & & \mathbf{H} \\
 \mathbf{v(1,1)} & \mathbf{v(1,2)} & \dots & \mathbf{v(1,n)} & = & \mathbf{w(1,1)} & \mathbf{w(1,2)} & * & \mathbf{h(1,1)} & \mathbf{h(1,2)} & \dots & \mathbf{h(1,n)} \\
 \mathbf{v(2,1)} & \mathbf{v(2,2)} & \dots & \mathbf{v(2,n)} & & \mathbf{w(2,1)} & \mathbf{w(2,2)} & & \mathbf{h(2,1)} & \mathbf{h(2,2)} & \dots & \mathbf{h(2,n)} \\
 \cdot & & & & & \cdot & \cdot & & & & & \\
 \cdot & & & & & \cdot & \cdot & & & & & \\
 \cdot & & & & & \cdot & \cdot & & & & & \\
 \mathbf{v(p,1)} & \mathbf{v(p,2)} & \dots & \mathbf{v(p,n)} & & \mathbf{w(p,1)} & \mathbf{w(p,2)} & & & & &
 \end{array}$$

Metagenes and metagene expression profiles

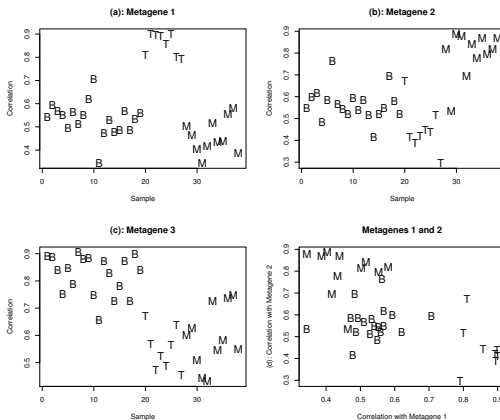
- ▶ $k = 2$
- ▶ Columns of W (w_1, w_2) are metagenes
- ▶ Rows of H (h_1, h_2) are metagene expression profiles
- ▶ Genes with highest metagene coefficients are most interesting
- ▶

<i>Coefficient</i>	$w_1(\%)$	$w_2(\%)$
< 1	73	84
< 3	95	91
> 10	1 (53)	1.5 (77)

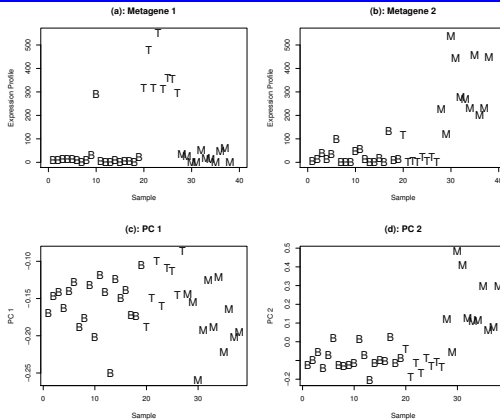
Metagenes and metagene expression profiles



Metagenes and metagene expression profiles



Metagenes and metagene expression profiles



Consensus clustering

- ▶ NMF may not converge to the same solution on each run
- ▶ Consensus clustering exploits this to evaluate consistency of the NMF algorithm across N runs
- ▶ Groups items into k clusters, where k is the pre-specified rank of the factorization
- ▶ Provides quantitative assessment of homogeneity of clustering
- ▶ Helps assess whether a given rank k provides a biologically meaningful decomposition of the data

Connectivity matrix

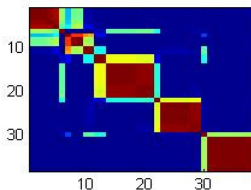
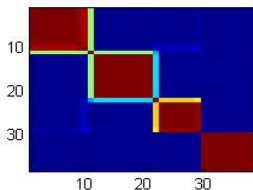
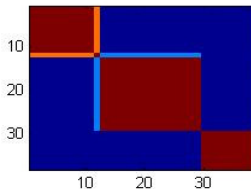
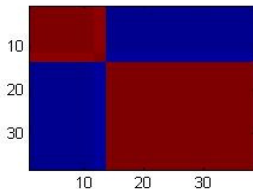
Connectivity Matrix

Sample	1	2	3		40
1	1	0	1		1
2	0	1	1		
3	1	0	1		
.	.				
.	.				
.	.				
.	.				
.	.				
40	1	0	0		1

Consensus clustering

- ▶ For each run, a binary connectivity matrix $C_{n \times n}$ is obtained where $C_{ij} = 1$ if items i and j cluster together and 0 otherwise
- ▶ Consensus matrix $\bar{C} = \text{mean connectivity matrix over } N = 50 - 100 \text{ runs}$
- ▶ $\bar{C}_{ij} \approx \text{probability that items } i \text{ and } j \text{ appear in the same cluster}$
- ▶ Apply hierarchical clustering (HC) to $\bar{C}$ to assign groups to each item and visualize clusters using re-ordered $\bar{C}$

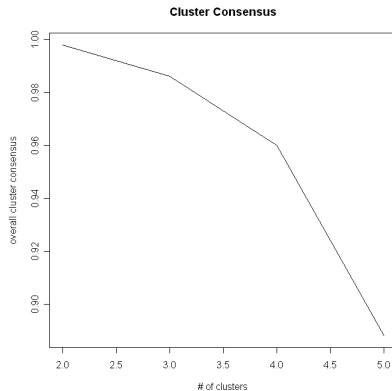
Heat map of consensus matrix (2-5 clusters)



Clustering accuracy

- ▶ Cophenetic correlation coefficient (ρ): Pearson correlation between $1 - \bar{C}$ and distance induced by HC using average linkage
- ▶ Overall cluster consensus $\kappa = \frac{\sum_{l=1}^k \sum_{i,j \in I_l, i < j} \bar{C}_{ij}}{\sum_{l=1}^k n_l(n_l-1)/2}$
- ▶ Other measures include normalized mutual information (NMI), adjusted Rand Index (ARI) etc.
- ▶ $0 < \rho, \kappa, NMI, ARI < 1$
- ▶ Plot ρ, κ, NMI or ARI vs. number of clusters (k)

κ vs. k



Some variations of NMF include

- ▶ Non-negative Sparse PCA (Zass & Sashua, 2006) can handle negative data but retains sparsity, PCs restricted to be non-negative and are disjoint, sparsity controllable
- ▶ NMF with orthogonality constraints (Li et al., 2008)
- ▶ NMF with sparsity constraints (wealth of literature)
- ▶ A generalized linear models approach to NMF via quasi-likelihood (Devarajan & Cheung, 2016)

Control of sparsity

- Sparsity in W (columns) and/or H (rows) can be incorporated explicitly via penalization

$$D(V||WH) = \sum_{i,j} \left\{ V_{ij} \log \left(\frac{V_{ij}}{(WH)_{ij}} \right) - V_{ij} + (WH)_{ij} \right\} + \lambda \sum_{a,j} H_{aj}$$

$$D(V||WH) = \sum_{i,j} [V_{ij} - (WH)_{ij}]^2 + \lambda \sum_{a,j} H_{aj}$$

- λ chosen via cross validation

Control of sparsity

► $D_\beta(V||WH) + \lambda_1\psi_1(H) + \lambda_2\psi_2(W), \quad \lambda_1, \lambda_2 > 0$

$$\psi_l(.) = \begin{cases} \sum_{a,j} H_{aj}, & (L_1 \text{ or } \textit{lasso} \text{ penalty}), l = 1, 2 \\ \sum_{i,a} W_{ia}^2, & (L_2 \text{ penalty}) \end{cases}$$

- λ_1, λ_2 chosen via cross validation
- Devarajan (*Neural Networks*, 2021)

Relationship to Generalized Linear Models (GLM)

- ▶ In the classical linear model $Y = X\beta + \epsilon$, $E(Y) = \mu = X\beta$
- ▶ For GLMs, $\eta = g(\mu) = X\beta$ is the linear predictor and $g(\cdot)$ is the link function
- ▶ Consider each column of V followed by each row and alternate between columns and rows
- ▶ Minimization of NMF cost function can be viewed as alternating non-negativity constrained GLM
- ▶ Potential connections to deep learning framework?

Relationship to GLM

- ▶ Lee & Seung (*Nature*, 1999; *NIPS*, 2001)
 - ▶ Non-negativity constrained alternating least squares (Gaussian GLM with identity link, Euclidean distance)
 - ▶ Non-negativity constrained alternating Poisson regression (Poisson GLM with identity link)
- ▶ Generalized to other members of the exponential family using quasi-likelihood (QL) (Devarajan & Cheung, *Neural Computation*, 2016)
- ▶ QL enables modeling of non-linear relationships and temporal structure via link functions (and further extensions using generalized estimating equations)
- ▶ Useful for modeling data with skewed distributions and signal-dependent noise (mean-variance relationship)

Summary of NMF

- ▶ Parts-based, local representation of data
- ▶ Sparse representation facilitates interpretation (although not guaranteed)
- ▶ Extracted components are non-orthogonal and overlapping
- ▶ Non-negativity of V , W & H correspond naturally to firing rates or muscle activities
- ▶ Guaranteed convergence but only to local optima
- ▶ Several methods exist for determining the optimal number of basis vectors
- ▶ Non-negativity constraints may not facilitate modeling of inhibitory drives
- ▶ Algorithms are available for modeling signal-dependent noise
- ▶ Implementation is easy and available in Matlab and R