Non-negative Matrix Factorization (NMF): Statistical learning of the mutational processes in human cancer

Asger Hobolth, October 21, 2018

Today:

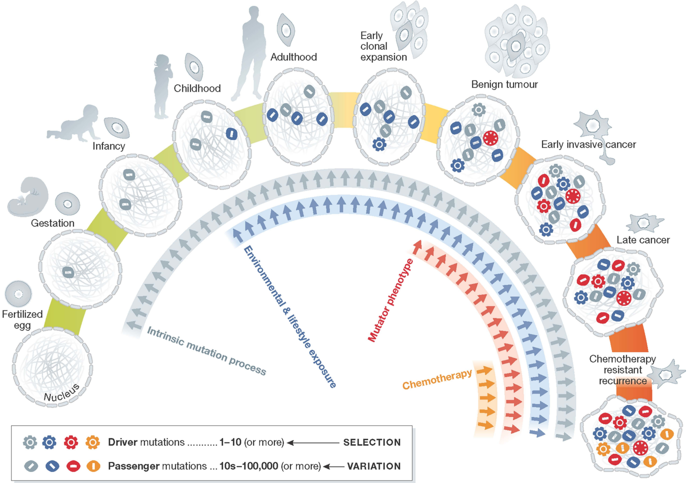
a. Background: Cancer genetics.
b. Signature Model: Non-negative Matrix Factorization (NMF).
c. Estimation: High-dimensional optimization problem.
d. Alternating non-negative least sequare (NLS).
e. Solutions to the NLS problem.
f. Test problem and comparison of algorithms.

Next few weeks:

Detailed solutions (algorithms) to the NLS problem.
Applications, Results and Conclusions.
Discussion, Critique and Perspectives.

a. Background: Cancer genetics

Cancer arises as an accumulation of driver mutations



Stratton (2013, Figure 1)

Sequencing Technology

The 1000$ genome is now available.

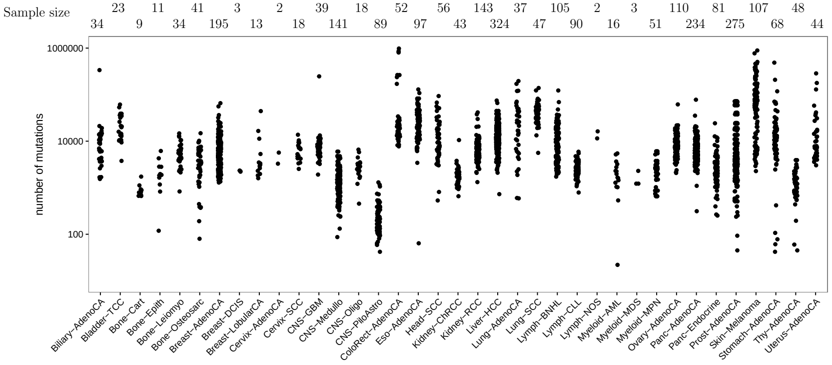
We sequence healthy tissue and the tumour, and identify the type of mutations along the genomes:

```
Healthy AAGCT...TCG...etc
 Tumour AAGCT...TAG...etc
```

We write `T[C>A]G`.

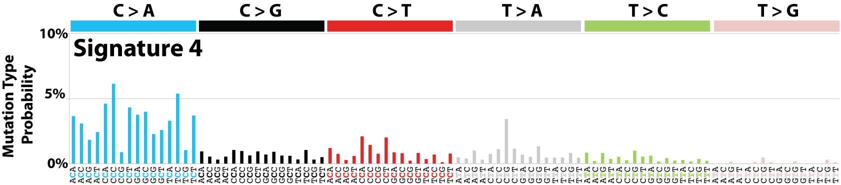
There are $4 \cdot [4 \cdot 3] \cdot 4 = 192$ mutation types.
Assuming strand-symmetry we get $192/2 = 96$ mutation types.

International effort to sequence 2500 tumours



Biological processes operating in human cancer I

Sequencing of lung cancer

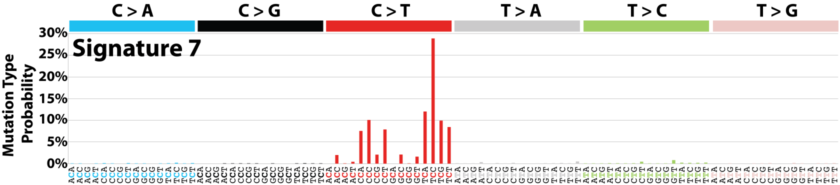


'Signature 4 is associated with smoking and its profile is similar to the mutational pattern observed in experimental systems exposed to tobacco carcinogens.'

`https://cancer.sanger.ac.uk/cosmic/signatures`
(COSMIC abbreviates 'catalogue of somatic mutations in cancer')
Alexandrov et al (2016): Signatures associated with tobacco smoking.

Biological processes operating in human cancer II

Sequencing of skin cancer



'Based on its prevalence in ultraviolet exposed areas and the similarity of the mutational pattern to that observed in experimental systems exposed to ultraviolet light Signature 7 is likely due to ultraviolet light exposure.'

`https://cancer.sanger.ac.uk/cosmic/signatures`

b. Signature Model: Non-negative Matrix Factorization (NMF).

Data: Count matrix $V \in \mathbb{R}^{M \times N}$

		Mutation types				
		A[C>A]A	A[C>A]C	A[C>A]G		T[T>G]T
		1	2	3	$\cdots$	N
Patients	1	$V_{1,1}$	$V_{1,2}$		$\cdots$	$V_{1,N}$
	2	$V_{2,1}$	$V_{2,2}$		$\cdots$	$V_{2,N}$
	3					
	$\vdots$	$\vdots$	$\vdots$			$\vdots$
	M	$V_{M,1}$	$V_{M,2}$		$\cdots$	$V_{M,N}$

Number of patients $M = 21$. Mutation types $N = 96$.

Non-negative Matrix Factorization

Matrix Factorization: $V \approx WH$

Non-negative: $W_{mk} \geq 0$ and $H_{mk} \geq 0$

Here $V \in \mathbb{R}^{M \times N}$, $W \in \mathbb{R}^{M \times K}$ and $H \in \mathbb{R}^{K \times N}$

Interpretation:
- Columns in H are the mutational signatures from exposures such as UV-light or tobacco.
- Rows in W are the weights of the signatures for each patient.

c. Estimation: High-dimensional optimization problem.

High-dimensional optimization problem

M = Number of patients = 21
N = Number of mutation types = 96
K = Number of signatures = 4

$$V_{(M\times N)} \approx W_{(M\times K)} H_{(K\times N)}$$

Number of data points = $M \cdot N \approx 2000$
Number of parameters = $M \cdot K + N \cdot K = (M+N) \cdot K \approx 500$

How to solve?

d. Alternating non-negative least sequare (NLS).

Alternating non-negative least squares (NLS)

$$V_{(M\times N)} \approx W_{(M\times K)} H_{(K\times N)}$$

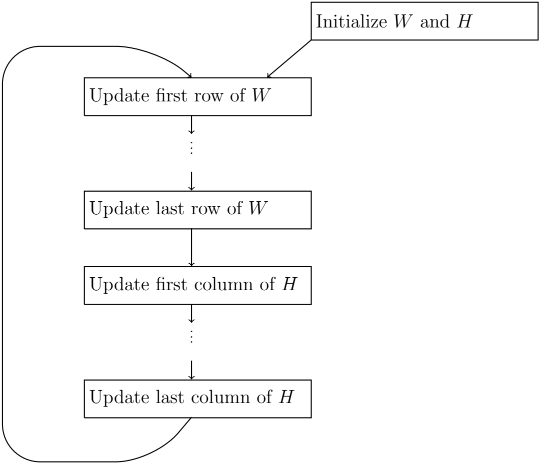
Cost (least square) function to minimize

$$\begin{aligned}
\| \, V - WH \, \|^2 &= \sum_{m=1}^{M} \left\{ \sum_{n=1}^{N} \Big(V_{mn} - (WH)_{mn} \Big)^2 \right\} \\
&= \sum_{n=1}^{N} \left\{ \sum_{m=1}^{M} \Big(V_{mn} - (WH)_{mn} \Big)^2 \right\}
\end{aligned}$$

where $(WH)_{mn} = \sum_{k=1}^{K} W_{mk} H_{kn}$

Strategy:
Alternate between updating rows m of W and columns n of H

Alternating non-negative least squares



Let $\mathrm{RSS}^{(t)} = \| V - W^{(t)} H^{(t)} \|$.
Stop when $\triangle\mathrm{RSS}^{(t)} = \mathrm{RSS}^{(t-1)} - \mathrm{RSS}^{(t)}$ is small.

Non-negative Least Square (NLS) problem

Update column n of H

Let $h = (H_{1n}, \ldots, H_{Kn}) \in \mathbb{R}^K$ and $v = (V_{1n}, \ldots, V_{Mn}) \in \mathbb{R}^M$. Then

$$\sum_{m=1}^{M} \Big(V_{mn} - (WH)_{mn} \Big)^2 = \| v - Wh \|^2$$

Problem:

$$\text{minimize} \;\; f(h) = \| v - Wh \|^2 \quad \text{subject to} \;\; h \geq 0$$

A small calculation:

$$(v - Wh)'(v - Wh) = h'W'Wh - 2v'Wh + v'v = \frac{1}{2}h'Ah + b'h + c$$

where $A = 2W'W$, $b = -2W'v$ and $c = v'v$.

Equivalent Problem:

$$\text{minimize} \;\; f(h) = \frac{1}{2}h'Ah + b'h \;\; \text{subject to} \;\; h \geq 0$$

e. Solutions to the NLS problem

Solutions to the non-negative least square problem

Three traditional iterative methods[1]:

1. Majorize-Minimize (MM)
2. Projected Coordinate Descent (PCD)
3. Projected Gradient Descent (PGD)

We compare to an alternative iterative method:

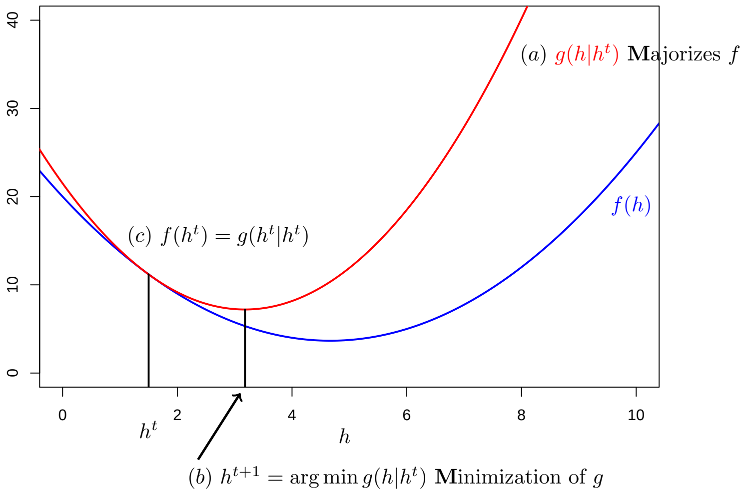
4. Expectation-maximization (EM)

And an alternative direct method from convex analysis:

5. Cone projection (CNP)

[1] Lange, Chi and Zhou (2014): Modern optimization for statisticians

1. Majorize-Minimize: Main idea



$$f(h^{t+1}) \underset{(a)}{\leq} g(h^{t+1}|h^t) \underset{(b)}{\leq} g(h^t|h^t) \underset{(c)}{=} f(h^t)$$

Problem:

$$\text{minimize} \;\; f(h) = \frac{1}{2} h' A h + b' h \;\; \text{subject to} \;\; h \geq 0$$

Note that

$$h' A h = \sum_{k=1}^{K} \sum_{l=1}^{K} A_{kl} h_k h_l$$

We use

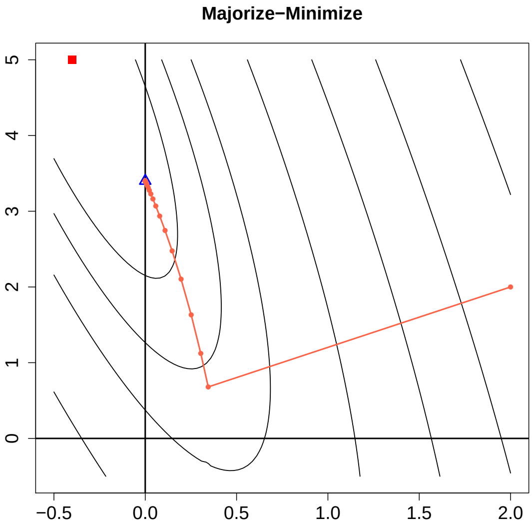
$$xy \leq \frac{1}{2}(\alpha x^2 + y^2/\alpha)$$

where $\alpha = y^t / x^t$.

The equality is derived from

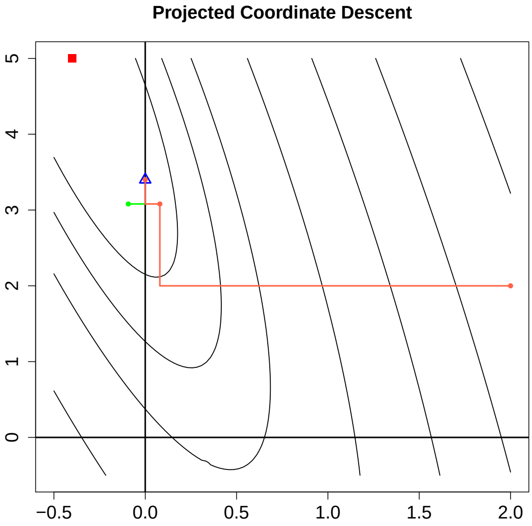
$$0 \leq (\sqrt{\alpha} x - y/\sqrt{\alpha})^2$$

$$W = \begin{bmatrix} 10 & 1 \\ 5 & 2 \end{bmatrix} \text{ and } v = (1, 8) \text{ and } h^0 = (2, 2)$$



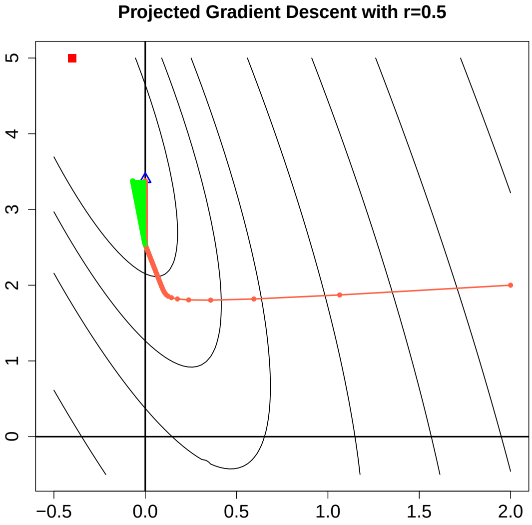
2. Projected Coordinate Descent

$$h_k^{t+1} = h_k^t - \frac{\nabla f(h^t)_k}{A_{kk}}$$

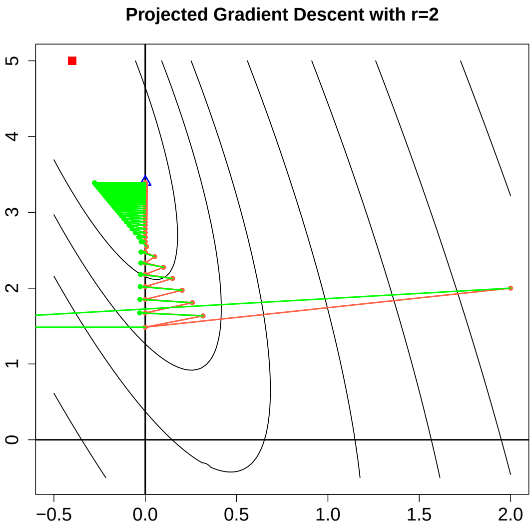


3. Projected Gradient Descent: Small step size

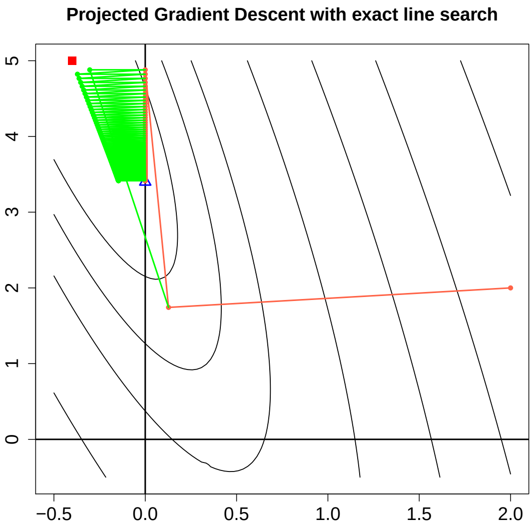
$$h^{t+1} = h^t - s\nabla f(h^t)$$



3. Projected Gradient Descent: Large step size



3. Projected Gradient Descent: Exact line search



4. Expectation-Maximization (EM) algorithm

Problem:

$$\text{minimize} \;\; f(h) = \| v - Wh \|^2 \;\;\; \text{subject to} \;\; h \geq 0$$

Corresponding statistical model

$$v_m \sim N((Wh)_m, \omega^2), \;\; m = 1, \ldots, M$$

We can view the problem as a Missing Data Problem:
Complete Data:

$$x_{mk} \sim N(w_{mk} h_k, \sigma_k^2) \;\; \text{and} \;\; y \sim N(0, \tau^2)$$

Observed Data:

$$v_m = y + \sum_{k=1}^{K} x_{mk} \sim N\Big(\sum_{k=1}^{K} w_{mk} h_k, \tau^2 + \sum_{k=1}^{K} \sigma_k^2 \Big) = N\Big((Wh)_m, \tau^2 + \sigma^2 \Big)$$

where $\sigma^2 = \sum_{k=1}^{K} \sigma_k^2$

The family of EM updates

Update rule from iteration t to iteration $t+1$:

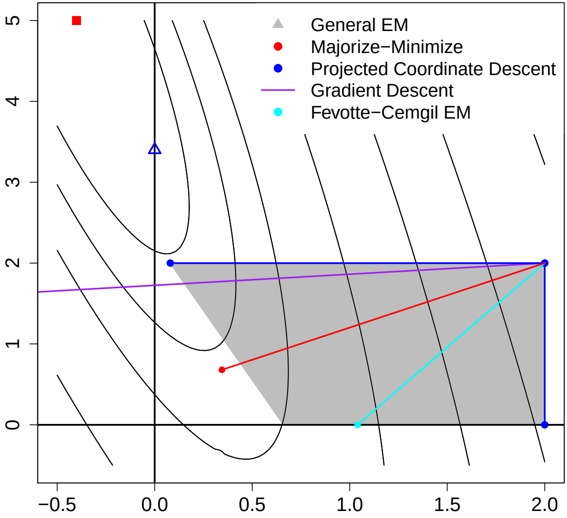
$$h_k^{t+1} = h_k^t - \frac{\sigma_k^2}{\tau^2 + \sigma^2} \frac{\nabla f(h^t)_k}{A_{kk}}$$

where $\sigma_k^2,\ k = 1, \ldots, K$ and τ^2 are free 'tuning parameters'

- Special Cases:
  Coordinate Descent ($\sigma_l^2 = 0, l \neq k,\ \tau^2$ technical)
  Gradient Descent ($\sigma_k^2 = A_{kk}, \tau^2 \geq 0$)
  Fevotte-Cemgil[2] ($\sigma_k^2 = 1,\ \tau^2 = 0$)
  Multiplicative EM ($\sigma_k^2 = h_k^t A_{kk}$)
- *Never* a Special Case:
  Majorize-Minimize.

[2] Fevotte and Cemgil (2009). NMF as probablistic inference in composite models. European Signal Processing Conference, Glasgow.

Overview



General EM
Majorize−Minimize
Projected Coordinate Descent
Gradient Descent
Fevotte−Cemgil EM
−0.5
0.0
0.5
1.0
1.5
2.0
0
1
2
3
4
5

5. Cone Projection

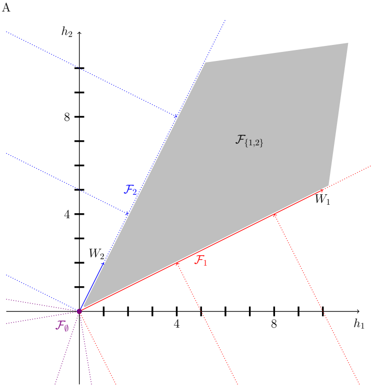
Problem:

$$\text{minimize} \quad \| \, v - Wh \, \|^2 \quad \text{subject to} \;\; h \geq 0$$

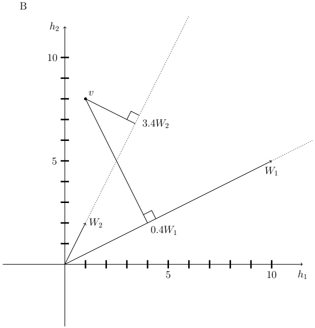
Example:

$$W = \begin{bmatrix} 10 & 1 \\ 5 & 2 \end{bmatrix}$$

Number of faces is 2^K



Brute Force Cone Projection



Subset $J \subseteq \{1,2\}$	Generator $X = W_J$	Coefficient $a = (X^T X)^{-1} X^T y$	Non-negative coefficients?	Squared distance $\|\|v - Xa\|\|^2$
$J = \{1,2\}$	W	$(-0.4, 5)$	No	0.0
$J = \{1\}$	W_1	0.4	Yes	45.0
$J = \{2\}$	W_2	3.4	Yes	7.2
$J = \emptyset$	$(0,0)^T$	0	Yes	65.0

A clever and efficient search in the index sets J is available (a modified algorithm from Meyer (2013): Quadratic programming with applications in statistics.)

f. Test problem and comparison of algorithms.

Test problem[3]

$W_{(100\times 50)}$	Entries from exponential with rate one.
$h_{(50\times 1)}$	Entries from standard uniform.
$e_{(100\times 1)}$	Entries from standard normal.

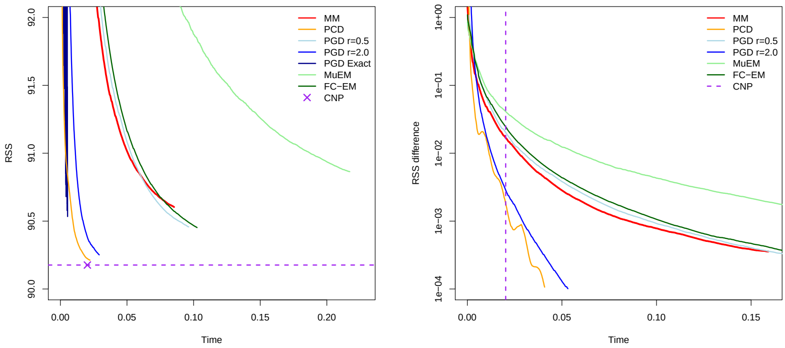
$$v = Wh + e$$

Aim: Recover the value of h, i.e.

$$\text{minimize } \; \mathrm{RSS}(h) = \parallel v - Wh \parallel^2 \quad \text{subject to } \; h \geq 0$$

[3]Modified from Lange, Chi and Zhou (2014)

Comparison on test problem



All algorithms are converging, but stopping criterion is not easy.
CNP attractive if stopping criterion is $\triangle\mathrm{RSS}^{(t)} < 0.001$.
PCD and PGD with $r = 2.0$ are also very fast.