Identifying Textual Clusters with Non-negative Matrix Factorization

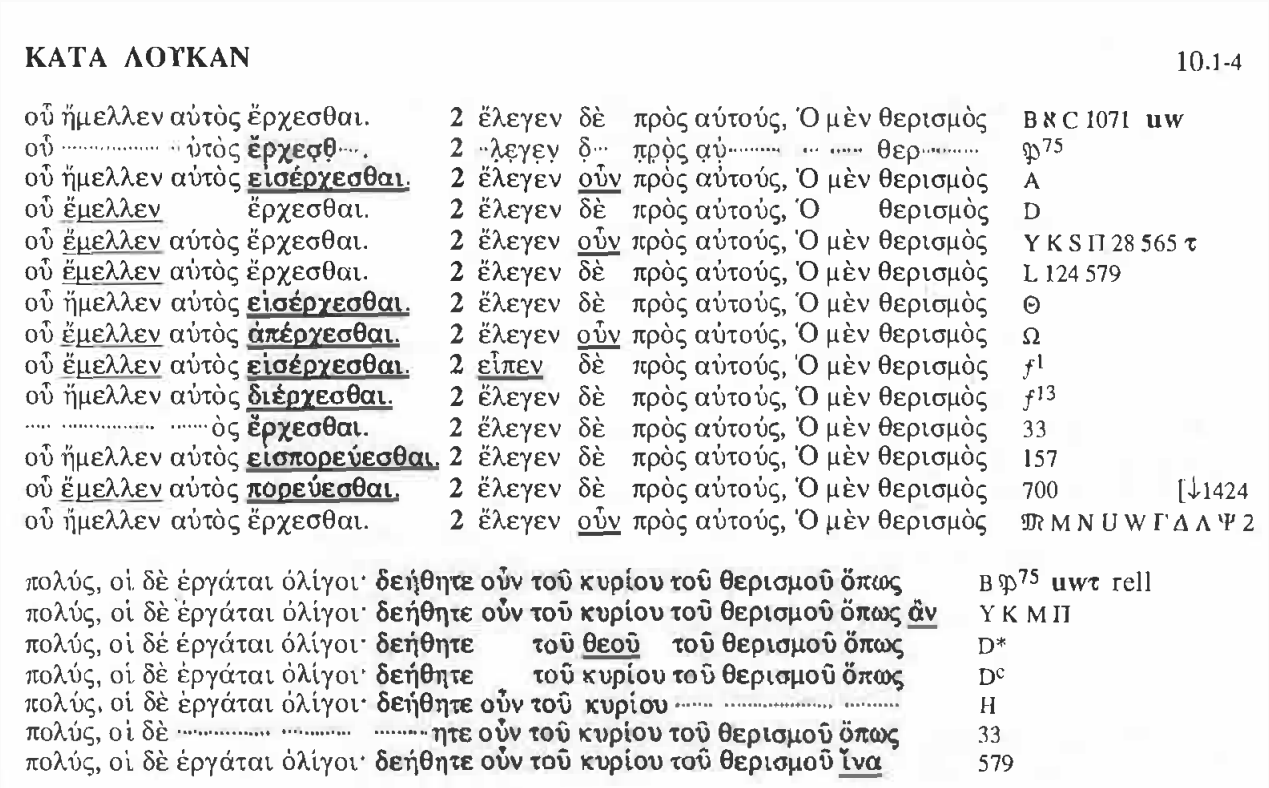
Joey McCollum*

16 September 2020

*Virginia Polytechnic Institute and State University

How Do We Compare Manuscripts?

- Start with collation—aligning texts at *variation units*



ΚΑΤΑ ΛΟΥΚΑΝ 10.1-4

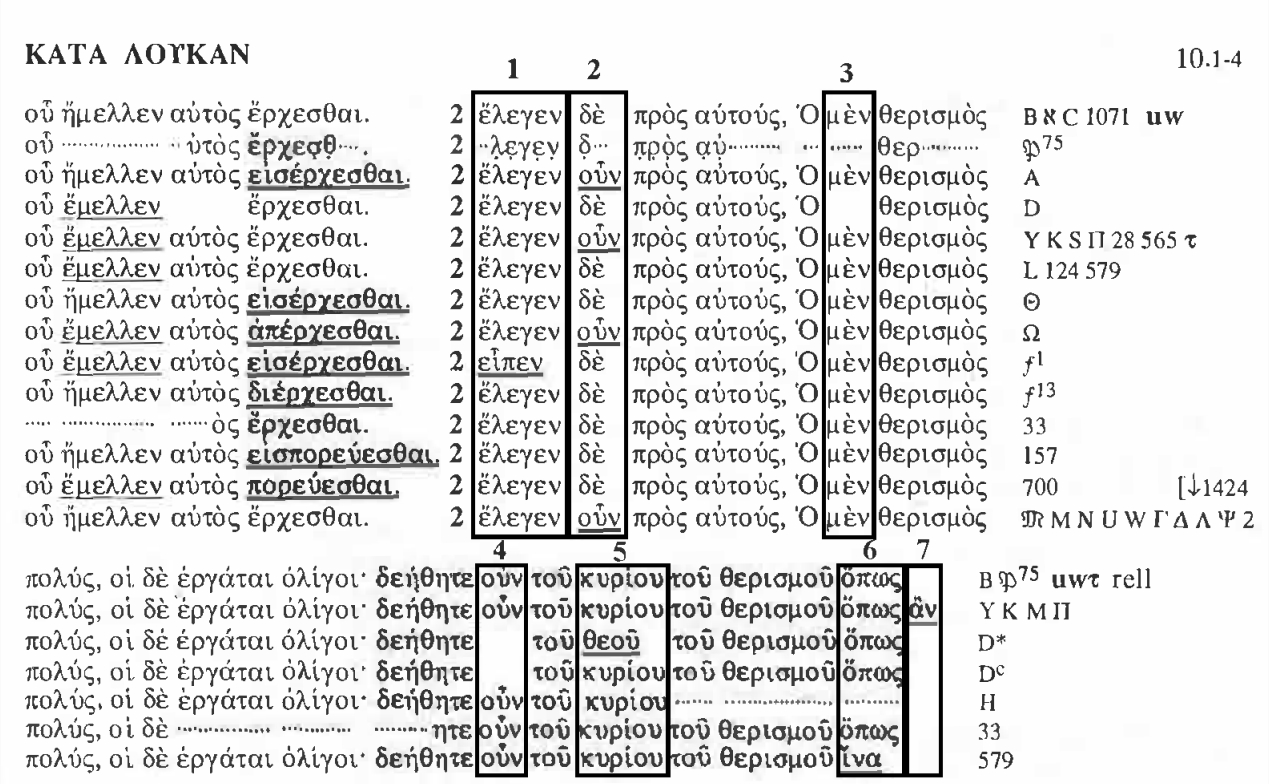
οὗ ἤμελλεν αὐτὸς ἔρχεσθαι.	2	ἔλεγεν	δὲ	πρὸς αὐτούς,	Ὁ μὲν θερισμὸς	B ℵ C 1071	**uw**
οὗ ····· ··ὐτὸς **ἔρχεσθ**···.	2	··λεγεν	δ···	πρὸς αὐ·····	··· ····· θερ·····	𝔓75	
οὗ ἤμελλεν αὐτὸς **εἰσέρχεσθαι.**	2	ἔλεγεν	**οὖν**	πρὸς αὐτούς,	Ὁ μὲν θερισμὸς	A	
οὗ **ἔμελλεν** ἔρχεσθαι.	2	ἔλεγεν	δὲ	πρὸς αὐτούς,	Ὁ θερισμὸς	D	
οὗ **ἔμελλεν** αὐτὸς ἔρχεσθαι.	2	ἔλεγεν	**οὖν**	πρὸς αὐτούς,	Ὁ μὲν θερισμὸς	Y K S Π 28 565 τ	
οὗ **ἔμελλεν** αὐτὸς ἔρχεσθαι.	2	ἔλεγεν	δὲ	πρὸς αὐτούς,	Ὁ μὲν θερισμὸς	L 124 579	
οὗ ἤμελλεν αὐτὸς **εἰσέρχεσθαι.**	2	ἔλεγεν	δὲ	πρὸς αὐτούς,	Ὁ μὲν θερισμὸς	Θ	
οὗ **ἔμελλεν** αὐτὸς **ἀπέρχεσθαι.**	2	ἔλεγεν	**οὖν**	πρὸς αὐτούς,	Ὁ μὲν θερισμὸς	Ω	
οὗ **ἔμελλεν** αὐτὸς **εἰσέρχεσθαι.**	2	**εἶπεν**	δὲ	πρὸς αὐτούς,	Ὁ μὲν θερισμὸς	f^{1}	
οὗ ἤμελλεν αὐτὸς **διέρχεσθαι.**	2	ἔλεγεν	δὲ	πρὸς αὐτούς,	Ὁ μὲν θερισμὸς	f^{13}	
····· ····· ·····ὸς **ἔρχεσθαι.**	2	ἔλεγεν	δὲ	πρὸς αὐτούς,	Ὁ μὲν θερισμὸς	33	
οὗ ἤμελλεν αὐτὸς **εἰσπορεύεσθαι.**	2	ἔλεγεν	δὲ	πρὸς αὐτούς,	Ὁ μὲν θερισμὸς	157	
οὗ **ἔμελλεν** αὐτὸς **πορεύεσθαι.**	2	ἔλεγεν	δὲ	πρὸς αὐτούς,	Ὁ μὲν θερισμὸς	700	[↓1424
οὗ ἤμελλεν αὐτὸς ἔρχεσθαι.	2	ἔλεγεν	**οὖν**	πρὸς αὐτούς,	Ὁ μὲν θερισμὸς	𝔐 M N U W Γ Δ Λ Ψ 2	

πολύς, οἱ δὲ ἐργάται ὀλίγοι· **δεήθητε οὖν τοῦ κυρίου τοῦ θερισμοῦ ὅπως**	B 𝔓75	**uwτ** rell
πολύς, οἱ δὲ ἐργάται ὀλίγοι· **δεήθητε οὖν τοῦ κυρίου τοῦ θερισμοῦ ὅπως ἂν**	Y K M Π	
πολύς, οἱ δέ ἐργάται ὀλίγοι· **δεήθητε τοῦ θεοῦ τοῦ θερισμοῦ ὅπως**	D*	
πολύς, οἱ δὲ ἐργάται ὀλίγοι· **δεήθητε τοῦ κυρίου τοῦ θερισμοῦ ὅπως**	D^{c}	
πολύς, οἱ δὲ ἐργάται ὀλίγοι· **δεήθητε οὖν τοῦ κυρίου** ····· ····· ·····	H	
πολύς, οἱ δὲ ····· ····· ·····ητε **οὖν τοῦ κυρίου τοῦ θερισμοῦ ὅπως**	33	
πολύς, οἱ δὲ ἐργάται ὀλίγοι· **δεήθητε οὖν τοῦ κυρίου τοῦ θερισμοῦ ἵνα**	579	

(Source: Swanson, *New Testament Greek Manuscripts, Luke*, 183)

How Do We Compare Manuscripts?

- Start with collation—aligning texts at *variation units*



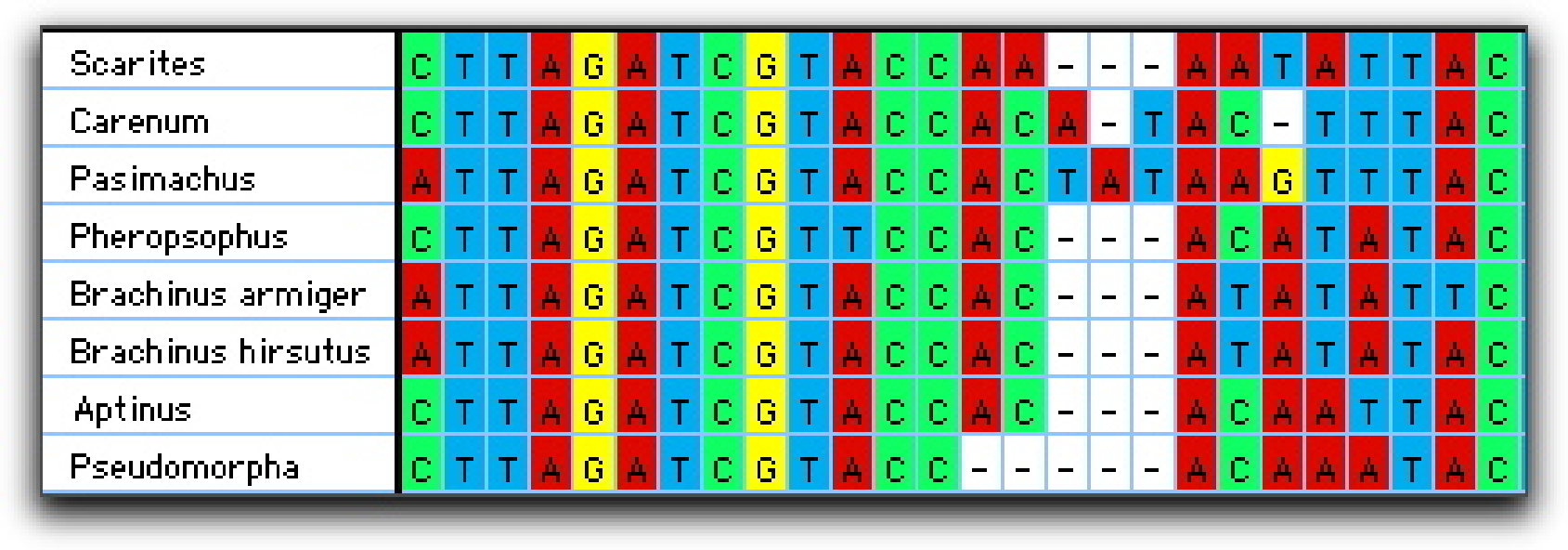
ΚΑΤΑ ΛΟΥΚΑΝ 10.1-4

	1	2			3		
οὗ ἤμελλεν αὐτὸς ἔρχεσθαι.	2	ἔλεγεν	δὲ	πρὸς αὐτούς, Ὁ	μὲν	θερισμὸς	B ℵ C 1071 **uw**
οὗ ····· ·ὐτὸς ἔρχεσθ···.	2	·λεγεν	δ···	πρὸς αὐ·····	···	θερ·····	𝔓75
οὗ ἤμελλεν αὐτὸς εἰσέρχεσθαι.	2	ἔλεγεν	οὖν	πρὸς αὐτούς, Ὁ	μὲν	θερισμὸς	A
οὗ ἔμελλεν ἔρχεσθαι.	2	ἔλεγεν	δὲ	πρὸς αὐτούς, Ὁ		θερισμὸς	D
οὗ ἔμελλεν αὐτὸς ἔρχεσθαι.	2	ἔλεγεν	οὖν	πρὸς αὐτούς, Ὁ	μὲν	θερισμὸς	Y K S Π 28 565 τ
οὗ ἔμελλεν αὐτὸς ἔρχεσθαι.	2	ἔλεγεν	δὲ	πρὸς αὐτούς, Ὁ	μὲν	θερισμὸς	L 124 579
οὗ ἤμελλεν αὐτὸς εἰσέρχεσθαι.	2	ἔλεγεν	δὲ	πρὸς αὐτούς, Ὁ	μὲν	θερισμὸς	Θ
οὗ ἔμελλεν αὐτὸς ἀπέρχεσθαι.	2	ἔλεγεν	οὖν	πρὸς αὐτούς, Ὁ	μὲν	θερισμὸς	Ω
οὗ ἔμελλεν αὐτὸς εἰσέρχεσθαι.	2	εἶπεν	δὲ	πρὸς αὐτούς, Ὁ	μὲν	θερισμὸς	f^1
οὗ ἤμελλεν αὐτὸς διέρχεσθαι.	2	ἔλεγεν	δὲ	πρὸς αὐτούς, Ὁ	μὲν	θερισμὸς	f^{13}
····· ····· ·····ὸς ἔρχεσθαι.	2	ἔλεγεν	δὲ	πρὸς αὐτούς, Ὁ	μὲν	θερισμὸς	33
οὗ ἤμελλεν αὐτὸς εἰσπορεύεσθαι.	2	ἔλεγεν	δὲ	πρὸς αὐτούς, Ὁ	μὲν	θερισμὸς	157
οὗ ἔμελλεν αὐτὸς πορεύεσθαι.	2	ἔλεγεν	δὲ	πρὸς αὐτούς, Ὁ	μὲν	θερισμὸς	700 [↓1424
οὗ ἤμελλεν αὐτὸς ἔρχεσθαι.	2	ἔλεγεν	οὖν	πρὸς αὐτούς, Ὁ	μὲν	θερισμὸς	𝔐 M N U W Γ Δ Λ Ψ 2

	4	5		6	7	
πολύς, οἱ δὲ ἐργάται ὀλίγοι· δεήθητε	οὖν	τοῦ	κυρίου	τοῦ θερισμοῦ	ὅπως	B 𝔓75 **uwτ** rell
πολύς, οἱ δὲ ἐργάται ὀλίγοι· δεήθητε	οὖν	τοῦ	κυρίου	τοῦ θερισμοῦ	ὅπως ἂν	Y K M Π
πολύς, οἱ δὲ ἐργάται ὀλίγοι· δεήθητε		τοῦ	θεοῦ	τοῦ θερισμοῦ	ὅπως	D*
πολύς, οἱ δὲ ἐργάται ὀλίγοι· δεήθητε		τοῦ	κυρίου	τοῦ θερισμοῦ	ὅπως	D^c
πολύς, οἱ δὲ ἐργάται ὀλίγοι· δεήθητε	οὖν	τοῦ	κυρίου	····· ·····	·····	H
πολύς, οἱ δὲ ····· ····· ·····ητε	οὖν	τοῦ	κυρίου	τοῦ θερισμοῦ	ὅπως	33
πολύς, οἱ δὲ ἐργάται ὀλίγοι· δεήθητε	οὖν	τοῦ	κυρίου	τοῦ θερισμοῦ	ἵνα	579

How Do We Compare Manuscripts?

- Comparable to DNA sequence alignment[1]
  - manuscripts ⟷ taxa / species
  - variation units ⟷ sites
  - variant readings ⟷ bases (A, C, G, T) and gaps (–)



(Source: http://www.sequence-alignment.com)

1. For the fascinating history of this relationship, see Lin, *The Erotic Life of Manuscripts*.

How Do We Compare Manuscripts?

- This provides a simple basis of comparison between pairs of manuscripts
  - Number of units where both agree
  - For a proportion, divide by number of units where the readings of both are known
- “Pre-genealogical coherence” in the Coherence-Based Genealogical Method (CBGM)

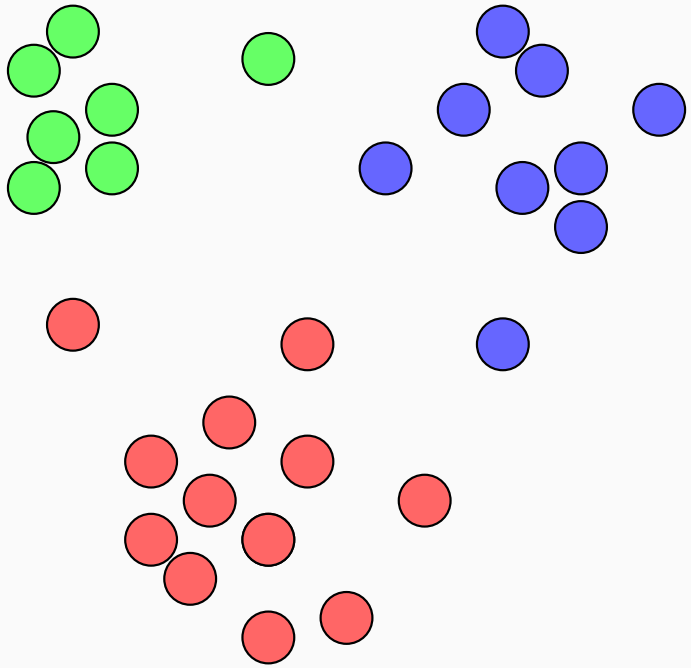
Genealogical comparisons for W1 = 5:

W2	DIR	NR	PASS	EQ	
35	>	4	115	101	(87.826%)
453	>	4	116	101	(87.069%)
03	>	7	116	98	(84.483%)
1611	<		116	98	(84.483%)
88	<		116	97	(83.621%)
1739	>	8	115	97	(84.348%)

- Can we use mutual agreement to classify manuscripts into groups?

The Quantitative Method

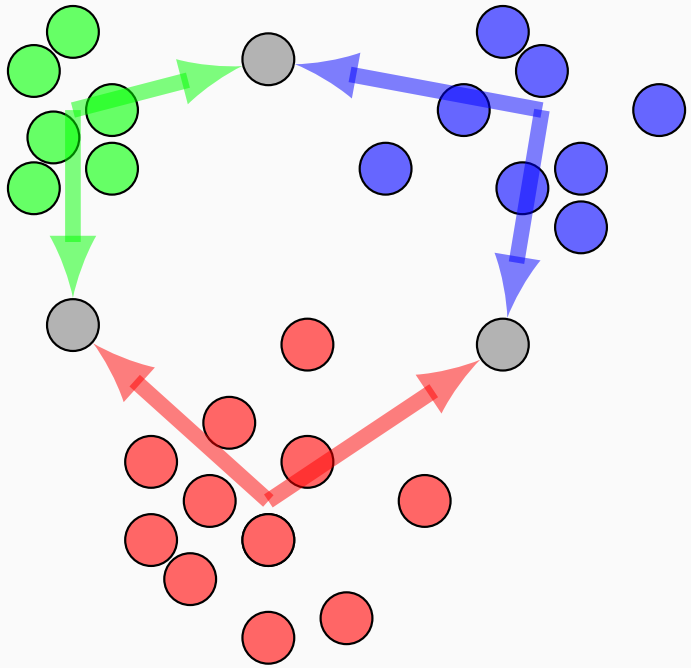
- Colwell and Tune: if manuscripts agree significantly more with one another than they do with other manuscripts, then they form a family, or *text-type*[2]
  - $\geq 70\%$ with one another, and $\geq 10\%$ more than with others



2. Colwell and Tune, “Quantitative Relationships.”

The Quantitative Method

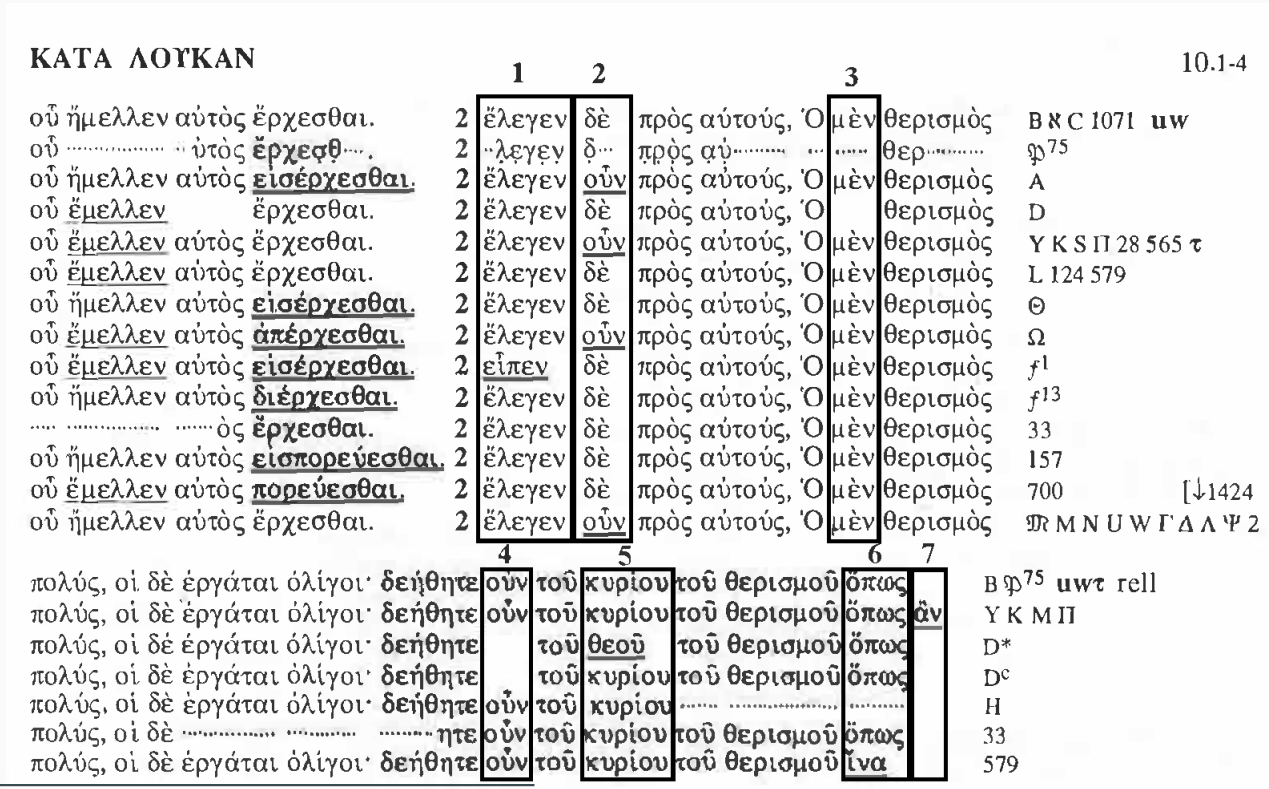
- Problems:
  - All units (including those involving singular readings and common scribal errors) have equal weight
  - Mixture in the transmission process is a problem[3]



3. Epp, “Textual Clusters.”

The Quantitative Method

- For efficiency and accuracy, comparisons should be done on the basis of *informative* points of variation[4]
- Specific readings, not the variation units containing them
- But how do we know which ones are the most informative?



ΚΑΤΑ ΛΟΥΚΑΝ 10.1-4

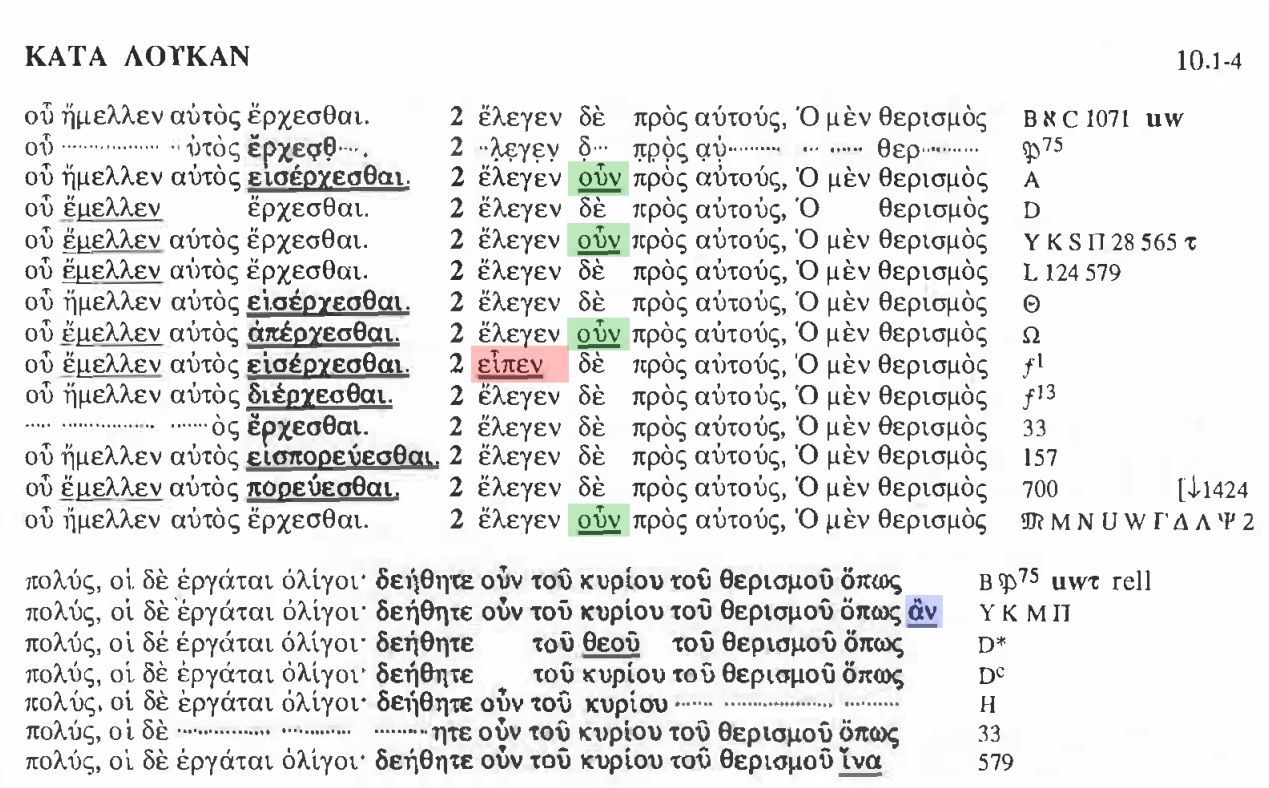
	1	2		3		
οὗ ἤμελλεν αὐτὸς ἔρχεσθαι.	2 ἔλεγεν	δὲ	πρὸς αὐτούς, Ὁ	μὲν	θερισμὸς	B ℵ C 1071 **uw**
οὗ ..... ὐτὸς ἔρχεσθ....	2 ..λεγεν	δ...	πρὸς αὐ.....	...	θερ.....	$\mathfrak{P}^{75}$
οὗ ἤμελλεν αὐτὸς εἰσέρχεσθαι.	2 ἔλεγεν	οὖν	πρὸς αὐτούς, Ὁ	μὲν	θερισμὸς	A
οὗ ἔμελλεν ἔρχεσθαι.	2 ἔλεγεν	δὲ	πρὸς αὐτούς, Ὁ		θερισμὸς	D
οὗ ἔμελλεν αὐτὸς ἔρχεσθαι.	2 ἔλεγεν	οὖν	πρὸς αὐτούς, Ὁ	μὲν	θερισμὸς	Y K S Π 28 565 τ
οὗ ἔμελλεν αὐτὸς ἔρχεσθαι.	2 ἔλεγεν	δὲ	πρὸς αὐτούς, Ὁ	μὲν	θερισμὸς	L 124 579
οὗ ἤμελλεν αὐτὸς εἰσέρχεσθαι.	2 ἔλεγεν	δὲ	πρὸς αὐτούς, Ὁ	μὲν	θερισμὸς	Θ
οὗ ἔμελλεν αὐτὸς ἀπέρχεσθαι.	2 ἔλεγεν	οὖν	πρὸς αὐτούς, Ὁ	μὲν	θερισμὸς	Ω
οὗ ἔμελλεν αὐτὸς εἰσέρχεσθαι.	2 εἶπεν	δὲ	πρὸς αὐτούς, Ὁ	μὲν	θερισμὸς	f^1
οὗ ἤμελλεν αὐτὸς διέρχεσθαι.	2 ἔλεγεν	δὲ	πρὸς αὐτούς, Ὁ	μὲν	θερισμὸς	f^{13}
....ὸς ἔρχεσθαι.	2 ἔλεγεν	δὲ	πρὸς αὐτούς, Ὁ	μὲν	θερισμὸς	33
οὗ ἤμελλεν αὐτὸς εἰσπορεύεσθαι.	2 ἔλεγεν	δὲ	πρὸς αὐτούς, Ὁ	μὲν	θερισμὸς	157
οὗ ἔμελλεν αὐτὸς πορεύεσθαι.	2 ἔλεγεν	δὲ	πρὸς αὐτούς, Ὁ	μὲν	θερισμὸς	700 [↓1424
οὗ ἤμελλεν αὐτὸς ἔρχεσθαι.	2 ἔλεγεν	οὖν	πρὸς αὐτούς, Ὁ	μὲν	θερισμὸς	𝔐 M N U W Γ Δ Λ Ψ 2

	4	5		6	7	
πολύς, οἱ δὲ ἐργάται ὀλίγοι·	δεήθητε	οὖν τοῦ	κυρίου	τοῦ θερισμοῦ	ὅπως	B $\mathfrak{P}^{75}$ **uwτ** rell
πολύς, οἱ δὲ ἐργάται ὀλίγοι·	δεήθητε	οὖν τοῦ	κυρίου	τοῦ θερισμοῦ	ὅπως ἂν	Y K M Π
πολύς, οἱ δέ ἐργάται ὀλίγοι·	δεήθητε	τοῦ	θεοῦ	τοῦ θερισμοῦ	ὅπως	D*
πολύς, οἱ δὲ ἐργάται ὀλίγοι·	δεήθητε	τοῦ	κυρίου	τοῦ θερισμοῦ	ὅπως	D^c
πολύς, οἱ δὲ ἐργάται ὀλίγοι·	δεήθητε	οὖν τοῦ	κυρίου	....	.....	H
πολύς, οἱ δὲ	.....ητε	οὖν τοῦ	κυρίου	τοῦ θερισμοῦ	ὅπως	33
πολύς, οἱ δὲ ἐργάται ὀλίγοι·	δεήθητε	οὖν τοῦ	κυρίου	τοῦ θερισμοῦ	ἵνα	579

4. Colwell, “Method in Locating.”

The Claremont Profile Method

- Start with an established set of manuscript groups[5]
- Filter out variation units involving common types of variation and singular / subsingular readings to get a set of *test passages*
- Readings supported by group manuscripts = the group's *profile*



ΚΑΤΑ ΛΟΥΚΑΝ 10.1-4

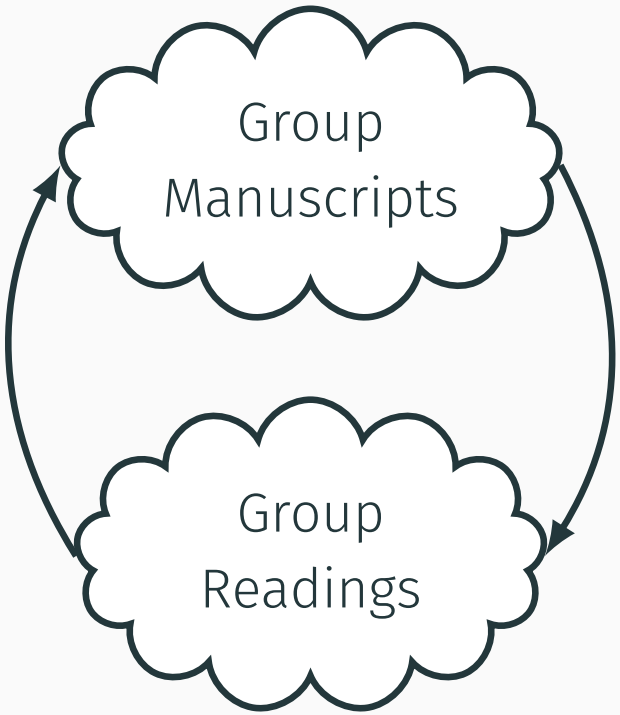
οὗ ἤμελλεν αὐτὸς ἔρχεσθαι.	2	ἔλεγεν	δὲ	πρὸς αὐτούς, Ὁ μὲν θερισμὸς	B ℵ C 1071 **uw**
οὗ ····· ·ὐτὸς **ἔρχεϙθ**···.	2	·λ̣εγεν	δ···	πρ̣ὸς α̣ὐ····· ··· ····· θερ·····	𝔓75
οὗ ἤμελλεν αὐτὸς **εἰσέρχεσθαι.**	2	ἔλεγεν	οὖν	πρὸς αὐτούς, Ὁ μὲν θερισμὸς	A
οὗ ἔμελλεν ἔρχεσθαι.	2	ἔλεγεν	δὲ	πρὸς αὐτούς, Ὁ θερισμὸς	D
οὗ ἔμελλεν αὐτὸς ἔρχεσθαι.	2	ἔλεγεν	οὖν	πρὸς αὐτούς, Ὁ μὲν θερισμὸς	Y K S Π 28 565 τ
οὗ ἔμελλεν αὐτὸς ἔρχεσθαι.	2	ἔλεγεν	δὲ	πρὸς αὐτούς, Ὁ μὲν θερισμὸς	L 124 579
οὗ ἤμελλεν αὐτὸς **εἰσέρχεσθαι.**	2	ἔλεγεν	δὲ	πρὸς αὐτούς, Ὁ μὲν θερισμὸς	Θ
οὗ ἔμελλεν αὐτὸς **ἀπέρχεσθαι.**	2	ἔλεγεν	οὖν	πρὸς αὐτούς, Ὁ μὲν θερισμὸς	Ω
οὗ ἔμελλεν αὐτὸς **εἰσέρχεσθαι.**	2	εἶπεν	δὲ	πρὸς αὐτούς, Ὁ μὲν θερισμὸς	f^{1}
οὗ ἤμελλεν αὐτὸς **διέρχεσθαι.**	2	ἔλεγεν	δὲ	πρὸς αὐτούς, Ὁ μὲν θερισμὸς	f^{13}
····· ····· ·····ὸς **ἔρχεσθαι.**	2	ἔλεγεν	δὲ	πρὸς αὐτούς, Ὁ μὲν θερισμὸς	33
οὗ ἤμελλεν αὐτὸς **εισπορεύεσθαι.**	2	ἔλεγεν	δὲ	πρὸς αὐτούς, Ὁ μὲν θερισμὸς	157
οὗ ἔμελλεν αὐτὸς **πορεύεσθαι.**	2	ἔλεγεν	δὲ	πρὸς αὐτούς, Ὁ μὲν θερισμὸς	700 [↓1424
οὗ ἤμελλεν αὐτὸς ἔρχεσθαι.	2	ἔλεγεν	οὖν	πρὸς αὐτούς, Ὁ μὲν θερισμὸς	𝔐 M N U W Γ Δ Λ Ψ 2

πολύς, οἱ δὲ ἐργάται ὀλίγοι· **δεήθητε οὖν τοῦ κυρίου τοῦ θερισμοῦ ὅπως**	B 𝔓75 **uwτ** rell
πολύς, οἱ δὲ ἐργάται ὀλίγοι· **δεήθητε οὖν τοῦ κυρίου τοῦ θερισμοῦ ὅπως ἂν**	Y K M Π
πολύς, οἱ δέ ἐργάται ὀλίγοι· **δεήθητε τοῦ θεοῦ τοῦ θερισμοῦ ὅπως**	D*
πολύς, οἱ δὲ ἐργάται ὀλίγοι· **δεήθητε τοῦ κυρίου τοῦ θερισμοῦ ὅπως**	D^{c}
πολύς, οἱ δὲ ἐργάται ὀλίγοι· **δεήθητε οὖν τοῦ κυρίου** ····· ····· ·····	H
πολύς, οἱ δὲ ····· ····· ·····**ητε οὖν τοῦ κυρίου τοῦ θερισμοῦ ὅπως**	33
πολύς, οἱ δὲ ἐργάται ὀλίγοι· **δεήθητε οὖν τοῦ κυρίου τοῦ θερισμοῦ ἵνα**	579

5. Wisse, *Profile Method*.

The Claremont Profile Method

- This allows us to isolate informative readings for group classification
- Also robust to mixture
- But it needs manuscript groups to be established first!
- “Good manuscripts have good readings, and good readings are found in good manuscripts”



Non-negative Matrix Factorization

- *Non-negative matrix factorization* (NMF), a machine learning technique, uses this circular relationship to solve both problems

- Represent our collation as a matrix $\mathbf{A}$ with a row for each variant reading and a column for each manuscript
- *m* rows by *n* columns

		$\mathfrak{P}^{75}$	A	B	D	K	f^1	579
Unit 1	ἔλεγεν	1	1	1	1	1	0	1
	εἶπεν	0	0	0	0	0	1	0
Unit 2	δὲ	1	0	1	1	0	1	1
	οὖν	0	1	0	0	1	0	0
Unit 3	μὲν	0	1	1	0	1	1	1
	omit	0	0	0	1	0	0	0
Unit 4	οὖν	1	1	1	0	1	1	1
	omit	0	0	0	1	0	0	0
Unit 5	κυρίου	1	1	1	0	1	1	1
	θεοῦ	0	0	0	1	0	0	0
Unit 6	ὅπως	1	1	1	1	1	1	0
	ἵνα	0	0	0	0	0	0	1
Unit 7	*omit*	1	1	1	1	0	1	1
	ἂν	0	0	0	0	1	0	0

Non-negative Matrix Factorization

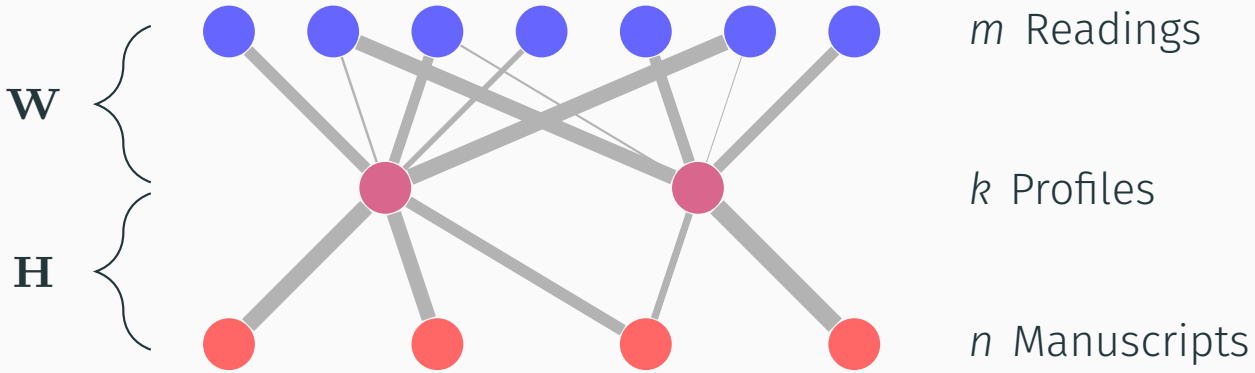
- The goal is to approximate this original matrix as the product of two smaller matrices with non-negative entries:

$$\mathbf{A} \approx \mathbf{WH}$$

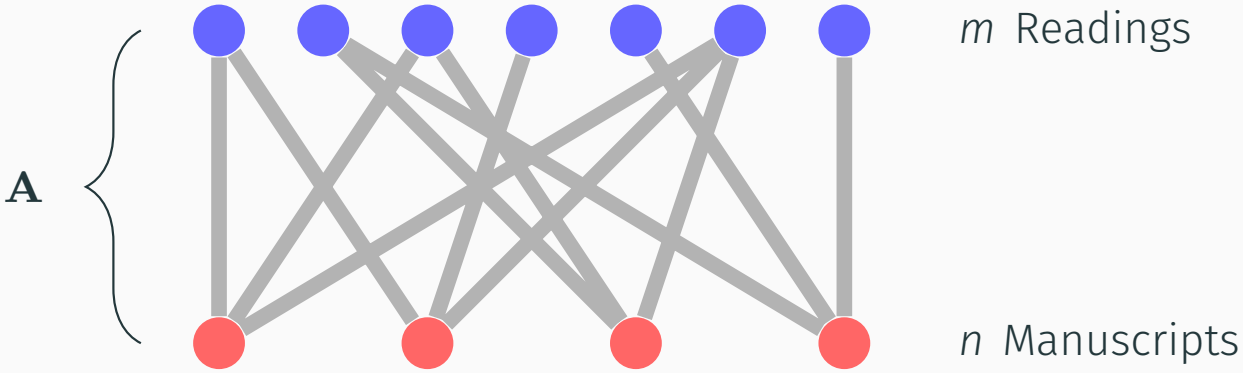
- Specify a number k of underlying textual profiles (there are metrics for finding good choices)
- $\mathbf{W}$: m rows and k columns; defines group readings
- $\mathbf{H}$: k rows and n columns; defines makeup of manuscripts in terms of profiles

Non-negative Matrix Factorization

- Use two sets of relationships with a few components...



...to reconstruct the large set of original relationships



Non-negative Matrix Factorization

- The process:
  1. Start with guesses for $\mathbf{W}$ and $\mathbf{H}$
  2. Fix $\mathbf{W}$, optimize the weights in $\mathbf{H}$ (Quantitative Method)
  3. Fix $\mathbf{H}$, optimize the weights in $\mathbf{W}$ (Claremont Profile Method)
  4. Repeat steps 2 and 3 until the difference between $\mathbf{A}$ and $\mathbf{WH}$ no longer decreases

- Guaranteed to terminate with *locally optimal* groupings in $\mathbf{W}$ and $\mathbf{H}$[6]

6. Grippo and Sciandrone, "On the Convergence."

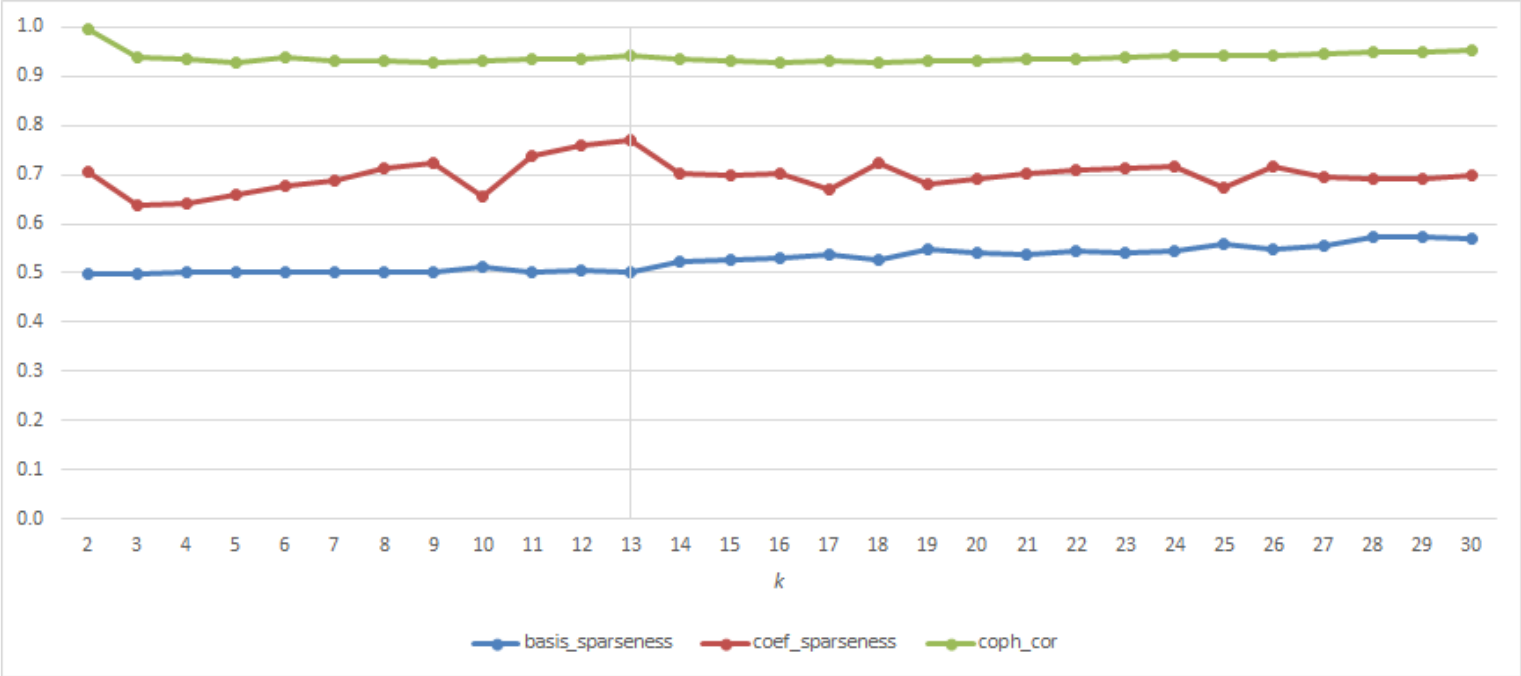
Results: Jude

- Tommy Wasserman's collation of Jude contains 1346 variant readings and 560 manuscripts (including lectionaries)[7]
- Filtering out 42 fragmentary manuscripts (< 300 known readings) yields a matrix $\mathbf{A}$ with $m = 1346$ rows and $n = 518$ columns
- The fragmentary manuscripts can be classified after groups are established[8]

7. Wasserman, *The Epistle of Jude*.
8. For details, see the appendix of McCollum, "Biclustering Readings and Manuscripts."

Results: Jude

- We select the number of profiles *k* based on several factors:
  - Overlap of readings in profiles
  - Mixture of profiles in manuscripts
  - Consistency of manuscript groupings when random starting points are used (the *cophenetic correlation coefficient*)[9]



9. Brunet et al., "Metagenes and Molecular Pattern Discovery."

Results: Jude

- The $k = 13$ groups identified by NMF correspond to groups in the Catholic Epistles identified in the literature

Members (by Gregory-Aland number)	Group
920, 1277, 1859, 1719, 452, 1857, 1871, 941, 1103, 1352, etc.	Kappa (von Soden)
141, 204, 394, 444, 1101, 1723, 1737, 1752, 1865, 2221, etc.	K^r (von Soden)
390, 1863, 912, 234, 1861, 2085, 1753, 2279, 42, 996, etc.	K^c (von Soden)
L606, L938, L145, L840, L740, L2106, L2394, L809, L1279, L62, etc.	Lectionary (Colwell)
606, 454, 641, 103, 221, 2125, 314, 250, 1888, 393, etc.	O, Θδ Commentaries (von Soden)
619, 1780, 1175, 330, 1769, 2516, 917, 451, 1162, 601, etc.	f^{1780} (unidentified)
1563, 1718, 1425, 1359, 1066, 0142, 056	f^{0142} (unidentified)

Results: Jude

- The $k = 13$ groups identified by NMF correspond to groups in the Catholic Epistles identified in the literature

Members (by Gregory-Aland number)	Group
03, 623, $\mathfrak{P}^{72}$, 81, 5, 326, 33, 1837, 93, 665, etc.	H (von Soden)
321, 918, 307, 453, 2197, 2818, 1678, 94, 2186, 1840, etc.	f^{453} (Spencer, Wachtel, Howe)
323, 1241, 322, 1739, 1881, 2298, 6	f^{1739} (Zuntz, Geer)
1505, 2495, 1611, 1292, 630, 2200, 1765, 1832, 2494, 876, etc.	f^{2138}/Harklean (Amphoux)
1843, 1869, 506, 1903, 489, 927, 203, 1868, 1729, 1873, etc.	I (von Soden)
915, 88, 459, 104, 1846, 1838, 1842, 1845	f^{915} (unidentified)

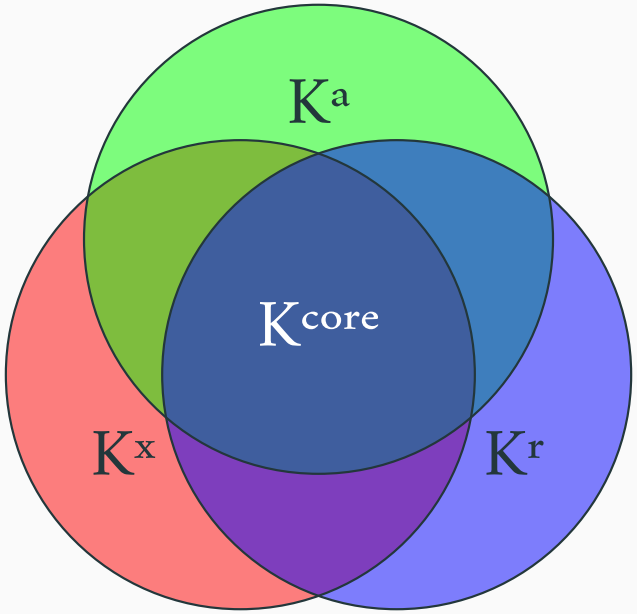
Results: John 18

- Applying NMF to Morrill's collation of all continuous-text manuscripts of John 18 illustrates some of the idiosyncrasies of the method and how to deal with them[10]
- Significantly larger and more "square" collation: $m = 1545$ variant readings and $n = 1610$ manuscripts after filtering out fragmentary manuscripts (< 350 known readings)
- (Recall that the collation matrix for Jude was 1346×518)

10. Morrill, "Complete Collation and Analysis."

Results: John 18

- Applying NMF to the matrix as-is separates readings common to multiple groups into their own “core” profiles
- No manuscripts belong to these profiles, but many appear “mixed” with it
- Symptom of volume and similarity of manuscripts, especially Byzantine ones

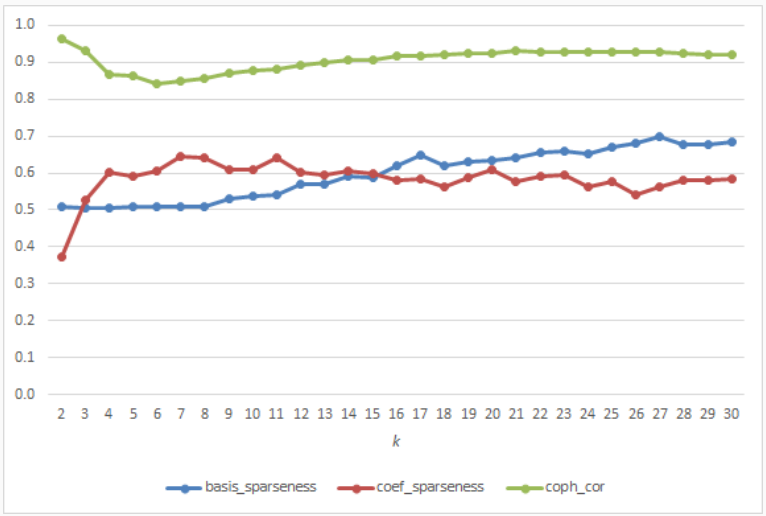


Results: John 18

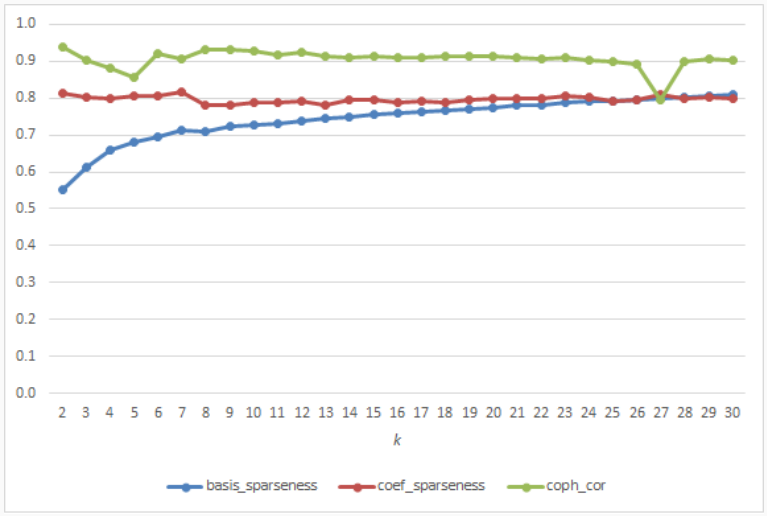
- To remedy this, weigh readings in the original matrix by their *inverse document frequency* (IDF)[11]

$$\log \frac{n}{\#\{\text{MSS with reading}\}}$$

- Removing singular readings is helpful in this setting
- Encourages NMF to isolate unique group readings in profiles



No weighting



IDF weighting

11. Jones, "Statistical Interpretation."

Results: John 18

- With $k = 12$, NMF identifies known groups from the literature

Members (by Gregory-Aland number)	Group
2605, 492, 1215, 2897, 1090, 1567, 1210, 851, 494, 2406, etc.	K^x (von Soden)
47, 1126, 61, 1138, 58, 56, 189, 1236, 825, 1614, etc.	K^r (von Soden)
2902, 1219, 1079, 489, 114, 2404, 389, 2193, 699, 1627, etc.	K^a (von Soden)
1534, 741, 857, 744, 2735, 1160, 817, 1261, 2470, 833, etc.	Θε Commentaries (von Soden)
892, 977, 555, 16, 152, 513, 1243, 829, 348, 1579, etc.	f^{16}+f^{1216} (Wisse)
1663, 1413, 2291, 86, 569, 71, 1170, 1014, 1531, 2705, etc.	M27+Cl1531 (Wisse)

Results: John 18

- With $k = 12$, NMF identifies known groups from the literature

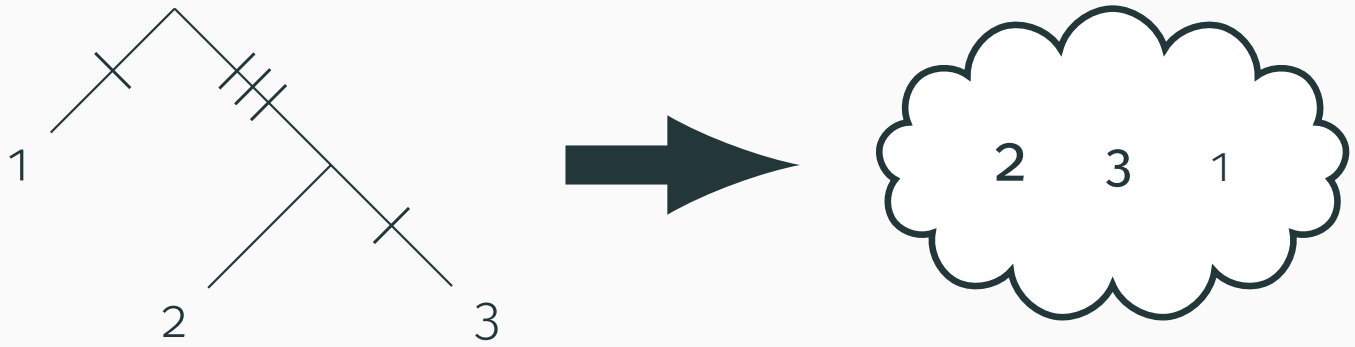
Members (by Gregory-Aland number)	Group
01, 032, 05, 579, 1654, 2561, 1242	Egyptian
1820, 2129, 865, 033, 019, 1819, 213, 03, 33, 1321, etc.	Alexandrian
1, 1582, 357, 138, 565, 209, 994, 2713, 2575, 1784, etc.	f^1 (Lake)
13, 788, 826, 828, 543, 69, 346, 1689, 124, 2786, etc.	f^{13} (Lake and Lake, Geerlings)
2524, 1001, 1268, 2397, 352, 2728, 132, 175, 1701, 2252, etc.	Cl1001+Cl352 (Wisse)
1446, 1050, 706, 1457, 827, 2620, 1128, 0211, 2707, 1402, etc.	Cl827 (Wisse)

Concluding Observations

- In John 18, Gregory-Aland 03 (Codex Vaticanus, B) stands out as an instructive example
- Appears to be mixed between the "Egyptian" and "Alexandrian" profiles, but could preserve a text earlier than both

	"03"
Kx	0.0290
Cl1001+Cl352	0.0000
Theophylact	0.0000
f13	0.0000
f1	0.0000
Alexandrian	1.1248
Egyptian	1.0556
Kr	0.0000
Ka	0.0000
M27+Cl1531	0.0000
f16+f1216	0.0000
Cl827	0.0000

- NMF identifies relationships, but not their directions
- Pre-genealogical, but not genealogical



Concluding Observations

- The advantage: few assumptions and editorial decisions are required
- Intended for use in “pre-processing” (manuscript and test reading selection)
- Useful for other applications (new manuscript classification)
- Work in progress: applying NMF to ~2000 manuscripts in the *pericope adulterae* (with Maurice A. Robinson)

References

Amphoux, Christian-B. “La Parenté Textuelle du $\mathbf{sy^h}$ et du Groupe 2138 dans l’Épître de Jacques.” *Bib* 62.2 (1981): 259–271.

Brunet, Jean-Philippe, Pablo Tamayo, Todd R. Golub, and Jill P. Mesirov. “Metagenes and Molecular Pattern Discovery Using Matrix Factorization.” *PNAS* 101.12 (2004): 4164–4169.

Colwell, Ernest C. “Is There a Lectionary Text of the Gospels?” *HTR* 25.1 (1932): 73–84.

———. “Method in Locating a Newly-Discovered Manuscript.” Pages 26–44 in *Studies in Methodology in Textual Criticism of the New Testament*. NTTSD 9. Leiden: Brill, 1969.

Colwell, Ernest C., and Ernest W. Tune. “The Quantitative Relationships between Text-types of New Testament Manuscripts.” Pages 56–62 in *Studies in Methodology in Textual Criticism of the New Testament*. NTTSD 9. Leiden: Brill, 1969.

Epp, Eldon Jay. “Textual Clusters: Their Past and Future in New Testament Textual Criticism.” Pages 519–577 in *The Text of the New Testament in Contemporary Research: Essays on the Status Quaestionis*. Edited by Bart D. Ehrman and Michael W. Holmes. 2nd ed. NTTSD 42. Leiden: Brill, 2012.

Geerlings, Jacob. “Family 13—The Ferrar Group: The Text according to John.” Studies and Documents. 21. 1962.

———. “Family 13—The Ferrar Group: The Text according to Luke.” Studies and Documents. 20. 1961.

———. “Family 13—The Ferrar Group: The Text according to Matthew.” Studies and Documents. 19. 1961.

Grippo, L., and S. Sciandrone. “On the Convergence of the Block Nonlinear Gauss-Seidel Method under Convex Constraints.” *Oper. Res. Lett.* 26 (2000): 127–136.

Jones, Karen Spärk. “A Statistical Interpretation of Term Specificity and Its Application in Retrieval.” *J. Doc.* 28.1 (1972): 11–21.

Lake, Kirsopp. “Codex 1 of the Gospels and Its Allies.” Texts and Studies. 7.3. Edited by J. Armitage Robinson (1902).

Lake, Kirsopp, and Silva Lake. “Family 13 (The Ferrar Group): The Text according to Mark, with a Collation of Codex 28 of the Gospels.” Studies and Documents. 11. 1941.

Lin, Yii-Jan. *The Erotic Life of Manuscripts: New Testament Textual Criticism and the Biological Sciences*. Oxford: Oxford University Press, 2016.

McCollum, Joey. “Biclustering Readings and Manuscripts via Non-negative Matrix Factorization, with Application to the Text of Jude.” *AUSS* 57.1 (2019): 61–89.

Morrill, M. Bruce. “A Complete Collation and Analysis of All Greek Manuscripts of John 18.” PhD diss., University of Birmingham, 2012.

Spencer, Matthew, Klaus Wachtel, and Christopher J. Howe. “The Greek *Vorlage* of the Syra Harclensis: A Comparative Study on Method in Exploring Textual Genealogy.” *TC* 7 (2002).

Swanson, Reuben J., ed. *New Testament Greek Manuscripts: Variant Readings Arranged in Horizontal Lines against Codex Vaticanus. Luke*. Sheffield: Sheffield Academic Press, 1995.

Thomas C. Geer, Jr. *Family 1739 in the Book of Acts*. SBLMS 48. Atlanta: Scholars Press, 1994.

Wasserman, Tommy. *The Epistle of Jude: Its Text and Transmission*. ConBNT 43. Stockholm: Almqvist; Wiksell International, 2006.

Wisse, Frederik. *The Profile Method for the Classification and Evaluation of Manuscript Evidence, as Applied to the Continuous Greek Text of the Gospel of Luke*. SD 44. Grand Rapids, MI: Wm. B. Eerdmans Publishing, 1982.

Zuntz, Günther. *The Text of the Epistles: A Disquisition upon the Corpus Paulinum, Schweich Lectures of 1946*. London: British Academy, 1953.