Andrews University Seminary Studies, Vol.57, No.1, 61–89.

BICLUSTERING READINGS AND MANUSCRIPTS VIA NON-NEGATIVE MATRIX FACTORIZATION, WITH APPLICATION TO THE TEXT OF JUDE

Joey McCollum
Virginia Polytechnic Institute and State University

Abstract

The text-critical practice of grouping witnesses into families or *texttypes* often faces two obstacles: the methodological question of how exactly to isolate the groups, given the chicken-and-egg relationship between "good" group readings and "good" group manuscripts, and contamination in the manuscript tradition. I introduce *non-negative matrix factorization* (NMF) as a simple, automated, and efficient solution to both problems. Within minutes, NMF can cluster hundreds of manuscripts and readings simultaneously, producing an output that details potential contamination according to an easy-to-interpret mixture model. I apply this method to Wasserman's extensive collation of the Epistle of Jude, showing that the resulting clusters correspond to human-identified textual families and their characteristic readings correctly divide witnesses into their groups. Due to its demonstrated accuracy, versatility, and speed, NMF could replace prior state-of-the-art classification methods and find fruitful application in a number of text-critical settings.

Keywords: New Testament, textual criticism, text families, manuscript relations, MSS classification, non-negative matrix factorization, Claremont Profile Method, Jude

Introduction

The analysis of genealogical relationships between manuscripts (hereafter MSS) played a prominent role in New Testament (hereafter NT) text-critical theory even before it was popularized in the work of Westcott and Hort.[1] Specifically, the principal step of classifying MSS into distinct families, or

[1] Brooke Foss Westcott and Fenton John Anthony Hort, *The New Testament in the Original Greek. Vol. 1: Text* (New York: Harper & Brothers, 1881).

texttypes, over a century-and-a-half earlier to the works of Mill, Bentley, and Bengel.[2] The underlying idea is that a large number of MSS can be reduced, on the basis of shared patterns or *profiles* of readings, to a smaller number of groups from which the textual critic can deduce a putative history of transmission.

The use of texttypes is not without obstacles, however. Deciding which MSS belong to which groups is already a nontrivial task, as it is intimately linked to the complementary task of assigning readings to groups. This connection has not always been obvious to textual critics; it has become apparent only through the shortcomings of methods that attempt to make either task depend entirely on the other. The earliest and simplest approaches to classifying MSS either ignored the relationship of readings to groups or postponed inferring it to a later stage,[3] but in practice, this was found insufficient. Most witnesses will agree on a majority of their readings, so weighing all readings equally only raises the question of just how different MSS need to be in order to to belong to different groups.[4] Later approaches, such as the Claremont Profile Method (CPM),[5] first grouped readings into profiles, and then attempted to classify MSS based on which profiles' readings they shared most. These approaches were more robust, but they left textual critics in another quandary. In order for readings to be assigned to groups, the

---

[2] Eldon Jay Epp, "Textual Clusters: Their Past and Future in New Testament Textual Criticism," in *The Text of the New Testament in Contemporary Research: Essays on the Status Quaestionis*, ed. Bart D. Ehrman and Michael W. Holmes, 2nd ed., NTTSD 42 (Leiden: Brill, 2012), 519–577, esp. 523–527.

[3] One of the earliest is the *quantitative method*, introduced in Ernest Cadman Colwell, "Method in Locating a Newly-Discovered Manuscript," in *Studies in Methodology in Textual Criticism of the New Testament*, NTTSD 9 (Leiden: Brill, 1969), 26–44; more recent studies exploring the same method, but with different similarity metrics and clustering rules, include J. C. Thorpe, "Multivariate Statistical Analysis for Manuscript Classification," *TC* 7 (2002) and Timothy J. Finney, "Mapping Textual Space," *TC* 15 (2010).

[4] See Gordon D. Fee, "The Text of John in Origen and Cyril of Alexandria: A Contribution to Methodology in the Recovery and Analysis of Patristic Citations," *Bib* 52.3 (1971): 357–394, esp. 364–365 and Bart D. Ehrman, "The Use of Group Profiles for the Classification of New Testament Documentary Evidence," *JBL* 106.3 (1987): 465–486, esp. 465–466. See Timothy J. Finney, "How to Discover Textual Groups," *Digital Studies / le Champ Numérique* 8.1 (2018): 7 for a statistical approach to establishing thresholds for dissimilarity.

[5] For introductory material, see Paul Robert McReynolds, "The Claremont Profile Method and the grouping of Byzantine New Testament Manuscripts" (PhD diss., Claremont Graduate School, 1969) and Frederik Wisse, *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: Eerdmans, 1982).

groups must already be established some other way, and if the only other way to do this was on the basis of MSS, then the whole process would beg the original question.[6] The root of the problem became a circular relationship: characteristic MSS of a given type are determined by which characteristic readings they have, and characteristic readings of a given type are determined by which characteristic MSS attest to them. The critical next step became the development of a method capable of solving both of these complementary problems simultaneously.

Even assuming a solution to the basic problem of isolating textual groups, traditional texttypes face another more robust threat. In constructing their genealogy of the NT text, Westcott and Hort overlooked the effects of *contamination*, or mixture of characteristic readings from different branches of transmission.[7] This oversight has proven to be problematic; as more MSS are discovered and studied, boundaries between the groups assigned to them become increasingly blurred.[8] Indeed, the problem of contamination among NT MSS has become so widely recognized that it has given rise to new text-critical methods, specifically tailored to account for it.[9]

---

[6] Indeed, the CPM has been criticized on the basis of its application with poorly-identified groups (W. Larry Richards. "A Critique of a New Testament Text-Critical Methodology: The Claremont Profile Method," *JBL* 96.4 [1977]: 555–566, esp. 562–565). Because of this, it is best used in conjunction with more quantitative methods (Ehrman, "Group Profiles," 469–471).

[7] Ernest Cadman Colwell, "Genealogical Method: Its Achievements and Its Limitations," *JBL* 66.2 (1947): 109–133, esp. 114–118.

[8] Epp, "Textual Clusters," 522.

[9] The most prominent of these is the *Coherence-Based Genealogical Method* (CBGM), and it has thus far been applied in the development of the *Editio Critica Maior* (ECM) for the General Epistles and Acts. The theoretical background for this method is detailed in Gerd Mink, "Problems of a Highly Contaminated Tradition: The New Testament. Stemmata of Variants as a Source of a Genealogy for Witnesses," in *Studies in Stemmatology II*, ed. Pieter van Reenen, August den Hollander, and Margot van Mulken (Amsterdam: John Benjamin, 2004), 13–85, and a student-friendly introduction can be found in Tommy Wasserman and Peter J. Gurry, *A New Approach to Textual Criticism: An Introduction to the Coherence-Based Genealogical Method*, RBS 80 (Atlanta: SBL Press, 2017). Another approach to the problem of contamination is explored in Matthew Spencer, Klaus Wachtel, and Christopher J. Howe, "Representing Multiple Pathways of Textual Flow in the Greek Manuscripts of the Letter of James Using Reduced Median Networks," *Computers and the Humanities* 38.1 (2004): 1–14.

While the matter of contamination has cast a shadow over texttype theory,[10] texttypes have not been rejected universally.[11] Additionally, the assumptions of other methods introduce limitations that texttype-based methods do not face. Perhaps most importantly, the prudent reduction of witnesses and readings to genealogically-significant groups may be necessary to make genealogical approaches more tractable and effective.

In what follows, I will present *non-negative matrix factorization* (NMF) as a simple, fast, and fully-automated method for classifying MSS and readings simultaneously. It is pre-genealogical, in the sense that it does not infer any prior–posterior relationships among readings or texts. As such, it is intended, not to replace genealogical methods, but to assist them.[12] In the first section that follows, I introduce some basic concepts behind how a broader class of methods, including NMF, approaches the classification problem and how NMF, in particular, classifies both readings and MSS in the presence of contamination. In the section after that, I describe my application of NMF to a full collation of the Epistle of Jude. In the last section, I show that NMF yields intuitive results that correspond to human classifications in existing literature. Finally, I conclude with a brief discussion of NMF's promise in more involved applications.[13]

Theoretical Basis

To describe the methodology behind NMF, a useful place to start is with a similar, but slightly broader, method known as *factor analysis*. Factor analysis has enjoyed much recent attention in NT text-critical studies, seeing extensive development and use at Andrews University in particular.[14] A comparison

[10] Klaus Wachtel, "Towards a Redefinition of External Criteria: The Role of Coherence in Assessing the Origin of Variants," in *Textual Variation: Theological and Social Tendencies?* ed. H.A.G. Houghton and David C. Parker, Texts and Studies 6 (Piscataway, NJ: Gorgias, 2008), 109–129, esp. 114.

[11] For a defense of its continued value, see Epp., "Textual Clusters."

[12] For more on this application, see the Conclusions section.

[13] The author would like to thank Stephen L. Brown for his feedback on this paper at every stage of its development, the referees for their thorough remarks and suggestions on the initially-submitted draft, and Brent Niedergall and Duncan Johnson for their comments on the second draft.

[14] A brief summary and assessment can be found in Thorpe, "Multivariate Statistical Analysis," 43–46. For a more comprehensive introduction, see Clinton S. Baldwin, "Factor Analysis: A New Method for Classifying New Testament Greek Manuscripts," *AUSS* 48.1 (2010): 29–53. For more specific applications to NT books and corpora, see Kenneth Keumsang Yoo, "The Classification of the Greek Manuscripts of 1 Peter with Special Emphasis on Methodology" (PhD diss., Andrews University, 2001) and Clinton S. Baldwin, "The So-Called Mixed Text: An Examination of the Non-Alexandrian and Non-Byzantine Text-Type in the Catholic Epistles"

of the two methods will provide some context for the underlying theory and advantages of NMF.

Factor analysis and NMF both rely on the same basic concepts to model and solve the problem of classifying MSS and readings. One key element is the *reading profile*, which I will define simply as *a set of readings from the collation with numerical weights assigned to them*. In factor analysis, these are called the *factors*. Intuitively, a reading's weight in a profile conveys how that reading is correlated to the group associated with that profile. Reading profiles in this context can be viewed as augmented forms of the group profiles used in the CPM. A specific advantage to this modification, as I will discuss shortly, is that the assignment of numerical weights to readings provides us with a mechanism of *combining* profiles in different ways. We can "mix" two reading profiles by adding the weights of their corresponding readings; we can "subtract" one profile from another by subtracting the weights of their corresponding readings, and we can "scale" a reading profile by multiplying all of its readings' weights by the same scaling factor. [15]

Factor analysis and NMF attempt to approximate every MS's text (i.e., its pattern of readings) using combinations of a small number of profiles, in which the profiles themselves are assigned weights to indicate how much textual material they contribute to the MS being approximated. The MSS that are predominantly described by the same profile can be understood as belonging to the cluster associated with it. Factor analysis and NMF iteratively adjust the weights of the readings in the profiles to ensure that the MSS' texts are covered as closely as possible and different clusters overlap as little as possible.

The main shortcoming of factor analysis is that in the presence of negative weights, its outputs become difficult to interpret. How exactly does a negatively-weighted reading relate to a group profile? What if a MS's text is approximated by a combination of profiles in which one profile is subtracted from another? What kind of contamination would this describe, if it can be said to describe contamination at all?

Non-negative matrix factorization, as its name suggests, avoids these ambiguities by adding the constraint that none of the weights assigned to readings or profiles can be negative. This change allows us to see combinations of readings or reading profiles as "sums of parts" or "mixtures of

(PhD diss., Andrews University, 2007).

[15] In the parlance of linear algebra, the mathematical expressions for these descriptions are called *linear combinations*. For example, in a collation with three variant readings r_1, r_2, and r_3, we would express the reading profile for cluster 1 as $F_1 = a_1 r_1 + a_2 r_2 + a_3 r_3$. The coefficients a_1, a_2, and a_3 are the weights assigned to the readings; they can be positive, negative, or zero. Meanwhile, if MS m_1 can be approximated using reading profiles 4 and 5, the corresponding expression is $m_1 \approx b_4 F_4 + b_5 F_5$, where b_4 and b_5 are weights assigned to the reading profiles.

components," which greatly facilitates the interpretation of outputs where contamination is involved.

As a consequence of its "sum of parts" model, NMF is also well-suited to identify common textual components shared by multiple textual groups. For example, multiple clusters associated with Byzantine subfamilies might have their own reading profiles with fewer distinctive readings, while their common Byzantine readings are assigned to a separate cluster's reading profile.[16] In situations like this, NMF may shed light on hierarchical structure in the MS data, in which ancestral material is inherited by later families.

Ever since it was first popularized, NMF has been applied to a variety of fields.[17] Applications most relevant to the one under discussion include classifying documents by their topics,[18] isolating gene expressions in DNA,[19] and determining mixture in human biological ancestry.[20] While I will summarize the basic principles behind NMF, I will do so primarily in terms of the present application, without delving too much into technical details.[21]

---

[16] The textual critic interpreting the cluster's output by NMF must therefore take care to distinguish between cases of shared ancestry and true instances of contamination. This is typically easy to spot: clusters representing common readings will not have "pure" representative MSS, but will instead share their most representative MSS with other clusters.

[17] See Suvrit Sra and Inderjit S. Dhillon, *Nonnegative Matrix Approximation: Algorithms and Applications*, technical report prepared for the Department of Computer Science, University of Texas at Austin (2006) for a detailed survey.

[18] Wei Xu, Xin Liu, and Yihong Gong, "Document Clustering Based on Non-negative Matrix Factorization," in Proceedings *of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York: ACM, 2003), 267–273.

[19] See Jean-Philippe Brunet et al., "Metagenes and Molecular Pattern Discovery Using Matrix Factorization," *Proceedings of the National Academy of Sciences of the United States of America* 101.12 (2004): 4164–4169 and Karthik Devarajan, "Nonnegative Matrix Factorization: An Analytical and Interpretive Tool in Computational Biology," *PLoS Computational Biology* 4.7 (2008): 1–12.

[20] Eric Frichot et al., "Fast and Efficient Estimation of Individual Ancestry Coefficients," *Genetics* 196.4 (2014): 973–983.

[21] For an accessible introduction, see Daniel D. Lee and H. Sebastian Seung, "Learning the Parts of Objects by Non-negative Matrix Factorization," *Nature* 401 (1999): 788–791. For a more technical overview of the software implementation of NMF used for this project, see Marinka Žitnik and Blaž Zupan, "NIMFA: A Python Library for Nonnegative Matrix Factorization," *Journal of Machine Learning Research* 13 (2012): 849–853. For a mathematical description of the specific methods used in our implementation of NMF, see Chih-Jen Lin, "Projected Gradient Methods for Nonnegative Matrix Factorization," *Neural Computation* 19.10 (2007): 2756–2779.

Our text-critical application at hand lends itself well to NMF, as one of the most natural ways to think of a collation of MSS would be as different readings in a data table, or *matrix*: each column representing a MS, and each row representing a variant reading.[22] If a given reading were found in a given MS, then the entry in the corresponding row and column would be 1; otherwise, it would be 0.[23] For future reference, I will designate the number of readings (i.e., rows) as m and the number of MSS (i.e., columns) as n, and I will describe the resulting table as an $m \times n$ matrix (see Table 1.).

Table 1. Matrix Representation of Part of a Collation[a]

	03	35	88	1505	1739
Jude 1:4/45–58, δεσποτην και κυριον ημων ιησουν χριστον	1	0	0	0	1
Jude 1:4/45–58, δεσποτην θεον και κυριον ημων ιησουν χριστον	0	1	0	1	0
Jude 1:4/45–58, δεσποτην θεον και κυριον ιησουν χριστον	0	0	1	0	0
Jude 1:13/8, απαφριζοντα	0	0	0	1	1

[22] For the purposes of this study, I do not encode data at the granularity level of *variation units*, or collections of exclusive variant readings at a location in the text. While we conventionally would include the index of a variation unit containing a given reading in that reading's row label (e.g., u4-8r3, "unit 4 through 8, reading 3"), this would not affect how the data is processed. The distinction between readings in the same variation unit is maintained by the constraint that in a given MS (i.e., column), at most one reading (i.e., row) in each variation unit can have a value of 1.

[23] It should be clarified that a scribal omission of any text at a variation unit also counts as a "reading," and so an omission at a variation unit will also label a row in the collation matrix. Meanwhile, *lacunae* (gaps of content caused by missing portions of a page or other damage) do not receive rows in the matrix, as they do not represent any reading copied by a scribe.

There is more than one way to encode lacunae and uncertain readings. One is to set the cells to 0 for all readings in variation units where a given MS is lacunose. Another is to set these cells with fractional values so that the values for all readings in each lacunose variation unit add to 1, the intuition being that each fractional value represents the probability of a given reading having been present. The latter approach is more robust, as the choice of coefficients can be tailored for specific situations (e.g., if a reading is ambiguous but can be narrowed down to a subset of the available readings, or if the space taken up by a lacuna rules out some readings, but not others). However, for this paper, I chose to take the former approach, as it is simpler and more suited to showing the power of NMF in the absence of human intervention.

While NMF can be applied to MSS with any number of lacunae, highly fragmentary witnesses tend to contribute more "noise" than information. In the appendix, I will show how to classify these types of witnesses in post-processing.

Jude 1:13/8, επαφριζοντα	1	1	1	0	0
Jude 1:16/14–16, επιθυμιας εαυτων	0	1	0	0	1
Jude 1:16/14–16, επιθυμιας αυτων	1	0	1	1	0
Jude 1:25/3, omit	1	0	1	1	1
Jude 1:25/3, σοφω	0	1	0	0	0

[a]Variation units have alternately-highlighted rows. Variants readings, including the variation unit indices, appear as the row labels, and MS IDs appear as column labels.

To run NMF on an input collation matrix, we must specify the number of clusters (e.g., texttypes or families) that we want to infer using the collation data.[24] Throughout this paper, I will designate this number *k*. A small choice of *k* will produce coarse groupings (e.g., for *k* = 3, the clusters will likely be "Byzantine," "Alexandrian," and "everything else"), while a larger choice will yield finer and more accurate groupings. The textual critic using NMF must decide on an agreeable compromise between succinctness and accuracy when setting this parameter: too low a choice for *k* will oversimplify and fail to capture the MS data accurately, while too high a choice will make the resulting textual groupings less succinct and more complex.

The output of NMF is two smaller matrices, which I will identify by the conventional shorthand *W* and *H*.[25] The first matrix *W* is called the *basis* matrix, and it describes the relationships between readings and the reading profiles of inferred textual clusters. It has *m* rows for the variant readings in the collation and *k* columns for the group reading profiles. A reading with a higher weight in a profile can be viewed as more representative of that profile's group than other readings. The second matrix *H* is called the *mixture* matrix, and it represents the relationship between MSS and textual clusters. It has *k* rows for the underlying textual clusters and *n* columns for the MSS in the collation. The values in this matrix tell us which clusters' reading profiles, when combined, best approximate the set of readings found in each MS. A MS with a high mixture weight from one cluster can be viewed as a pure representative of that cluster, while a MS with lower weights, spread across multiple clusters, can be viewed as a witness with mixed textual components (see Table 2.).

[24] For exploratory tasks, we are unlikely to know this number ahead of time. For details on how to determine the best one, see section entitled "Classification of Readings."

[25] As its name suggests, NMF factors the original collation matrix into the matrix product of *W* and *H*. The product of the two matrices captures the process described in section 2: it approximates the MS collation data using only weighted combinations of group profiles of readings.

Table 2. NMF Output Matrices for the Collation data in Table 1 for $k = 3$ Clusters[a]

	Profile 1	Profile 2	Profile 3
Jude 1:4/45–58, δεσποτην και κυριον ημων ιησουν χριστον	0.2473	0.0000	0.6385
Jude 1:4/45–58, δεσποτην θεον και κυριον ημων ιησουν χριστον	0.0520	0.7389	0.4017
Jude 1:4/45–58, δεσποτην θεον και κυριον ιησουν χριστον	0.4057	0.0000	0.0000
Jude 1:13/8, απαφριζοντα	0.0000	0.0000	1.0164
Jude 1:13/8, επαφριζοντα	0.7549	0.7389	0.0000
Jude 1:16/14–16, επιθυμιας εαυτων	0.0000	0.7389	0.5996
Jude 1:16/14–16, επιθυμιας αυτων	0.8251	0.0000	0.2881
Jude 1:25/3, omit	0.7247	0.0000	0.9168
Jude 1:25/3, σοφω	0.0000	0.7389	0.0000

	03	35	88	1505	1739
Profile 1	1.1638	0.0000	1.3523	0.3871	0.0000
Profile 2	0.0000	1.3535	0.0000	0.0000	0.0000
Profile 3	0.2029	0.0000	0.0000	0.7641	1.0992

[a]The top matrix (the *basis* matrix *W*) contains the weights of readings in each group's reading profile, with higher weights indicating precedence within the profile. The bottom matrix (the *mixture* matrix *H*) shows the makeup of each MS in terms of weighted contributions from different groups. In this example, MSS 03 and 88 are almost purely represented by Profile 1, as they share its most characteristic readings—ἐπαφρίζοντα at Jude 1:13/8, ἐπιθυμίας αὐτῶν at Jude 1:16/14–16, and omission at Jude 1:25/3.

How does NMF assign accurate weights to readings in its profiles (matrix *W*) and to profiles with a mixture that models the texts of MSS (matrix *H*)? In a nutshell, it starts with "guesses" for the weights in one matrix and then uses these to find the best weights for the other matrix. We can get a more practical idea of this by observing how a chain of more traditional methods is typically applied. To start, suppose we make an initial "guess" for the mixture matrix *H* by assigning MSS to clusters according to a simple approach like the quantitative method. This initial guess for *H* will not be particularly accurate, primarily because of its hard assignment of MSS to different clusters, with no mixture. But then suppose we use the CPM to determine group profiles of readings. Using the initial group assignments we just made in *H*, we can determine which readings are more or less representative of each cluster (based on agreement among the MSS in each cluster) and adjust the weights of these readings appropriately in *W*. At this point in the CPM, we

would refine our classifications of the MSS in *H* using the new weights for representative readings in *W*: MSS with more representative readings would arise as purer representatives of certain clusters, and MSS with readings from different groups would be recognized as mixed.

The guiding principle of NMF is that once accurate weights are known for one matrix, they can be used to refine the weights in the other. In other words, NMF uses the circular relationship between characteristic MSS and characteristic readings to its advantage. It adjusts weights for readings in their group profiles and weights for group mixture in MSS so that the original collation data can be estimated as accurately as possible using combinations of the inferred group reading profiles. Speaking in terms of existing methodology, we could say that it continually iterates the steps of the CPM, re-weighting characteristic group readings in their profiles based on the weights of the group MSS that attest to them, and then vice-versa, until the results no longer increase in accuracy.[26] This approach of iterative refinement is so powerful that, even if the initial guesses for the weights of *W* or *H* are completely random, NMF will typically climb up to a reasonable choice of weights before it can no longer improve them.

Application

Data

We applied NMF to Tommy Wasserman's comprehensive collation of Jude.[27] I considered this a good testing ground for several reasons. First, the size of the collation, which might be prohibitive for more complex, human-supervised methods, can be handled efficiently and automatically by NMF. Second, the collation covers virtually all readings and MSS.[28] We can, therefore, avoid

[26] One may wonder if the process thus described can get caught in an infinite loop. It turns out that this is impossible; for a mathematical proof of this, see L. Grippo and S. Sciandrone, "On the Convergence of the Block Nonlinear Gauss-Seidel Method under Convex Constraints," *Operations Research Letters* 26 (2000): 127–136.

It should be noted that while NMF will always reach a stopping point, the choice of weights it ends up with may not result in the most accurate approximation of the collation data. To find the set of weights that achieves the highest possible accuracy, NMF may need to be restarted many times with different starting points; see section entitled "How Many Groups?" for more details.

[27] Tommy Wasserman, *The Epistle of Jude: Its Text and Transmission*, ConBNT 43 (Stockholm: Almqvist & Wiksell International, 2006); for the digital dataset, see Tommy Wasserman, "Transcription of the Manuscripts Containing the New Testament Letter of Jude," 2012, http://dx.doi.org/10.17026/dans-xcz-cqbr.

[28] Wasserman notes that his apparatus does not record the most frequent orthographic variants, such as instances of movable *nu*, final vowel elisions in prepositions and conjunctions, cases of itacism, and other common vowel interchanges (*The Epistle of Jude*, 129–130). This is actually good for our purposes, since such readings are

any existing biases associated with previous selections of "significant" readings and MSS, in order to verify whether NMF will come to the same conclusions independently. Third, to the best of my knowledge, no other application of this scale has been done with Wasserman's data. It is hoped that my work will spark continued research involving Wasserman's collation and encourage collations of equal scale elsewhere in the NT.[29]

Wasserman's collation covers 560 MSS, including 3 papyri and 38 lectionaries,[30] across 360 variation units. I encoded all unambiguous readings as described in section 2. The result was a 1346 × 560 matrix with 178,887 non-zero entries.

Because NMF attempts to partition the collation data into underlying groups that can be added and mixed together, highly lacunose MSS can negatively influence the process. To account for this, I treated all MSS with fewer than 300 readings as fragmentary and postponed their classification to a later step.[31] Filtering these out, we are left with a 1346 × 518 matrix with 172,932 non-zero entries. The excluded MSS and their classifications are listed in the appendix.

How Many Groups?

A natural question to arise from this process would be how many clusters NMF should fit to the data. The process of determining the right number is called *rank estimation*, and one of the most popular metrics used in this process is called the *cophenetic correlation coefficient*.[32] In terms of my application, this value measures the frequency with which NMF assigns the same MSS to the same groups over many runs with random initial choices of weights. If NMF's navigation of the solution space always leads it to the same solution or

polygenetic and are typically considered unimportant for MS classification (W. Larry Richards, *The Classification of the Greek Manuscripts of the Johannine Epistles* [Missoula, MT: Scholars Press, 1977], 27–28).

[29] For other such collations, see Michael Bruce Morrill, "A Complete Collation and Analysis of All Greek Manuscripts of John 18" (PhD. diss., University of Birmingham, 2012) and Matthew S. Solomon, "The Textual History of Philemon" (PhD diss., New Orleans Baptist Theological Seminary, 2014).

[30] This figure excludes correctors' hands, alternate readings, and commentary readings.

[31] I chose a threshold of 300 as a simple compromise to achieve sufficient information on readings for classification purposes and to avoid setting aside too many MSS for classification later.

[32] I will not elaborate on the technical details of this metric here. See J. P. Brunet et al., "Metagenes and Molecular Pattern Discovery Using Matrix Factorization," *Proceedings of the National Academy of* Sciences *of the United States in America* 101.12 (2004): 4164–4169 for an introduction.

small set of solutions (in which case the coefficient will be high), then we can have higher confidence that there is an underlying structure to the data that is accurately captured by k clusters. After repeating this process for all values of k we are interested in, the rule of thumb is to "select values of k where the magnitude of the cophenetic correlation coefficient begins to fall."[33] For data with a hierarchical structure, such as MSS with different tiers of common ancestry, multiple such values of k may be suitable for uncovering branches of the text at different granularities (e.g., several Byzantine subfamilies might emerge from what was previously a broadly Byzantine cluster).

Beside the cophenetic correlation coefficient, other factors may influence the decision of how many clusters are best. One is the *sparsity* of the matrices W and H; higher sparsity in the output matrix W (respectively, H) basically means that fewer readings (respectively, MSS) are assigned high weights for each group in each column (respectively, row), or, put more simply, that the groups have less overlap.[34] Other factors include how accurately W and H approximate the original data set and whether the choice of k clusters achieves an agreeable balance between detail and succinctness.

Implementation

For reasons of space, I will not detail our software implementation of NMF, nor the specifications of our hardware here. However, for those interested in reusing or adapting the code for similar work, I have made the collation dataset, code, and implementation details available for free at https://github.com/jjmccollum/jude-nmf.

Results

Table 3 gives summary statistics for the rank estimation and factorization results for $2 \leq k \leq 30$. In general, NMF isolated groups that explained over 95% of the variance in the observed data (in this case, readings in MSS), and it did so in about 2.5 minutes, on average.

[33] Brunet, "Metagenes and Molecular Pattern," 4165.

[34] For more technical detail, see Patrik O. Hoyer, "Non-negative Matrix Factorization with Sparseness Constraints," *Journal of Machine Learning Research* 5 (2004): ed. Peter Dayan, 1457–1469.

Table 3: Summary Statistics for NMF Results[a]

k	TIME	COPH	DIST	EVAR	*W*.SPAR	*H*.SPAR
2	0.7311	0.9970	9220.0095	0.9467	0.4967	0.7051
3	3.2303	0.9363	8741.7228	0.9494	0.4993	0.6385
4	4.2654	0.9335	8383.2904	0.9515	0.5015	0.6406
5	12.3092	0.9266	8127.9438	0.9530	0.5011	0.6594
6	17.0137	0.9381	7896.8484	0.9543	0.5019	0.6783
7	24.6415	0.9321	7694.5026	0.9555	0.5025	0.6862
8	28.7929	0.9311	7491.7511	0.9567	0.5020	0.7131
9	47.2967	0.9277	7336.0094	0.9576	0.5026	0.7250
10	44.2114	0.9314	7216.6958	0.9583	0.5121	0.6542
11	77.2047	0.9355	7060.6949	0.9592	0.5026	0.7365
12	77.1497	0.9354	6886.0044	0.9602	0.5038	0.7586
13	113.6054	0.9400	6761.8625	0.9609	0.5035	0.7682
14	139.0699	0.9343	6671.0795	0.9614	0.5238	0.7006
15	157.3397	0.9303	6567.2208	0.9620	0.5283	0.6976
16	133.7074	0.9268	6445.4954	0.9627	0.5319	0.7033
17	226.0826	0.9323	6380.0739	0.9631	0.5391	0.6682
18	302.9293	0.9259	6251.1086	0.9639	0.5272	0.7222
19	372.9483	0.9300	6176.3545	0.9643	0.5488	0.6816
20	291.5081	0.9304	6111.1190	0.9647	0.5408	0.6927
21	177.2737	0.9359	6021.3340	0.9652	0.5370	0.7007
22	300.1050	0.9356	5931.9671	0.9657	0.5435	0.7084
23	193.2883	0.9385	5845.8003	0.9662	0.5422	0.7127
24	237.4277	0.9410	5758.9796	0.9667	0.5435	0.7176
25	269.2819	0.9427	5708.4434	0.9670	0.5594	0.6736
26	274.7705	0.9428	5614.3632	0.9675	0.5487	0.7155
27	225.5836	0.9455	5536.6625	0.9680	0.5540	0.6941
28	216.4106	0.9483	5452.1027	0.9685	0.5718	0.6921
29	182.8073	0.9494	5385.8597	0.9689	0.5717	0.6917
30	179.8480	0.9520	5328.1415	0.9692	0.5695	0.6978

[a]Here, k indicates the rank (i.e., number of clusters) of the NMF run, TIME gives the running time in seconds for the best NMF run, COPH gives the value of the cophenetic correlation coefficient (see section "How Many Groups?" for details), DIST gives the error of NMF's approximation of the collation data, EVAR gives the proportion of explained variance, and *W*.SPAR and *H*.SPAR measure the sparseness of the output matrices *W* and *H*, respectively. Best ranks, according to the cophenetic correlation coefficient rule of thumb, are highlighted.

The numbers of clusters that provide the best fit, according to the rule of thumb, are 2, 6, 11, 13, 17, and 21. Because the factorization for 13 clusters had the highest *H* sparsity (i.e., best separation between MS groups), I chose to examine the NMF results for this number of groups in detail.

Classification of Manuscripts

In order to describe the textual groups represented by the clusters, it is instructive to look at their most representative readings and witnesses. In what follows, all MS numbers follow the Gregory-Aland numbering system.[35]

Cluster 1 appears to represent a large subfamily of the Byzantine texttype.[36] Its strongest representative is the tenth-century MS 920, which assigns this cluster a weight of 3.7179. Other strong tenth-century representatives include MSS 1871 (3.3434), 605 (1.7590), 1880 (1.2326), and 82 (1.1676). The only older cluster member is the ninth-century MS 1841 (1.7781). Notably, all of these older MSS, with the possible exception of 920, have nontrivial mixture contributions from cluster 11, which contains more familiar and probably older Byzantine witnesses. Apart from this, we do not recognize this specific family in the literature. Given its common, undistinctive readings and its size, cluster 1 is best described as a "general Byzantine" cluster. I will therefore designate it as "K."

Cluster 2 represents f^{1739}, a well-known textual family.[37] NMF identified the following MSS as members of this cluster: 323 (with weight 2.8466 for this cluster), 1241 (2.8002), 322 (2.7957), 1739 (2.8466), 1881 (2.5119), 2298 (2.4219), and 6 (1.7270).[38] This group has been identified independently in 1 Peter[39] and 2 Peter,[40] and in the General Epistles it has been shown to share important readings with the old Georgian versions.[41] Its namesake

[35] Kurt Aland et al., *Kurzgefasste Liste der griechischen Handschriften des Neuen Testaments*, ANT 1 (Berlin: de Gruyter, 1994). The *Liste* can be consulted online at http://ntvmr.uni-muenster.de/liste.

[36] We can conclude that a cluster contains a MS if that MS's largest mixture contribution comes from that cluster.

[37] See Thomas C. Geer, Jr., Family 1739 in the Book of Acts, SBLMS 48 (Atlanta: Scholars Press, 1994) and Günther Zuntz, The Text of the Epistles: A Disquisition upon the Corpus Paulinum, Schweich Lectures of 1946 (London: British Academy, 1953).

[38] While NMF classifies majuscule 04 / C as Alexandrian (cluster 7) in Jude, it also shows it to have strong mixture (0.7604) with this cluster.

[39] See Yoo, "Classification," 112–116, who classifies majuscule 04 / C as belonging to this group in 1 Peter.

[40] Terry Dwain Robertson, "Relationships among the Non-Byzantine Manuscripts of 2 Peter," *AUSS* 39.1 (2001): 41–59, esp. 45–47.

[41] Christian-B. Amphoux and Dom B. Outtier, "Les Leçons des Versions Géorgi-

is a consistently-cited witness in NA[28]. Scholars have conjectured that its exemplar dates back as far as the fourth century.[42] Further evidence for the family's antiquity has been found in its close similarity to the text used by Origen.[43] The connection with Origen has led some to posit that *f*[1739] represents the controversial "Caesarean" texttype in the General Epistles.[44] While the cluster is small, its members are remarkably cohesive, with the top three witnesses showing almost no mixture with any other cluster.

Cluster 3 represents the group of lectionaries. The existence of a distinct lectionary textual group has long been recognized,[45] but a thorough examination of this group in the General Epistles was delayed for some time. The first and perhaps most extensive work in this area was done by Junack.[46] Junack's work confirmed the existence of a large and cohesive textual family among the Byzantine lectionaries. At least in the context of Jude, our results, based on Wasserman's complete collation, should give additional weight to these findings. Our results also agree with Junack's identification of *l*596 as an exceptionally non-Byzantine lectionary; NMF classified this MS as a strong representative of the Alexandrian cluster (7), with a weight of 1.3997 for that cluster. This cluster also contains non-lectionary MSS, though they are lower on the list due to mixture.

Cluster 4 is the majority subgroup $\mathbf{K}^{r}$, also known as *f*[35], as can be seen from the overlap between the top MSS in the mixture matrix and the list of collated MSS for 2 John–Jude in Pickering's edition.[47] This cluster is by far

ennes de l'Épître de Jacques," *Bib* 65.3 (1984): 365–376, esp. 372–373.

[42] Thomas C. Geer, Jr, "Codex 1739 in Acts and Its Relationship to Manuscripts 945 and 1891," *Bib* 69.1 (1988): 27–46, esp. 27.

[43] K. W. Kim, "Codices 1582, 1739, and Origen," *JBL* 69.2 (1950): 167–175, esp. 168–170. While the strongest connection between 1739 and Origen appears in Romans, notes in Jas 2:13 and 1 John 4:3 indicate a similar relationship in the General Epistles. In his conclusion, Kim goes on to suggest that GA 1582, a copy of the gospels apparently written by the same scribe as 1739 and also sharing many readings with Origen, was originally part of the same codex as 1739 (Kim, "Codices," 175). For additional discussion on 1582, see Amy S. Anderson, *The Textual Tradition of the Gospels: Family 1 in Matthew*, NTTST 32 (Leiden: Brill, 2004).

[44] Amphoux and Outtier, "Versions Géorgiennes," 374–375.

[45] Ernest Cadman Colwell, "Is There a Lectionary Text of the Gospels?" *HTR* 25.1 (1932): 73–84.

[46] Klaus Junack, "Zu den griechischen Lektionaren und ihrer Überlieferung der Katholischen Briefe," in Die alten Übersetzungen des Neuen Testaments: die Kirchenväterzitate und Lektionare: der gegenwärtige Stand ihrer Erforschung und ihre Bedeutung für die griechische Textgeschichte, ed. Kurt Aland, ANT 5 (Berlin: de Gruyter, 1972), 498–591.

[47] *The Greek New Testament According to Family* 35, ed. Wilbur N. Pickering, 2nd

the largest, and it exhibits strong agreement among its purest representatives. However, despite this agreement, its only witnesses predating the eleventh century are the tenth-century MSS 457 (with a moderate weight of 0.8178 for this cluster), 1891 (0.7658), and 450 (0.4717). One possible reason for this is that the family originated later in the history of NT transmission. It has been suggested that it was "produced out of the K^x type with lectionary and liturgical interests in mind."[48] Of course, even if this is the case, the family surely predates the tenth century. Indeed, it just falls short of dominating the makeup of the ninth-century majuscule 020 / L^{ap}, which NMF assigned a K^r mixture weight of 0.4812 and an Alexandrian mixture weight of 0.4824.

Cluster 5 corresponds to f^{2138}. The group is small, and its leading representatives are the following: 1505 (weight 2.3894 for this cluster), 2495 (2.3642), 1611 (2.2124), 1292 (2.1500), 630 (1.9693), and 2200 (1.8282). These first six MSS consistently have small but noticeable mixture components from cluster 7 (Alexandrian), while five other MSS have largely Byzantine affinities and the remaining two have very strong mixtures? with cluster 6 (f^{453}). These subgroups of witnesses may represent localized branches of the family or different stages of its development. The f^{2138} group has been identified specifically in Jude through factor analysis,[49] and in the General Epistles, its core members have been shown to have a connection to the Harklean Syriac version.[50]

Cluster 6 undoubtedly represents f^{453}, another recognized group.[51] The earliest of its witnesses is the tenth-century MS 307 (weight 2.2161 for this cluster). Other notable group members include 321 (2.2676), 918 (2.2268), 453 (2.2054), 2197 (2.1783), and 2818 (2.0642). The aforementioned MSS, including 307, are all pure representatives of the group, with virtually no mixture from other clusters.[52]

---

ed. (Wilbur N. Pickering, 2015), 722.

[48] *The New Testament in the Original Greek: Byzantine Textform*, ed. Maurice A. Robinson and William G. Pierpont (Southborough, MA: Chilton, 2005), 557.

[49] It corresponds to group 3 in Baldwin, "The So-Called Mixed Text," 106.

[50] See Christian-B. Amphoux, "La Parenté Textuelle du sy^h et du Groupe 2138 dans l'Épître de Jacques," *Bib* 62.2 (1981): 259–271; Barbara Aland and Andreas Karl Juckel, *Das Neue Testament in syrischer* Überlieferung, vol. 1 ANTF 7 (Berlin: de Gruyter, 1986); and Matthew Spencer, 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).

[51] Spencer, Wachtel, and Howe, "Greek Vorlage."

[52] This group was independently identified in the General Epistles through stemmatic methods by Spencer, Wachtel, and Howe, who noted that it "contains states of text that are thought to be important for the formation of the Byzantine text" (Spencer, Wachtel, and Howe, "Greek Vorlage").

Cluster 7 is clearly Alexandrian. Not surprisingly, its top representatives are 03 / B (weight 1.7075 for this cluster), 𝔓[72] (1.6416), 81 (1.5978), 5 (1.5827), 326 (1.5766), and 33 (1.5653). Majuscules 01 / 044 ,(1.3291) א / Ψ (1.3074), 02 / A (1.3013), and 04 / C (0.8790) also fall under this cluster, but as the other columns of the mixture matrix show, these MSS also share some elements with other clusters.

Cluster 8 is von Soden's K^c Byzantine subgroup,[53] as can be seen from the presence of the following K^c MSS in the cluster: 390 (mixture weight 2.0269 for this cluster), 912 (1.9854), 234 (1.9735), 2085 (1.9573), 1753 (1.8504), 42 (1.8063), 996 (1.7051), 1594 (1.6357), 1405 (1.5897), 51 (1.3048), and 223 (1.2764).[54] The cluster as established by NMF has no witnesses from earlier than the tenth century, and of its purest representatives, the oldest is the eleventh-century MS 42.

Cluster 9 appears to represent a "commentary" text group. Of its strongest witnesses, the top MS, 606 (mixture weight 2.0461 for this cluster), belongs to von Soden's Οθδ group, with Pseudo-Oecumenius' commentary on Acts and the General Epistles and Theodoret's commentary on the Pauline epistles; MSS 454 (2.0119), 641 (2.0045), 103 (1.9162), 314 (1.6596), 250 (1.5903), 1862 (1.5384), and 327 (1.4548) belong to the O group, having only Pseudo-Oecumenius's commentary; MS 018 / K^{ap} (1.3648) belongs to the $A^{\pi\rho}$ group, with Andreas the Presbyter's commentary on Acts and the General Epistles.[55] The non-commentary MSS in the cluster could either represent copies of only the text from the commentary, or the text on which the commentary was based. The group appears to be a relatively old Byzantine group, with ninth-century MSS 1862 and 018 appearing as prominent representatives. As it lacks an existing siglum, I will designate it *Comm*.

Cluster 10 represents a particularly "Alexandrian" branch of the Byzantine texttype. Three notable MSS—the minuscule 1066 (weight 1.5542 for this cluster) and the closely-related majuscules 0142 (1.2691) and 056 (1.1165), all of which contain the Pseudo-Oecumenius commentary—are tenth-century witnesses to the text of this cluster. The text itself shares several Alexandrian readings, which implies that the text at least incorporated elements from an ancient tradition. In addition, the strongest representatives of the cluster, 1563 (1.9423), 1718 (1.8537), 1425 (1.8438), and 1359

[53] Hermann Freiherr von Soden, Die Schriften des Neuen Testaments in ihrer ältesten erreichbaren Textgestalt hergestellt auf Grund ihrer Textgeschichte, vol. 1 (Göttingen: Vandenhoeck & Ruprecht, 1911): 1761.

[54] These MSS are von Soden's δ366, α366, δ365, α465, α395, α107, δ383, δ375, α555, δ364, and α186, respectively.

[55] Robert Waltz, *The Encyclopedia of New Testament Textual Criticism*, (available online at https://books.google.com/books/about/The_Encyclopedia_of_New_Testament_Textua.html?id=pefhAAAAQBAJ), 199–200.

(1.7900), all exhibit small elements of mixture from the Alexandrian cluster. If the underlying text had ever been widespread, few of its witnesses seem to have survived, as this cluster is small. Lacking an existing siglum, I will designate it f^{0142} after its oldest member.

Cluster 11 represents another of the older Byzantine subgroups. Its relative age is attested by the presence of the ninth-century MSS 1424 (weight 1.1035 for this cluster) and 049 (1.0603), both of which contain mixture from the *Comm* cluster. The prominence of MS 1780 (1.5711) may also be an indicator of an earlier text, as 1780 belongs to the older K^a family (also known as von Soden's I^κ group or Family Π) elsewhere.[56] Similarly, MS 1175 (1.4728) is a major witness to this Byzantine subgroup, although it also contains some mixture from the K^c cluster. This adds some detail to the findings of Richards, who has shown that 1175 is Alexandrian in James–2 Peter and Byzantine in 1 John–Jude.[57] As 1424 and 1175 are consistently-cited witnesses throughout the NT in NA[28], this cluster may be of special interest to future research into the text they carry. Lacking an existing name for this group, I will refer to it as f^{1780}.

Cluster 12 contains several MSS associated with von Soden's I group. The MSS with the highest mixture weights for this cluster are 1843 (1.6896), 1869 (1.5543), 506 (1.5086), 1903 (1.4808), 489 (1.4778), 927 (1.4493), 203 (1.4455), 1868 (1.4379), 1729 (1.4229), and 1873 (1.3229). Given the moderate size of the cluster and the consistent von Soden classifications of its members, I will tentatively use von Soden's classification and label this cluster a "Western" branch of the Byzantine texttype in Jude.

Cluster 13 is a curious group consisting of just a few MSS. It appears to be closely related to the Alexandrian text, as many members of that cluster feature large mixture weights from this one. The top two MSS, 915 (weight 2.7366) and 88 (2.6297), agree on many readings in Jude. In the General Epistles, they and a few other MSS with high weights from this cluster—1846 (1.8525), 621 (0.7650), 442 (0.7624), and 1243 (0.5928)—read δι' ὕδατος καὶ πνεύματος καὶ αἵματος in 1 John 5:6. In 1 Corinthians, 88 and 915 attest to the placement of 14:34–35 at the end of the chapter, a transposi-

[56] See Silva Lake, *Family Π and the Codex Alexandrinus: The Text according to Mark*, SD 5 (London: Christophers, 1936); Jacob Geerlings, *Family Π in Luke*, SD 22 (Salt Lake City: University of Utah Press, 1962); Jacob Geerlings, *Family Π in John*, SD 23 (Salt Lake City: University of Utah Press, 1963); Russell N. Champlin, *Family Π in Matthew*, SD 24 (Salt Lake City: University of Utah Press, 1964); and Tommy Wasserman, "The Patmos Family of New Testament MSS and Its Allies in the Pericope of the Adulteress and Beyond," *TC* 7. However, while K^a / Family Π is a known family in the gospels, it does not appear to exist at all in the corpus of the General Epistles. Any relationship in Jude suggested by MS 1780, therefore, is speculative.

[57] W. Larry Richards, "Gregory 1175: Alexandrian or Byzantine in the Catholic Epistles?" *AUSS* 21.2 (1983): 155–168, esp. 157.

tion associated with Western witnesses.[58] This variant has led to much debate over whether or not these two witnesses have a common source in a localized Western text and whether or not they implicate 1 Cor 14:34–35 as an interpolation.[59] On the basis of these readings, one might conjecture that this small handful of witnesses attests to a "Western" text of Jude, but a cursory examination of its agreements and disagreements with the Latin text of Jude in the ECM[60] indicates that a strong Western connection is unlikely.[61] As this cluster seems unidentified in the literature, I will designate it f^{915} here.

There are a few observations to make here. First, NMF reveals a surprising number of Byzantine subgroups. In particular, the Byzantine texttype splits into the common group K, the lectionary group, the von Soden groups K^r and K^c, the commentary group, an Alexandrian-Byzantine group f^{0142}, an older Byzantine group f^{1780}, and a Western-Byzantine group corresponding to von Soden's I group. Based on reading patterns, the Byzantine MSS clearly do not form a monolithic group in Jude.

Second, NMF identifies smaller and subtler textual groups that are underrepresented or entirely excluded from the most popular critical apparatuses. Table **4** details the amount of representation each NMF cluster receives in the ECM's MS list and the NA^{28} consistently-cited witnesses list for Jude.[62] Naturally, the ECM, given its wider selection of data, offers a reasonable sampling from all the clusters identified by NMF, although it does noticeably favor Alexandrian witnesses. The NA^{28} apparatus in Jude clearly overrepresents the Alexandrian cluster, and while its *Byz* siglum may correctly cover Byzantine support at most variation units, it ignores much of the variety within or close to the Byzantine tradition, (*Lect*, K^r, K^c, and I) leaving the

[58] Gordon D. Fee, *The First Epistle to the Corinthians* (Grand Rapids: Eerdmans, 1987), 699.

[59] See Curt Niccum, "The Voice of the Manuscripts on the Silence of Women: The External Evidence for 1 Cor 14.34–5," *NTS* 43.2 (1997): 242–255; Philip B. Payne, "MS. 88 as Evidence for a Text without 1 Cor 14.34–5," *NTS* 44.1 (1998): 152–158; Jennifer Shack, "A Text without 1 Corinthians 14.34–35? Not according to the Manuscript Evidence," *JGRChJ* 10 (2014): 90–112; and Philip B. Payne, "Vaticanus Distigme-obelos Symbols Marking Added Text, Including 1 Corinthians 14.34–5," *NTS* 63 (2017): 604–625.

[60] *Novum Testamentum Graecum Editio Critica* Maior *IV, Catholic Letters, Part 1: Text*, ed. Barbara Aland et al., 2nd ed. (Stuttgart: Deutsche Bibelgesellschaft, 2014).

[61] In Jude, f^{915} unambiguously disagrees with the Latin tradition more often than it agrees, and the only reasonably exclusive point of agreement between the two is the reading τρόπον ἐκπορνεύσασαι in Jude 7/24-28.

[62] See *Novum Testamentum Graecum Editio Critica Maior IV, Catholic Letters, Part 2: Supplementary Material*, ed. Barbara Aland et al., 2nd ed. (Stuttgart: Deutsche Bibelgesellschaft, 2014), 9 and *Novum Testamentum Graece*, ed. Barbara Aland et al., 28th ed. (Stuttgart: Deutsche Bibelgesellschaft, 2012), 66*.

reader uninformed when there are disagreements within the tradition, with the only information offered being the *Byz*^{pt} siglum. Given the precedent of human classifications of MSS before NA^{28}, this data highlights the need for tools like NMF in witness selection for critical editions.

Table 4. Distribution of ECM and NA^{28} Consistently-cited Witnesses in Jude among Clusters Identified by NMF[a]

Cluster ID	MSS	ECM Witnesses	NA28 Witnesses
K	102	11	1
f^{1739}	7	6	1
Lect	39	7	0
K^{r}	143	8	0
f^{2138}	13	10	2
f^{453}	18	11	1
Alex	35	34	14
K^{c}	23	2	0
Comm	25	2	0
f^{0142}	10	7	1
f^{1780}	50	9	1
I	45	7	0
f^{915}	8	8	1

[a]Witnesses which are too lacunose to be included for NMF are excluded, as is the *Byz* siglum.

Third, if we cross-reference our results with Wasserman's collation, we see that NMF assigns higher weights to more evenly-divided readings than it does to rarer readings exclusive to groups. This is to be expected, as NMF aims to minimize the number of misclassified readings.[63] It also dovetails with NMF's isolation of Byzantine subfamilies, which are better distinguished by patterns of readings than by individual readings. For this reason, a reading with a high weight may represent multiple clusters, and patterns of more common readings may identify clusters better than group-exclusive readings. While this approach may not cluster readings as sparsely as we would like, it can help us identify potentially-early divisions in the scribal tradition, helping us to determine where different families side in these splits. I will address

[63] Of course, we can encourage NMF to isolate more characteristic group readings by weighting readings or variation units in the collation matrix according to their genealogical significance, but since my focus in this paper is on the use of NMF as a tool for pre-genealogical analysis, I will restrict this discussion to this note.

the variation units containing the most characteristic group readings in the following section.

Classification of Readings

In what follows, I will use Wasserman's division of variation units to reference the readings in question. Support for readings will be denoted by the group sigla introduced in the previous section. If a cluster has a reading profile with an assigned weight at least twice the value of its weight for any other reading in the variation unit, I consider the cluster to support a given reading. If the cluster does not have a high enough weight for any one reading, then it will be classified as being split between the readings with the highest weights.

Variation Unit: Jude 1:1/4–8

Table 5. Jude 1:1/4–8

Variants	Witnesses
Ἰησοῦ Χριστοῦ δοῦλος	f^{1739}, K^{r}, f^{2138}, f^{453}, *Alex*, *Comm*^{pt}, f^{0142}, I, f^{915}
Χριστοῦ Ἰησοῦ δοῦλος	K, *Lect*, K^{c}, *Comm*^{pt}, f^{1780}

An application of the CPM found that this variation unit contained a primary reading for one group identified by factor analysis in Variation Unit: Jude.[64] The transposition that occurs here is a common one throughout the Pauline and General Epistles. The order Ἰησοῦ Χριστοῦ has the earliest and most diverse support, while Χριστοῦ Ἰησοῦ finds its support among families that have close ties to the Byzantine text. The Robinson-Pierpont edition (RP) is probably correct in adopting Ἰησοῦ Χριστοῦ for its text, but the Byzantine support for Χριστοῦ Ἰησοῦ might merit it a place in the margin of that edition.

Variation Unit: Jude 1:4/48–58

Table 6. Jude 1:4/48–58

Variants	Witnesses
δεσπότην καὶ κύριον ἡμῶν Ἰησοῦν Χριστὸν	f^{1739}, *Lect*, f^{453}, *Alex*^{pt}, f^{0142}
δεσπότην θεὸν καὶ κύριον ἡμῶν Ἰησοῦν Χριστὸν	K, K^{r}, f^{2138}, *Alex*^{pt}, *Comm*^{pt}, f^{1780}, I
θεὸν καὶ δεσπότην τὸν κύριον ἡμῶν Ἰησοῦν Χριστὸν	K^{c}
δεσπότην καὶ θεὸν καὶ κύριον ἡμῶν Ἰησοῦν Χριστὸν	*Comm*^{pt}
δεσπότην θεὸν καὶ κύριον Ἰησοῦν Χριστὸν	f^{915}

[64] Baldwin, "The So-Called Mixed Text," 124, 240 (unit 2). Baldwin's group A2 reads Χριστοῦ Ἰησοῦ.

Part of this variant (the inclusion or omission of θεὸν) has been shown to contain a primary reading for a group identified by factor analysis in Jude.[65] Although the various readings in this unit are separated primarily by minor additions and omissions, Wasserman rightly points out that many of these changes were probably not accidental. Indeed, changes involving words like θεὸν and καὶ were likely prompted by "the question (of) whether the whole phrase refers to Jesus Christ, or if the first part refers (only) to God."[66] The ambiguity is preserved in the reading of f^{1739} *et al*., and, to some extent, the reading found in some of the commentary cluster. The two Byzantine readings and the reading of f^{915} clarify the phrase in different ways, with the more widespread Byzantine reading and the f^{915} reading making a distinction between God and Jesus, while the K^{c} reading treats Jesus as the sole referent.

Variation Unit: Jude 1:5/12–20

Table 7. Jude 1:5/12–20

Variants	Witnesses
πάντα, ὅτι ὁ κύριος ἅπαξ	f^{2138}
ἅπαξ πάντα, ὅτι Ἰησοῦς	Alexpt
πάντα, ὅτι Ἰησοῦς ἅπαξ	f^{1739}, f^{915} pt
πάντα, ὅτι ὁ Ἰησοῦς ἅπαξ	f^{915} pt
ἅπαξ πάντα, ὅτι ὁ θεὸς	Alexpt
πάντα, ὅτι ὁ θεὸς ἅπαξ	Alexpt, f^{915} pt
ὑμᾶς ἅπαξ τοῦτο, ὅτι ὁ κύριος	K, Lect, K^{r}, Alexpt, K^{c}, f^{1780}, I
ὑμᾶς τοῦτο ἅπαξ, ὅτι ὁ κύριος	Comm
ἅπαξ τοῦτο, ὅτι ὁ κύριος	f^{453}, f^{0142}

Part of this variant (the inclusion or omission of ὑμᾶς) has been shown to contain a primary reading for a group identified by factor analysis in Jude.[67] Wasserman describes this variant as "one of the textually most difficult passages in Jude, and in the whole NT."[68] His decision to adopt ὑμᾶς ἅπαξ πάντα, ὅτι ὁ κύριος, a composite reading not found in any surviving Greek witness, attests to the thorny nature of this textual problem. As NMF identifies, this variation unit divides the textual tradition both between, and within,

[65] Baldwin, "The So-Called Mixed Text," 124, 244 (unit 34). Baldwin's group B3 adds θεὸν.

[66] The Epistle of Jude, 251.

[67] Baldwin, "The So-Called Mixed Text," 124, 244–245 (unit 59). Baldwin's group A1 adds ὑμᾶς.

[68] The Epistle of Jude, 255.

several branches. The most widely-attested reading is ὑμᾶς ἅπαξ τοῦτο, ὅτι ὁ κύριος, thanks to its support from several Byzantine subfamilies. The remaining Byzantine-related groups read ἅπαξ τοῦτο, ὅτι ὁ κύριος, which differs from the first reading only in the absence of ὑμᾶς.

While the variants involving word order and the presence or absence of ὑμᾶς are more common, and therefore less significant genealogically, the variants involving the choice between πάντα and τοῦτο and the subject of the clause introduced here are more significant. When these considerations are taken into account, the differences between the readings with Byzantine and Byzantine-related support (all of which feature τοῦτο and ὁ κύριος) become minor variations on one widely-accepted reading. The support for τοῦτο over πάντα in part of the Alexandrian cluster is likely an indication of contamination, as the rest of the cluster supports readings with πάντα.

Variation Unit: Jude 1:9/24–28

Table 8. Jude 1:9/24–28

Variants	Witnesses
τοῦ Μωϋσέως σώματος	K, f^{453}, *Alex*, K^{c}, f^{0142}, f^{1780}, $f^{915\ pt}$
τοῦ Μωσέως σώματος	f^{1739}, *Lect*, K^{r}, f^{2138}, *Comm*, I, $f^{915\ pt}$

This variant has been shown to contain a primary reading for a group identified by factor analysis in Jude.[69] This variant is orthographic in nature, and as the even division of NMF-assigned support indicates, both spellings of Moses's name likely arose in more than one stream of transmission independently. Even the Byzantine groups are divided here, as the margin of Robinson-Pierpont (RP) correctly notes.

Variation Unit: Jude 1:12/42–46

Table 9. Jude 1:12/42–46

Variants	Witnesses
δὶς ἀποθανόντα, ἐκριζωθέντα	K, f^{1739}, K^{r}, f^{2138}, *Alex*, K^{c}, *Comm*, f^{0142}, f^{1780}, f^{915}
δὶς ἀποθανόντα, καὶ ἐκριζωθέντα	*Lect*, f^{453}, I

[69] Baldwin, "The So-Called Mixed Text," 124, 247 (unit 124). Baldwin's group A1 reads Μεωϋσέως, but this is likely a typographical error; existing transcriptions and images of the witnesses listed in support of this reading have Μωϋσέως (up to minor orthographic variation). Baldwin appears to have split the witnesses to Μωϋσέως into two separate groups.

As the variation unit concerns the last two items in a list of qualities, the addition of a final καὶ would not be uncommon among scribes. This variation could very well have arisen independently on separate occasions.

Variation Unit: Jude 1:13/30–34

Table 10. Jude 1:13/30–34

Variants	Witnesses
εἰς αἰῶνα τετήρηται	$f^{1739\,pt}$, K^{r}, f^{2138}, f^{453}, *Alex*, K^{c}, f^{0142}, $f^{1780\,pt}$, I^{pt}, f^{915}
εἰς τὸν αἰῶνα τετήρηται	K, *Lect*, *Comm*, $f^{1780\,pt}$, I^{pt}
εἰς αἰῶνας τετήρηται	$f^{1739\,pt}$

This variant has been shown to contain a primary reading for a group identified by factor analysis in Jude.[70] All of the variant readings in this unit differ in only small ways (the addition or omission of an article or a single letter), but these differences have an effect on the stylistic smoothness of the phrase. It is worth noting that the reading εἰς τὸν αἰῶνα τετήρηται has decent support from clusters with Byzantine connections. RP is probably correct in adopting εἰς αἰῶνα τετήρηται for its text, but εἰς τὸν αἰῶνα τετήρηται might be good to include in the margin.

Variation Unit: Jude 1:15/14–18

Table 11. Jude 1:15/14–18

Variants	Witnesses
πάντας τοὺς ἀσεβεῖς	*Lect*, f^{2138}, f^{453}, *Alex*
πάντας τοὺς ἀσεβεῖς αὐτῶν	K, K^{r}, K^{c}, *Comm*, f^{1780}, I, f^{915}
πάντας ἀσεβεῖς	f^{1739}
N/A (omits in an overlapping variation unit)	f^{0142}

Part of this variant (the inclusion or omission of αὐτῶν) has been shown to contain a primary reading for a group identified by factor analysis in Jude.[71] The NA[27] and NA[28] reading πᾶσαν ψυχὴν is supported by only 3 Greek witnesses; it is not listed here because NMF classifies it as a weak reading (with weight 0.0498) in the Alexandrian profile. The most characteristic reading of this cluster (and of three other clusters) is πάντας τοὺς ἀσεβεῖς,

[70] Baldwin, "The So-Called Mixed Text," 124, 253 (unit 193). Baldwin's group A1 reads εἰς αἰῶνα.

[71] Baldwin, "The So-Called Mixed Text," 124, 255–256 (unit 220). Baldwin's group B3 adds αὐτῶν.

the reading preferred by Wasserman.[72] According to NMF, the most representative readings for the clusters that have any reading at all here are slight variations on the same idea. Most of the Byzantine clusters and some of the less-Byzantine clusters are agreed on the more expansive reading πάντας τοὺς ἀσεβεῖς αὐτῶν. Meanwhile, following the pattern we have observed up to this point, f^{1739} is isolated in supporting a much simpler construction.

Variation Unit: Jude 1:16/14–16

Table 12. Jude 1:16/14–16

Variants	Witnesses
ἐπιθυμίας ἑαυτῶν	f^{1739}, *Lect*, K^{r}, f^{1780}, I, $f^{915\ \text{pt}}$
ἐπιθυμίας αὐτῶν	K, f^{2138}, f^{453}, *Alex*, K^{c}, *Comm*, f^{0142}, $f^{915\ \text{pt}}$

This variant has been shown to contain primary readings for multiple groups identified by factor analysis in Jude.[73] As the readings and their external support suggest, the history of this variant is likely a complicated one. The Byzantine clusters are sharply divided on this issue, as the RP margin correctly notes, and the non-Byzantine clusters are also scattered. The situation suggests that both readings likely arose multiple times independently, a conclusion supported by the reasonable transcriptional probability of the one-letter change from αὐτῶν and ἑαυτῶν and vice-versa.

Variation Unit: Jude 1:25/10–20

Table 13. Jude 1:25/10–20

Variants	Witnesses
διὰ Ἰησοῦ Χριστοῦ τοῦ κυρίου ἡμῶν	f^{1739}, f^{2138}, f^{453}, *Alex*, I^{pt}, f^{915}
om.	K, *Lect*, K^{r}, K^{c}, *Comm*, f^{0142}, f^{1780}, I^{pt}

This variant has been shown to contain primary readings for multiple groups identified by factor analysis in Jude.[74] In a reversal of the situation usually associated with the Byzantine text, the Byzantine clusters omit what seems like a common doxological expansion to the text, while the non-Byzantine clusters include it.

[72] The Epistle of Jude, 301–304.

[73] Baldwin, "The So-Called Mixed Text," 125, 257 (unit 242). Baldwin's groups A4, B1, and B3 read ἑαυτῶν, ἑαυτῶν, and αὐτῶν, respectively.

[74] Baldwin, "The So-Called Mixed Text," 125, 267–268 (unit 313). Baldwin's groups A1 and B2 add, and M omits.

Variation Unit: Jude 1:25/32–38

Table 14. Jude 1:25/32–38

Variants	Witnesses
πρὸ παντὸς τοῦ αἰῶνος	$f^{2138\ pt}$, *Alex*, $f^{0142\ pt}$
πρὸ παντὸς αἰῶνος	f^{1739}, $f^{2138\ pt}$, f^{453}, f^{915}
om.	K, *Lect*, K^{r}, K^{c}, *Comm*, $f^{0142\ pt}$, f^{1780}, I

This variant has been shown to contain primary readings for multiple groups identified by factor analysis in Jude.[75] This variant effectively repeats the situation of the previous one: the Byzantine clusters (with the partial exception of the f^{0142} group) omit the longer phrase, while the non-Byzantine clusters include it, up to smaller variations.

Summary

In this paper, I have shown how non-negative matrix factorization, or NMF, can efficiently classify both MSS and readings in a collation, even in the presence of contamination. Specifically, because NMF models the classification problem in terms of additive mixture between weighted profiles of readings, it simplifies the process for users to identify common ancestral textual components and potential cases of contamination in its output tables.

On the practical side, I have demonstrated that NMF is able to factor a complete collation matrix of 518 MSS of Jude in minutes. Using NMF, we are able to classify many previously-unclassified MSS and verify several existing group classifications. Our classifications included the small, but well-known groups f^{1739}, f^{2138}, and f^{453}. Distinct textual families for lectionaries and commentaries were isolated. Well-known Alexandrian MSS classified in the same group were found, and a less-documented group f^{915} that exhibits notable textual peculiarities elsewhere in the NT was isolated. Clusters that offer empirical justification for von Soden's K^{r} and K^{c} groups, as well as for numerous branches of the Byzantine text were identified. In addition, the discussion of determinative readings identified by NMF verified the choices for the textual and marginal readings of Jude in the RP Byzantine text and proposed additional marginal readings based on the readings of the identified Byzantine subgroups.

Conclusions

NMF has tremendous potential as a tool for fast, automated, texttype-based classification, and it should be implemented in further studies. The weights that populate NMF's output classification tables furnish an instant guide to

[75] Baldwin, "The So-Called Mixed Text," 125, 268 (unit 314). Baldwin's groups A1 and B2 support the longest reading.

pure and mixed witnesses, which can be of tremendous use in witness and variation unit selection for the construction of future critical texts of the NT. Applied to complete collations or to collations with a high volume of MSS (e.g., *Text und Textwert*), NMF can distill massive datasets to more tractable ones with minimal loss of information. Because datasets of this size are present and multiplying in the INTF's Virtual Manuscript Room (VMR),[76] an NMF module would be a fitting addition to this collaborative research environment.

While NMF is not meant to make inferences regarding prior and posterior textual relationships, it could potentially facilitate more complex genealogical methods like the Coherence-Based Genealogical Method (CBGM) by giving simple and easy-to-interpret indications of pre-genealogical coherence and contamination. Checking for contamination in a MS is as easy as looking at its column in the mixture matrix (H). To estimate pre-genealogical coherence for a given variant reading, one can simply check whether any group's reading profile closely splits the weight assigned to a given reading with another reading in the same variation unit.

NMF should be implemented in future text-critical applications and improved with continued research. In light of the present work reported in this article, we can hope to find MS classifications from NMF examined further and perhaps used as starting points for new research on the complex textual history of the NT. It certainly deserves our greatest effort.

[76] Accessible at http://ntvmr.uni-muenster.de/.

APPENDIX

Classification of Lacunose Manuscripts

As explained earlier, in the process of data selection, I regarded the texts of correctors and witnesses with fewer than three hundred readings as fragmentary and therefore secondary to our application. Because of their age, most papyri and majuscules are so lacunose that they must be excluded in this way. This leaves us with an unfortunate situation, in which we have nothing to say about the MSS in which we are most interested.

Thankfully, a simple solution is available. Once NMF on the primary set of witnesses has produced a basis matrix *W* for reading profiles, we can use this matrix to classify the secondary witnesses by whatever readings they do have, as we would in the confirmatory step of the CPM. While mathematical details are beyond the scope of this discussion, it will suffice to say that freely-available software libraries can handle this task within seconds.[77]

For the sake of space, I will not list the mixture weights of all secondary MSS. The weights of GA 2138 and the consistently-cited NA[28] witnesses $\mathfrak{P}^{74}$, $\mathfrak{P}^{78}$, 025, and 1852 are summarized below.

The Papyrus $\mathfrak{P}^{74}$

The papyrus $\mathfrak{P}^{74}$ has positive weights for the following groups: 0.0061 for f^{1739}, 0.0247 for f^{2138}, 0.0010 for *Alex*, 0.0057 for K^{c}, 0.0002 for *Comm*, 0.0423 for f^{0142}, 0.0040 for f^{1780}, 0.0124 for I, and 0.0122 for f^{915}. The precise textual complexion of this witness is elusive, in part because of its extremely fragmentary state and in part because where the MS's readings can be deduced, they are assigned low weights by NMF (meaning they are not important to any group's reading profile). Indeed, one of the only places where $\mathfrak{P}^{74}$'s reading is unambiguous is in Jude 1:12/16, where it reads σπιλάδες with virtually all other MSS.

The Papyrus $\mathfrak{P}^{78}$

This papyrus fares significantly better, and with surprising results: its positive mixture weights are 0.0015 for *Lect*, 0.1274 for f^{2138}, 0.0174 for *Comm*, 0.0311 for f^{0142}, 0.0012 for f^{1780}, and 0.0237 for f^{915}. The high weight for f^{2138} comes from the reading ἐπέχουσαι in Jude 1:7/50. Without further readings available, we can only conjecture a genealogical relationship between this witness and the family in question.

---

[77] Implementation details and code can be found at https://github.com/jjmccollum/jude-nmf.

The Majuscle 025 / P

This majuscule has even better results: 0.3062 for *Lect*, 0.1884 for K^r, 0.1320 for f^{0142}, and 0.2126 for f^{1780}. Given the groups that best fit its extant readings, we can confidently classify this as a broadly Byzantine witness, but it is difficult to tell whether the nearly equal mixture from the clusters involved is due to contamination or simply because the gaps in the witness prevent a more certain classification.

The Minuscle 1852

In contrast, we can confidently declare MS 1852 to be anything but Byzantine: it has positive group weights of 0.2790 from f^{1739}, 0.5953 from f^{2138}, 0.6042 from *Alex*, and 0.3392 from f^{915}. Again, it is unclear whether the mixture observed here is real or only apparent due to the lacunose nature of the MS.

The Minuscle 2138

As we would expect, this minuscule is strongly classified as a member of the cluster bearing its name: it has mixture weights of 0.1631 from f^{1739}, 1.9754 from f^{2138}, 0.0854 from *Alex*, and 0.0192 from K^c. The strength of the classification is helped by the fact that 2138 falls just below the threshold of minimum extant readings, being extant in 282 variation units.