NUMERIC METHODS FOR NONNEGATIVE MATRIX FACTORIZATION [*]

LARSON HOGSTROM[†]

Abstract. This report introduces the framework for parts-based representations using NMF and focuses on the algorithms and numerical aspects of computation.

Key words. dimensionality reduction, matrix factorization, least squares, nonnegativity, data representation

1. Introduction. Nonnegative matrix factorization (NMF) is a dimensionality reduction technique that can be applied in a variety of domains. Like principal component analysis (PCA), NMF seeks to describe a large matrix of data using a small set of information dense components. Unlike PCA, which imposes orthogonality constraints on components, NMF requires nonnegativity after factorization. This procedure allows for only additive combinations of positive components to estimate the input matrix. The endpoint is a parts-based representation of the starting data where each component is more easily interpreted than PCA. A classical example of NMF is the factorization of a large database of human face images. Factorization using PCA will yield a series of ?eigenfaces? which can be added or subtracted, but individual components are not easy to interpret on their own. The NMF ”parts-based” representation however, yields components like noses, eyes, ears, and mouths which can be added linearly to construct a face.

A number of algorithms have been proposed in order to construct the NMF factorization with positivity constraints. Iterative algorithms, for example, build the WH factorization while converging on a local maximum of a specified object function. This project will examine two of such objective functions 1) the conventional least squares difference between the input and the factorization and 2) an objective function based on the Kullback-Leibler divergence2. The flop counts of algorithmic update rules for both approaches will both be assessed analytically and the different conditions for monotonic convergence will be examined. Both algorithms will be implemented in python using the ?numpy? and ?scipy? packages. The two approaches will be compared as follows: 1) In numeric simulations of input matrices with known contributions of signal and noise 2) using real-world gene expression data of cancer cells that have been exposed to different drugs grouped into known pharmacological classes. In simulations, the two algorithms will be compared by their accuracy after a fixed number of iterations. The approaches will also be compared in their ability to classify the gene expression profiles into known pharmacological groups.

2. Problem Definition and Multiplicative Update Rules. Text leading to definition...

definition For a nonnegative matrix $\mathbf{A} \in \mathbf{R^{mxn}}$, select a low-rank approximation of size k such that there are two nonnegative matrices $\mathbf{W} \in R^{mxk}$ and $\mathbf{H} \in R^{kxn}$ which minimizes a function such as

$$f(\mathbf{W},\mathbf{H}) = \frac{1}{2}||\mathbf{A}\ \text{-}\ \mathbf{WH}||_F^2$$

[*] A final project for 18.335, Fall 2015

[†] (hogstrom@mit.edu).

Other commonly used objective functions include Euclidian distance and Kullback-Leibler (KL) divergence. KL can be extended to a more general information-based framework using Renyi's divergence. (Devarajan, 2005). Here, a single parameter α is used to represent a continuum of distance measures and KL airises as a special case as $\alpha \to 1$.

$$KL(V||WH) = \sum_{ij} [V_{ij} \log \frac{V_{ij}}{(WH)_{ij}} - V_{ij} + (WH)_{ij}]$$

The top matter of a journal paper falls into a standard format. It begins of course with the `\documentclass` command

```
\documentclass{siamltex1213}
```

Other class options can be included in the bracketed argument of the command, separated by commas.

Suggested optional arguments include:

final Without this option, lines which extend past the margin will have black boxes next to them to help authors identify lines that they need to fix, by re-writing or inserting breaks. `final` turns these boxes off, so that very small margin breaks which are not noticible will not cause boxes to be generated.

oneeqnum Normally `siamltex.cls` numbers equations, tables, figures, and theorem environments with a decimal number, composed of the section of the paper, a period, and the number of the enumerated object (example: 1.2). The sequence of numbering is also restarted with each new section, so that, for example, the last equation of section 3 may be 3.10, but the first equation of section 4 would be 4.1. Using `oneeqnum` numbers all equations consecutively throughout a paper with a single digit.

onethmnum Using `onethmnum` numbers all theorem-like environments consecutively throughout a paper with a single digit.

onefignum Using `onethmnum` numbers all figures consecutively throughout a paper with a single digit.

onetabnum Using `onethmnum` numbers all tables consecutively throughout a paper with a single digit.

The title and author parts are formatted using the `\title` and `\author` commands as described in Lamport [3]. The `\date` command is not used. `\maketitle` produces the actual output of the commands.

The addresses and support acknowledgments are put into the `\author` commands via `\thanks`. If support is overall for the authors, the support acknowledgment should be put in a `\thanks` command in the `\title`. Specific support should go following the addresses of the individual authors in the same `\thanks` command.

Sometimes authors have support or addresses in common which necessitates having multiple `\thanks` commands for each author. Unfortunately LaTeX does not normally allow this, so a special procedure must be used. An example of this procedure follows. Grant information can also be run into both authors' footnotes.

```
\title{TITLE OF PAPER}

\author{A.~U. Thorone\footnotemark[2]\ \footnotemark[5]
\and A.~U. Thortwo\footnotemark[3]\ \footnotemark[5]
\and A.~U. Thorthree\footnotemark[4]}

\begin{document}
\maketitle

\renewcommand{\thefootnote}{\fnsymbol{footnote}}

\footnotetext[2]{Address of A.~U. Thorone}
\footnotetext[3]{Address of A.~U. Thortwo}
\footnotetext[4]{Address of A.~U. Thorthree}
\footnotetext[5]{Support in common for the first and second
authors.}

\renewcommand{\thefootnote}{\arabic{footnote}}
```

Notice that the footnote marks begin with `[2]` because the first mark (the asterisk) will be used in the title for date-received information by SIAM, even if not already used for support data. This is just one example; other situations follow a similar pattern.

Following the author and title is the abstract, key words listing, and AMS subject classification number(s), designated using the `{abstract}`, `{keywords}`, and `{AMS}` environments. If there is only one AMS number, the commands `\begin{AM}` and `\end{AM}` are used instead of `{AMS}`. This causes the heading to be in the singular. Authors are responsible for providing AMS numbers. They can be found in the Annual Index of Math Reviews, or through `e-Math` (`telnet e-math.ams.com`; login and password are both `e-math`).

Left and right running heads should be provided in the following way.

```
\pagestyle{myheadings}
\thispagestyle{plain}
\markboth{A.~U. THORONE AND A.~U. THORTWO}{SHORTER PAPER TITLE}
```

3. Equations and mathematics.

One advantage of LaTeX is that it can automatically number equations and refer to these equation numbers in text. While plain TeX's method of equation numbering (explicit numbering using `\leqno`) works in the SIAM macro, it is not preferred except in certain cases. SIAM style guidelines call for aligned equations in many circumstances, and LaTeX's `{eqnarray}` environment is not compatible with `\leqno` and LaTeX is not compatible with the plain TeX command `\eqalign` and `\leqalignno`. Since SIAM may have to alter or realign certain groups of equations, it is necessary to use the LaTeX system of automatic numbering.

Sometimes it is desirable to designate subequations of a larger equation number. The subequations are designated with (roman font) letters appended after the number. SIAM has supplemented its macros with the `subeqn.clo` option which defines the environment `{subequations}`.

```
\begin{subequations}\label{EKx}
\begin{equation}
 y_k =  B  y_{k-1} +  f, \qquad k=1,2,3,\ldots
```

```
\end{equation}
for  any initial vector $ y_0$.   Then
\begin{equation}
 y_k\rightarrow  u \mbox{\quad iff\quad} \rho( B)<1.
\end{equation}
\end{subequations}
```

All equations within the `{subequations}` environment will keep the same overall number, but the letter designation will increase.

Clear equation formatting using TeX can be challenging. Aside from the regular TeX documentation, authors will find Nicholas J. Higham's book *Handbook of Writing for the Mathematical Sciences* [2] useful for guidelines and tips on formatting with TeX. The book covers many other topics related to article writing as well.

Authors commonly make mistakes by using `<`, `>`, `\mid`, and `\parallel` as delimiters, instead of `\langle`, `\rangle`, `|`, and `\|`. The incorrect symbols have particular meanings distinct from the correct ones and should not be confused.

Table 1
Illustration of incorrect delimiter use.

<table>
<tr><th colspan="2">Wrong</th><th colspan="2">Right</th></tr>
<tr><td><code>&lt;x, y&gt;</code></td><td>$< x, y >$</td><td><code>\langle x, y\rangle</code></td><td>$\langle x, y\rangle$</td></tr>
<tr><td><code>5 &lt; \mid A \mid</code></td><td>$5 <\mid A \mid$</td><td><code>5 &lt; |A|</code></td><td>$5 < |A|$</td></tr>
<tr><td><code>6x = \parallel x</code>
<code>- 1\parallel_{i}</code></td><td>$6x =\parallel x - 1 \parallel_i$</td><td><code>6x = \|x - 1\|_{i}</code></td><td>$6x = \|x-1\|_i$</td></tr>
</table>

Another common author error is to put large (and even medium sized) matrices in-line with the text, rather than displaying them. This creates unattractive line spacing problems, and should be assiduously avoided. Text-sized matrices (like $\left(\begin{smallmatrix} a & b \\ b & c \end{smallmatrix}\right)$) might be used but anything much more complex than the example cited will not be easy to read and should be displayed.

More information on the formatting of equations and aligned equations is found in Lamport [3]. Authors bear primary responsibility for formatting their equations within margins and in an aesthetically pleasing and informative manner.

The SIAM macros include additional roman math words, or "log-like" functions, to those provided in standard TeX. The following commands are added: `\const`, `\diag`, `\grad`, `\Range`, `\rank`, and `\supp`. These commands produce the same word as the command name in math mode, in upright type.

4. Special fonts. SIAM supports the use of the AMS-TeX fonts (version 2.0 and later). The package `amsfonts` can be included with the command `\usepackage{amsfonts}`. This package is part of the AMS-LaTeXdistribution, available from the AMS or from the Comprehensive TeX Archive Network (anonymous ftp to ftp.shsu.edu). The blackboard bold font in this font package can be used for designating number sets. This is preferable to other methods of combining letters (such as I and R for the real numbers) to produce pseudo-bold letters but this is tolerable as well. Typographically speaking, number sets may simply be designated using regular bold letters; the blackboard bold typeface was designed to fulfil a desire to simulate the limitations of a chalk board in printed type.

4.1. Punctuation. All standard punctuation and all numerals should be set in roman type (upright) even within italic text. The only exceptions are periods and commas. They may be set to match the surrounding text.

References to sections should use the symbol §, generated by `\S`. (If the reference begins a sentence, the term "Section" should be spelled out in full.) Authors should not redefine `\S`, say, to be a calligraphic S, because `\S` must be reserved for use as the section symbol.

Authors sometimes confuse the use of various types of dashes. Hyphens (`-`, -) are used for some compound words (many such words should have no hyphen but must be run together, like "nonzero," or split apart, like "well defined"). Minus signs (`$-$`, $-$) should be used in math to represent subtraction or negative numbers. En dashes (`--`, –) are used for ranges (like 3–5, June–August), or for joined names (like Runge–Kutta). Em dashes (`---`, —) are used to set off a clause—such as this one—from the rest of the sentence.

4.2. Text formatting. SIAM style preferences do not make regular use of the `{enumerate}` and `{itemize}` environments. Instead, `siamltex.cls` includes definitions of two alternate list environments, `{remunerate}` and `{romannum}`. Unlike the standard itemized lists, these environments do not indent the secondary lines of text. The labels, whether defaults or the optional user-defined, are always aligned flush right.

The `{remunerate}` environment consecutively numbers each item with an arabic numeral followed by a period. This number is always upright, even in slanted environments. (For those wondering at the unusual naming of this environment, it comes from Seroul and Levy's [4] definition of a similar macro for plain TEX: `\meti` which is `\item` spelled backwards. Thus `{remunerate}` a portion of `{enumerate}` spelled backwards.)

The `{romannum}` environment consecutively numbers each item with a lower-case roman numeral enclosed in parentheses. This number will always be upright within slanted environments (as in theorems).

5. Theorems and Lemmas. Theorems, lemmas, corollaries, definitions, and propositions are covered in the SIAM macros by the theorem-environments `{theorem}`, `{lemma}`, `{corollary}`, `{definition}` and `{proposition}`. These are all numbered in the same sequence and produce labels in small caps with an italic body. Other environments may be specified by the `\newtheorem` command. SIAM's style is for Remarks and Examples to appear with italic labels and an upright roman body.

```
\begin{theorem}
Sample theorem included for illustration.
Numbers and parentheses, like equation $(3.2)$, should be set
in roman type.  Note that words (as opposed to ''log-like''
functions) in displayed equations, such as
$$ x^2 = Y^2 \sin z^2 \mbox{ for all } x $$
will appear in italic type in a theorem, though normally
they should appear in roman.\end{theorem}
```

This sample produces Theorem 4.1 below.

Theorem 5.1. *Sample theorem included for illustration. Numbers and parentheses, like equation* (3.2)*, should be set in roman type. Note that words (as opposed to* *"log-like" functions) in displayed equations, such as*

$$x^2 = Y^2 \sin z^2 \textit{ for all } x$$

will appear in italic type in a theorem, though normally they should appear in roman.

Proofs are handled with the `\begin{proof}` `\end{proof}` environment. A "QED" box □ is created automatically by `\end{proof}`, but this should be preceded with a `\qquad`.

Named proofs, if used, must be done independently by the authors. SIAM style specifies that proofs which end with displayed equations should have the QED box two ems (`\qquad`) from the end of the equation on line with it horizontally. Below is an example of how this can be done:

```
{\em Proof}. Proof of the previous theorem
                  .
                  .
                  .
thus,
$$
a^2 + b^2 = c^2 \qquad\endproof
$$
```

6. Figures and tables. Figures and tables sometimes require special consideration. Tables in SIAM style are need to be set in eight point size by using the `\footnotesize` command inside the `\begin{table}` environment. Also, they should be designed so that they do not extend beyond the text margins.

SIAM style requires that no figures or tables appear in the references section of the paper. LaTeX is notorious for making figure placement difficult, so it is important to pay particular attention to figure placement near the references in the text. All figures and tables should be referred to in the text.

SIAM supports the use of `epsfig` for including PostScript figures. All PostScript figures should be sent in separate files. See the `epsfig` documentation (available via anonymous ftp from CTAN: ftp.shsu.edu) for more details on the use of this style option. It is a good idea to submit high-quality hardcopy of all PostScript figures just in case there is difficulty in the reproduction of the figure. Figures produced by other non-TeX methods should be included as high-quality hardcopy when the manuscript is submitted.

PostScript figures that are sent should be generated with sufficient line thickness. Some past figures authors have sent had their line widths become very faint when SIAM set the papers using a high-quality 1200dpi printer.

Hardcopy for non-PostScript figures should be included in the submission of the hardcopy of the manuscript. Space should be left in the `{figure}` command for the hardcopy to be inserted in production.

7. Bibliography and BibTeX. If using BibTeX, authors need not submit the `.bib` file for their papers. Merely submit the completed `.bbl` file, having used `siam.bst` as their bibliographic style file. `siam.bst` only works with BibTeX version 99i and later. The use of BibTeX and the preparation of a `.bib` file is described in greater detail in [3].

If not using BibTeX, SIAM bibliographic references follow the format of the following examples:

```
\bibitem{Ri} {\sc W. Riter},
{\em Title of a paper appearing in a book}, in The Book
Title, E.~D. One, E.~D. Two, and A.~N. Othereditor, eds.,
Publisher, Location, 1992, pp.~000--000.

\bibitem{AuTh1} {\sc A.~U. Thorone}, {\em Title of paper
with lower case letters}, SIAM J. Abbrev. Correctly, 2
(1992), pp.~000--000.

\bibitem{A1A2} {\sc A.~U. Thorone and A.~U. Thortwo}, {\em
Title of paper appearing in book}, in Book Title: With All
Initial Caps, Publisher, Location, 1992.

\bibitem{A1A22} \sameauthor, % generates the 3 em rule
{\em Title of Book{\rm :} Note Initial Caps and {\rm ROMAN
TYPE} for Punctuation and Acronyms}, Publisher,
Location, pp.~000--000, 1992.

\bibitem{AuTh3} {\sc A.~U. Thorthree}, {\em Title of paper
that's not published yet}, SIAM. J. Abbrev. Correctly, to appear.
```

Other types of references fall into the same general pattern. See the sample file or any SIAM journal for other examples. Authors must correctly format their bibliography to be considered as having used the macros correctly. An incorrectly formatted bibliography is not only time-consuming for SIAM to process but it is possible that errors may be introduced into it by keyboarders/copy editors.

As an alternative to the above style of reference, an alphanumeric code may be used in place of the number (e.g., [AUTh90]). The same commands are used, but `\bibitem` takes an optional argument containing the desired alphanumeric code.

Another alternative is no number, simply the authors' names and the year of publication following in parentheses. The rest of the format is identical. The macros do not support this alternative directly, but modifications to the macro definition are possible if this reference style is preferred.

8. Conclusion. Many other style suggestions and tips could be given to help authors but are beyond the scope of this document. Simple mistakes can be avoided by increasing your familiarity with how LaTeX functions. The books referred to throughout this document are also useful to the author who wants clear, beautiful typography with minimal mistakes.

Appendix. The use of appendices. The `\appendix` command may be used before the final sections of a paper to designate them as appendices. Once `\appendix` is called, all subsequent sections will appear as

Appendix A. Title of appendix. Each one will be sequentially lettered instead of numbered. Theorem-like environments, subsections, and equations will also have the section number changed to a letter.

If there is only *one* appendix, however, the `\Appendix` (with a capital letter) should be used instead. This produces only the word **Appendix** in the section title, and does not add a letter. Equation numbers, theorem numbers and subsections of the appendix will have the letter “A” designating the section number.

If you don’t want to title your appendix, and just call it **Appendix A.** for example, use `\appendix\section*{}` and don’t include anything in the title field. This works opposite to the way `\section*` usually works, by including the section number, but not using a title.

Appendices should appear before the bibliography section, not after, and any acknowledgments should be placed after the appendices and before the bibliography.

REFERENCES

[1] M. Goossens, F. Mittelbach, and A. Samarin, *The LaTeX Companion*, Addison-Wesley, Reading, MA, 1994.

[2] N. J. Higham, *Handbook of Writing for the Mathematical Sciences*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1993.

[3] L. Lamport, LaTeX: *A Document Preparation System*, Addison-Wesley, Reading, MA, 1986.

[4] R. Seroul and S. Levy, *A Beginner’s Book of* TeX, Springer-Verlag, Berlin, New York, 1991.