Constrained Versions of DEDICOM for Use in Unsupervised Part-Of-Speech Tagging

Daniel M. Dunlavy, Peter A. Chew

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy’s National Nuclear Security Administration under contract DE-AC04-94AL85000.

Approved for public release; further dissemination unlimited.
Constrained Versions of DEDICOM for Use in Unsupervised Part-Of-Speech Tagging

Daniel M. Dunlavy  
Scalable Analysis & Viz Department  
Sandia National Laboratories  
P.O. Box 5800  
Albuquerque, NM 87185-1327  
dmdunla@sandia.gov

Peter A. Chew  
Galisteo Consulting Group, Inc.  
4004 Carlisle Blvd NE, Suite H  
Albuquerque, NM 87107  
pachew@galisteoconsulting.com

Abstract

This report describes extensions of DEDICOM (DEcomposition into DIrectional COMponents) data models [3] that incorporate bound and linear constraints. The main purpose of these extensions is to investigate the use of improved data models for unsupervised part-of-speech tagging, as described by Chew et al. [2]. In that work, a single domain, two-way DEDICOM model was computed on a matrix of bigram frequencies of tokens in a corpus and used to identify parts-of-speech as an unsupervised approach to that problem. An open problem identified in that work was the computation of a DEDICOM model that more closely resembled the matrices used in a Hidden Markov Model (HMM), specifically through post-processing of the DEDICOM factor matrices. The work reported here consists of the description of several models that aim to provide a direct solution to that problem and a way to fit those models. The approach taken here is to incorporate the model requirements as bound and linear constraints into the DEDICOM model directly and solve the data fitting problem as a constrained optimization problem. This is in contrast to the typical approaches in the literature, where the DEDICOM model is fit using unconstrained optimization approaches, and model requirements are satisfied as a post-processing step.
This page intentionally left blank.
1 Notation

In this report, we use the following conventions for notation. Matrices are denoted as bolded capital letters, e.g., $A$. Scalar matrix elements are denoted in capital letters with subscripts referencing row and column indices, respectively, e.g., $A_{ij}$ is the element of $A$ in row $i$ and column $j$. The Frobenius norm of a matrix is denoted as $\| \cdot \|_F$ and defined as $\sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} |A_{ij}|^2}$ for a matrix $A \in \mathbb{R}^{m \times n}$. Column-stochastic (row-stochastic) matrices are those whose columns (rows) sum to 1. Doubly stochastic matrices have columns and rows that sum to 1.
2 Description of the DEDICOM Models

The DEDICOM model was originally developed to identify latent factors contributing to asymmetric relationships amongst a group of individuals [3]. In that work, the relationships consisted of how much person $i$ liked person $j$, denoted as $X_{ij}$, and all ratings were captured in a data matrix $X$. The goal in that original work was to identify the underlying relationships described at “higher levels” (i.e., latent factors) than those observed (liking ratings). Since that original work, there have been many uses of the DEDICOM model to identify latent factors for asymmetric relational data. Below we present the original model and several extensions that incorporate constraints on the latent factors that we propose may help identify the relationships in different ways.

2.1 Original DEDICOM Model

Given a data matrix $X \in \mathbb{R}^{n \times n}$ that describes asymmetric relations between $n$ objects and $k \ll n$, the single domain, two-way DEDICOM model is a latent factor model of the form

$$X = ARA^T + E$$

(1)

where $A \in \mathbb{R}^{n \times k}$ describes the relationships between the objects and latent factors, $R \in \mathbb{R}^{k \times k}$ describes the relationships between the latent factors, and $E \in \mathbb{R}^{n \times n}$ accounts for the errors in the model.

These models are often fit by finding the best approximation

$$X \approx ARA^T$$

(2)

by solving the following minimization problem

$$\min_{A,R} \|X - ARA^T\|_F^2.$$  

(3)

Note that there are no constraints imposed on the factors $A$ and $R$ in this form of the model.

This is the model employed in the work of Chew et al. [2] and used as the basis for an unsupervised approach to part-of-speech tagging.

2.2 Nonnegative DEDICOM Model

DEDICOM is often applied to nonnegative data (ratings, counts, affinities, etc.), yet nothing in the original model requires the latent factors to be nonnegative. To aid in interpreting these model factors, one can include nonnegativity constraints on the DEDICOM factor
matrices by solving the following problem

\[
\min_{\mathbf{A}, \mathbf{R}} \|\mathbf{X} - \mathbf{A}\mathbf{R}^T\|_F^2
\]

s.t. \(A_{ij} \geq 0\)
\(R_{ij} \geq 0\).

The use of nonnegative DEDICOM models has been shown to improve interpretability of the latent factors in analyzing relationships of senders and recipients of email messages [1].

### 2.3 DEDICOM Model with Stochastic Factors

One of the goals of the work reported here was to extend the models identified in the part-of-speech work by Chew et al. [2] by incorporating constraints directly into the model formulation. As mentioned above, the relationship between the DEDICOM model and HMMs was established in that work. Specifically, a DEDICOM model with column-stochastic \(\mathbf{A}\) and row-stochastic \(\mathbf{R}\) approximate the emission and transition probability matrices, respectively, of a \(k\)-state HMM. Incorporating these constraints into the model, we get the following problem

\[
\min_{\mathbf{A}, \mathbf{R}} \|\mathbf{X} - \mathbf{A}\mathbf{R}^T\|_F^2
\]

s.t.
\[
\sum_{i=1}^{n} A_{ij} = 1 \quad \forall j \in \{1, \ldots, k\}
\]
\[
\sum_{j=1}^{n} R_{ij} = 1 \quad \forall i \in \{1, \ldots, k\}
\]
\(A_{ij} \geq 0\)
\(R_{ij} \geq 0\).

In [2], satisfaction of the constraints above was achieved by normalizing the columns and rows of solution factors \(\mathbf{A}\) and \(\mathbf{R}\), respectively, for an unconstrained version of this formulation.

### 2.4 DEDICOM Model with Fully Stochastic Factors

Another goal of the work reported here was to implement the model identified as an open problem in the part-of-speech work by Chew et al. [2]. Specifically, a DEDICOM model with column-stochastic \(\mathbf{A}\) and doubly-stochastic \(\mathbf{R}\) would approximate the emission and
transition probability matrices, respectively, of a \( k \)-state HMM better than the model in Section 2.3. Incorporating these constraints into the model, we get the following problem

\[
\min_{A,R} \| X - ARA^T \|_F^2 \tag{6}
\]

\[
\text{s.t.} \quad \sum_{i=1}^n A_{ij} = 1 \quad \forall j \in \{1, \ldots, k\}
\]
\[
\sum_{j=1}^n R_{ij} = 1 \quad \forall i \in \{1, \ldots, k\}
\]
\[
\sum_{i=1}^n R_{ij} = 1 \quad \forall j \in \{1, \ldots, k\}
\]
\[
A_{ij} \geq 0 \quad R_{ij} \geq 0 \quad B_{ij} \geq 0.
\]

The original suggestion in [2] was to use Sinkhorn–Knopp diagonalization [7] of \( R \) as a post-processing step of the DEDICOM model fit, but the model above incorporates the stochastic matrix constraints directly into the model.

### 2.5 Weak DEDICOM Model with Stochastic Factors

In the original DEDICOM work [3], the notion of a weak model was introduced, when an extra factor, \( B \), was used to allow for relationships between objects from different domains. Combining the weak DEDICOM model with the constraints of stochastic factor matrices, we get the following problem

\[
\min_{A,R} \| X - ARB^T \|_F^2 \tag{7}
\]

\[
\text{s.t.} \quad \sum_{i=1}^n A_{ij} = 1 \quad \forall j \in \{1, \ldots, k\}
\]
\[
\sum_{j=1}^n R_{ij} = 1 \quad \forall i \in \{1, \ldots, k\}
\]
\[
A_{ij} \geq 0 \quad R_{ij} \geq 0 \quad B_{ij} \geq 0.
\]

This model will compute approximations to the HMM probability matrices, with enough degrees of freedom in the model to capture the count information accurately in the factor \( B \), while allowing for satisfaction of the stochastic constraints on the factors \( A \) and \( R \).
3 Fitting the DEDICOM Models

For the work performed by Chew et al. [2] and Bader et al. [1], MATLAB was used for the data fitting of the models in (3) and (4), respectively. The former used an alternating least squares approach, whereas the latter used a multiplicative update approach that is often employed in nonnegative matrix factorization. One of the goals of the work reported here was to experiment with different constraints added to the original DEDICOM model. There exists previous efforts attempting to address DEDICOM modeling under specific constraints [6, 8], but to our knowledge no general framework for exploring constrained DEDICOM models exists. Thus, we explored the use of the Pyomo modeling framework [5] for model specification and data fitting, as this framework allows for rapid prototyping of model formulations and incorporation of various constraints without the need to implement solvers specific to each new model.

For the models described above, we prototyped each model in Pyomo and used the IPOPT solver to fit the models to data. The data used in developing the DEDICOM models above is a subset of the car data example prepared by Harshman et al. [4] to demonstrate the use of DEDICOM in market analysis applications. The specific subset of data used is presented in Appendix A. We note some of the models failed to converge to a good fit, and this may be due to infeasability issues associated with the addition of the constraints. See Appendices B–E for example output from the runs on the car data for each of the models described above. These are only example runs, as each run uses a different starting point for fitting the model via optimization. Although this has not been proven, the empirical evidence from our experiments indicate that some of the models may not be feasible. More work is needed to better understand these constraints from an optimization perspective.
4 Conclusions and Future Work

This report identifies several variants of the DEDICOM model that may be useful for unsupervised part-of-speech tagging. Using the Pyomo modeling framework, we have identified which of the models may have feasible solutions when fitting the models, as identified empirically using a small data set. Several challenges have been identified throughout this work, but there is promise in some of the results of these models.

One of the challenges in working with Pyomo is that the interface does not provide for efficient sparse computational kernels. For the problem of part-of-speech tagging, where the data matrix is very sparse, containing bigram frequencies of tokens in a text corpus, overcoming this challenge is crucial. We have been in contact with the Pyomo developers, who have indicated that such a capability will be made available in an upcoming release of Pyomo. We will continue to pursue the incorporation of this capability and identify improvements in efficiency as a result. In order to address much larger data problems than were used in our experiments, these capabilities will have to be included in Pyomo, or an alternative approach will need to be developed.

Another challenge that was not addressed in this work is the applicability of the various models to the problem of part-of-speech tagging. The focus of the work reported here was on modeling and data fitting, but whether any one of these models will help improve part-of-speech tagging remains to be seen. To that end, metrics for assessing the utility of a particular model will need to be developed. The metrics used by Chew et al. [2] required a subject matter expert.
References


A Car Buying Data Set

\[ X = [ 
3762 \ 4216 \ 3088 \ 3387 \ 4021 \ 3523 \ 4718 \ 3487 \ 4411 \ 4116 \\
4618 \ 5164 \ 3842 \ 4183 \ 4899 \ 4267 \ 5822 \ 4233 \ 5429 \ 5114 \\
3220 \ 3692 \ 2751 \ 2989 \ 3466 \ 3011 \ 4138 \ 3010 \ 3918 \ 3671 \\
3600 \ 4078 \ 3034 \ 3301 \ 3849 \ 3349 \ 4582 \ 3335 \ 4307 \ 4044 \\
4336 \ 4806 \ 3553 \ 3882 \ 4587 \ 4008 \ 5416 \ 3961 \ 5025 \ 4727 \\
3844 \ 4222 \ 3116 \ 3409 \ 4047 \ 3541 \ 4766 \ 3489 \ 4397 \ 4142 \\
4940 \ 5588 \ 4144 \ 4516 \ 5284 \ 4604 \ 6272 \ 4580 \ 5892 \ 5524 \\
3868 \ 4266 \ 3163 \ 3452 \ 4075 \ 3558 \ 4820 \ 3513 \ 4457 \ 4205 \\
4900 \ 5522 \ 4131 \ 4484 \ 5211 \ 4526 \ 6228 \ 4505 \ 5833 \ 5501 \\
4350 \ 4976 \ 3713 \ 4032 \ 4673 \ 4058 \ 5584 \ 4055 \ 5279 \ 4953 
] ; 
\]
B Results of the Nonnegative DEDICOM Model

Number of Iterations....: 104

Objective..............: 1.0654187181450950e-21 1.3364612400412071e-19
Constraint violation...: 1.8189894035458565e-12 1.8189894035458565e-12
Overall NLP error.......: 2.5059039427241684e-09 3.1434059057531967e-07

Number of objective function evaluations = 114
Number of objective gradient evaluations = 75
Number of equality constraint evaluations = 114
Number of inequality constraint evaluations = 0
Total CPU secs in IPOPT (w/o function evaluations) = 0.126
Total CPU secs in NLP function evaluations = 0.020

EXIT: Optimal Solution Found.

A = [
59949.45 26854.16 32991.50
40560.32 47601.66 51364.50
33291.56 42876.10 24946.77
35786.95 42030.81 33427.73
46872.71 34730.72 50036.85
41212.55 25817.38 48439.88
55697.63 53402.68 44749.85
35117.58 31426.90 48792.65
34247.94 60472.61 52692.16
41229.70 58205.57 35999.58
];

R = [
2.29e-07 1.78e-07 2.530e-07
2.24e-07 4.08e-07 1.985e-07
4.40e-07 2.24e-07 2.292e-07
];
C Results of the DEDICOM Model with Stochastic Factors

Number of Iterations....: 20000

<table>
<thead>
<tr>
<th></th>
<th>(scaled)</th>
<th>(unscaled)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Objective....</td>
<td>1.4997190149302384e+07</td>
<td>1.8812475323284910e+09</td>
</tr>
<tr>
<td>Constraint violation....</td>
<td>1.1148821810451366e+01</td>
<td>1.1148821810451366e+01</td>
</tr>
<tr>
<td>Overall NLP error.........</td>
<td>8.9357233256128961e+02</td>
<td>1.1208971339648817e+05</td>
</tr>
</tbody>
</table>

Number of objective function evaluations = 38252
Number of objective gradient evaluations = 18848
Number of equality constraint evaluations = 38253
Number of inequality constraint evaluations = 0
Total CPU secs in IPOPT (w/o function evaluations) = 21.673
Total CPU secs in NLP function evaluations = 5.826

EXIT: Maximum Number of Iterations Exceeded.

A = [
0.33 3.00 0.00
1.43 1.20 0.26
0.42 0.00 0.00
0.23 0.00 0.10
0.97 0.00 0.47
0.37 0.17 0.00
1.92 0.00 0.92
0.34 0.12 0.13
1.86 0.20 0.61
1.21 0.11 0.55
];

R = [
11.61 2.78 11.03
0.00 0.03 0.00
0.00 0.00 0.00
];
**D DEDICOM Model with Fully Stochastic Factors**

Number of Iterations....: 20000

<table>
<thead>
<tr>
<th></th>
<th>(scaled)</th>
<th>(unscaled)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Objective.................</td>
<td>1.5100312417665865e+07</td>
<td>1.8941831896720061e+09</td>
</tr>
<tr>
<td>Constraint violation....</td>
<td>1.7472939885166665e+01</td>
<td>1.7472939885166665e+01</td>
</tr>
<tr>
<td>Overall NLP error........</td>
<td>1.388943027115387e+02</td>
<td>1.7422980053213541e+04</td>
</tr>
</tbody>
</table>

Number of objective function evaluations = 28151
Number of objective gradient evaluations = 18946
Number of equality constraint evaluations = 28155
Number of inequality constraint evaluations = 0
Total CPU secs in IPOPT (w/o function evaluations) = 21.371
Total CPU secs in NLP function evaluations = 5.625

EXIT: Maximum Number of Iterations Exceeded.

A = 

\[
\begin{bmatrix}
0.12 & 0.14 & 0.07 \\
0.15 & 1.20 & 0.93 \\
0.07 & 0.00 & 0.00 \\
0.00 & 3.00 & 0.00 \\
0.18 & 1.02 & 0.00 \\
0.15 & 0.00 & 0.00 \\
0.17 & 1.87 & 0.27 \\
0.11 & 0.04 & 0.07 \\
0.09 & 1.66 & 0.17 \\
0.11 & 1.23 & 0.00
\end{bmatrix}
\]

R = 

\[
\begin{bmatrix}
0.52 & 0.12 & 0.34 \\
0.00 & 1.60 & 0.03 \\
0.47 & 0.02 & 1.25
\end{bmatrix}
\]
E Weak DEDICOM Model with Stochastic Factors

Number of Iterations....: 1408

<table>
<thead>
<tr>
<th></th>
<th>(scaled)</th>
<th>(unscaled)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Objective...........</td>
<td>2.8227071452245087e-06</td>
<td>3.5408038429696236e-04</td>
</tr>
<tr>
<td>Dual infeasibility..</td>
<td>5.2126351172678061e-07</td>
<td>6.5387294911007360e-05</td>
</tr>
<tr>
<td>Constraint violation.</td>
<td>5.2750692702829838e-11</td>
<td>5.2750692702829838e-11</td>
</tr>
<tr>
<td>Complementarity......</td>
<td>2.5059035596800618e-09</td>
<td>3.143405425262692e-07</td>
</tr>
<tr>
<td>Overall NLP error....</td>
<td>5.2126351172678061e-07</td>
<td>6.5387294911007360e-05</td>
</tr>
</tbody>
</table>

Number of objective function evaluations = 1497
Number of objective gradient evaluations = 1225
Number of equality constraint evaluations = 1497
Number of inequality constraint evaluations = 0
Number of equality constraint Jacobian evaluations = 1410
Number of inequality constraint Jacobian evaluations = 0
Number of Lagrangian Hessian evaluations = 1408
Total CPU secs in IPOPT (w/o function evaluations) = 2.643
Total CPU secs in NLP function evaluations = 0.391

EXIT: Solved To Acceptable Level.

A = [
    0.11 0.09 0.06
    0.11 0.10 0.11
    0.05 0.09 0.08
    0.07 0.09 0.09
    0.12 0.08 0.09
    0.11 0.06 0.08
    0.11 0.12 0.11
    0.10 0.07 0.09
    0.09 0.12 0.14
    0.07 0.12 0.11
];

R = [
    0.68 0.10 0.21
    0.65 0.03 0.30
    0.78 0.03 0.18
];

B = [
    12093.96 69270.78 5427.89
    8023.23 46380.53 30469.32
];
<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>9240.90</td>
<td>20222.90</td>
<td>16208.16</td>
<td></td>
</tr>
<tr>
<td>8600.25</td>
<td>30810.27</td>
<td>19938.38</td>
<td></td>
</tr>
<tr>
<td>8018.27</td>
<td>62998.07</td>
<td>22998.48</td>
<td></td>
</tr>
<tr>
<td>5915.20</td>
<td>63186.25</td>
<td>21245.36</td>
<td></td>
</tr>
<tr>
<td>12865.90</td>
<td>52032.96</td>
<td>22772.32</td>
<td></td>
</tr>
<tr>
<td>5216.88</td>
<td>52171.48</td>
<td>25643.13</td>
<td></td>
</tr>
<tr>
<td>8028.80</td>
<td>29762.18</td>
<td>37938.46</td>
<td></td>
</tr>
<tr>
<td>11408.23</td>
<td>25016.69</td>
<td>24760.31</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
DISTRIBUTION:

1  MS 0899    Technical Library, 9536 (electronic copy)