Factorizing Complex Discrete Data "with Finesse" – Supplementary Material

September 22, 2016

1 OMF Synthetic Experiments (dimensions $n = 50, m = 50$)

Figure 1 shows results for Ordinal Matrix Factorization (OMF) experiments on *small* synthetic data, primarily done in order to enable a comparison with GreEss (which is unable to be executed on the larger data in the paper due to a prohibitive run-time complexity). Here we generate synthetic data with $n = m = 50$; otherwise the parameters are identical to those used in the paper. We note that GreEss has an advantage in that it permits its "usage" matrix to contain entries over the full ordinal scale (i.e. it has more flexibility in modeling the data). Despite this, it only manages to outperform Finesse for the zero-noise case.

The k plot is worth discussing. Typically, as k increases, the reconstruction becomes more complicated (we see this with the Asso and PaNDa+ results). This time, however, we have data set dimensions of $n = m = 50$, which means that, for $k = 50$, there are as many patterns as observations. In this case Finesse is simply able to "cheat" to achieve a perfect decomposition (in terms of reconstruction error) because its initial basis matrix will always simply be a permutation of the data matrix. It is then left only with the trivial task of finding the "usage" matrix which represents the permutation (i.e. each row of the usage matrix will just have a single $\mathfrak{t}$ value in a different column). Such a case is only of interest for those curious as to the operation of Finesse – practically k is chosen much lower than n and m. The GreEss result in the k plot is also noteworthy: given the example in the paper, we can appreciate that its results improve as we continue to take more of its basis vectors. However, we also see that it requires *more* than 50 basis vectors to precisely reproduce the data sets at $k = 50$ (i.e. it still has a reasonable error for the case where there are as many patterns as observations).

Finally, we include results from Non-negative Matrix Factorization (NMF) here for reference, noting that it is a technique from linear algebra and uses the classical matrix product. The factors it produces are thus not the kind we are looking for (they lack the interpretation benefits), so its results here should be taken with a "grain of salt".

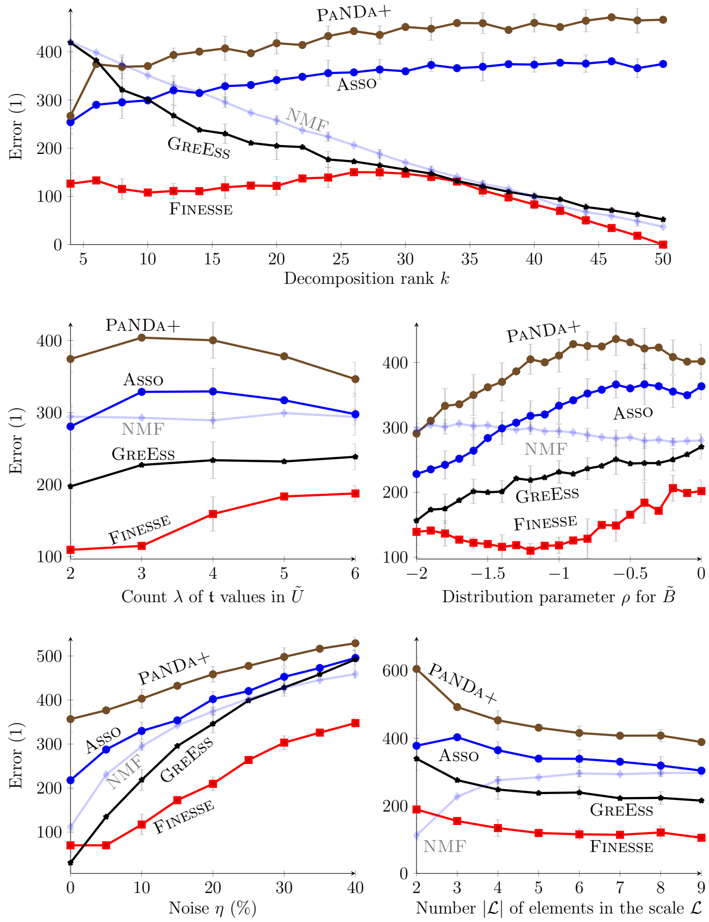

Figure 1: *Ordinal Matrix Factorization for small synthetic data ($n = m = 50$) with varying data-generation parameters k, λ, ρ, η and $|\mathcal{L}|$.*

2 BMF Synthetic Experiments

In Figure 2 we present a comparison between FINESSE, ASSO and PANDA+ on Boolean Matrix Factorization (BMF) problems. Here the generative model is similar to that for the ordinal experiments in the paper. Specifically, we generate "ground-truth" usage $\tilde{U}$ and basis $\tilde{B}$ matrices of order k (i.e. k columns in $\tilde{U}$, k rows in $\tilde{B}$). For the usage matrix $\tilde{U}$ we have the parameter λ, which represents the average number of true values $\mathfrak{t}$ per matrix row. For the basis matrix we have the parameter ρ which again defines the probability mass function (in the same way as in the paper, just with two scale entries this time), controlling the distribution of entries in $\tilde{B}$. That is, smaller values of ρ yield fewer $\mathfrak{t}$ values in $\tilde{B}$ and hence a sparser data matrix $\tilde{D} = \tilde{U} \odot \tilde{B}$; larger values of ρ yield more $\mathfrak{t}$ values in $\tilde{B}$ and hence a denser data matrix. As described in the paper, the data matrix D actually provided to each algorithm (along with the parameter k) is produced by replacing η% of the entries in $\tilde{D}$ with a uniform random choice from $\{\mathfrak{f}, \mathfrak{t}\}$ (note that this will leave the entry unchanged in half of the cases on average, so the "effective" noise percentage is approximately $\eta/2$).

We choose the same defaults of $n = 8000$, $m = 100$, $k = 16$, $l = 3$ and $\eta = 10\%$. The default for ρ is again chosen based on the goal of having a default data matrix with an *equal* distribution of $\mathfrak{t}$ and $\mathfrak{f}$ values. Our default of $\rho = -1.8$ achieves this.

Overall we see similar behavior to that for the ordinal experiments in the paper. The error tends to increase for larger k (more complicated reconstruction) and larger noise amounts η. ASSO and PANDA+ seem to prefer sparser data matrices (smaller ρ value), whereas FINESSE seems to have optimal performance around $\rho = -1.5$, where there are slightly more $\mathfrak{t}$ than $\mathfrak{f}$ values on average in the data matrix D.

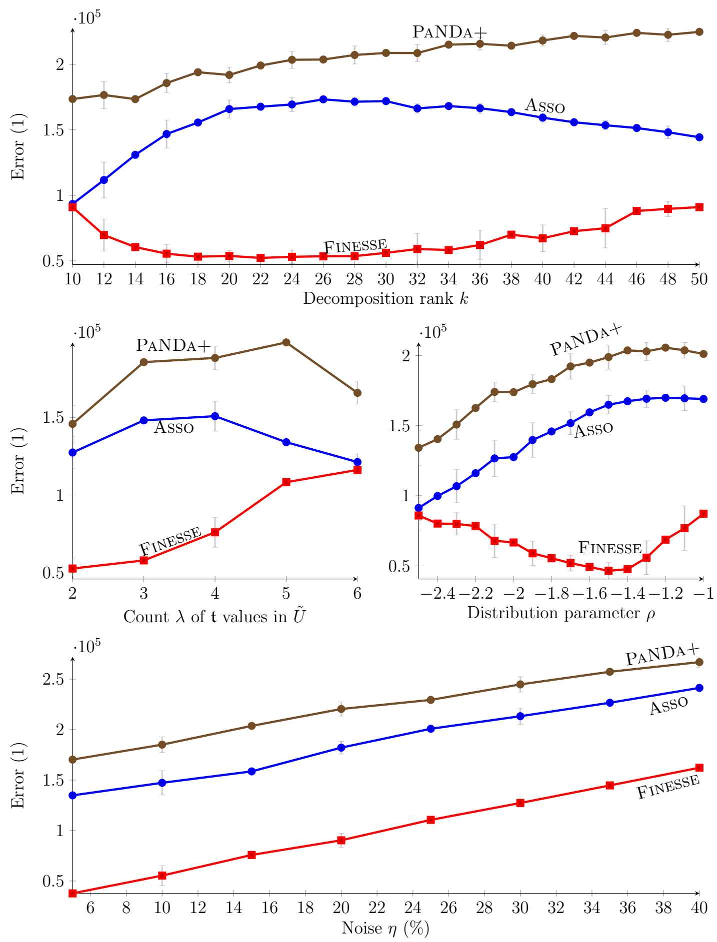

Figure 2: *Binary (Boolean) Matrix Factorization for synthetic data with varying data-generation parameters k, λ, ρ and η.*

3 Model-order selection for Yummly data

We note that the ontology used was not the *official* Yummly version (which is proprietary and not publicly-available). It was instead constructed by hand[1] from a tractable number (693) of ingredients existing in the set of 296 matches that were returned from an API-search for main-course recipes of American cuisine and tagged with the holiday *summer*.

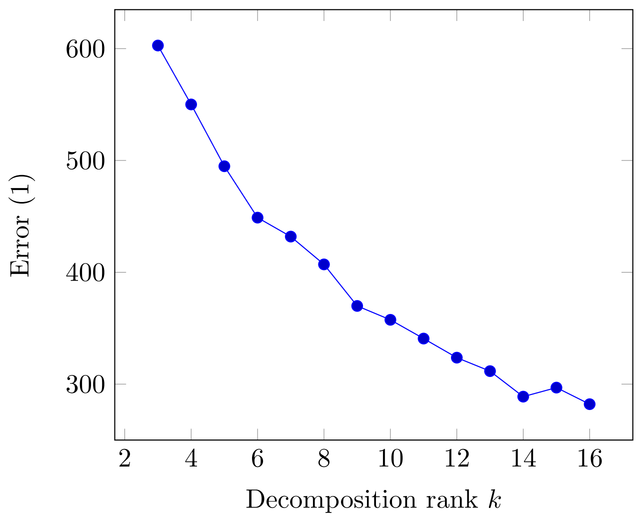

Figure 3: *Model-order selection for Yummly data. The "elbow" point is seen at $k = 6$.*

[1]We thank the student Georg Kastner for this work.

4 Model-order selection for IMDB data

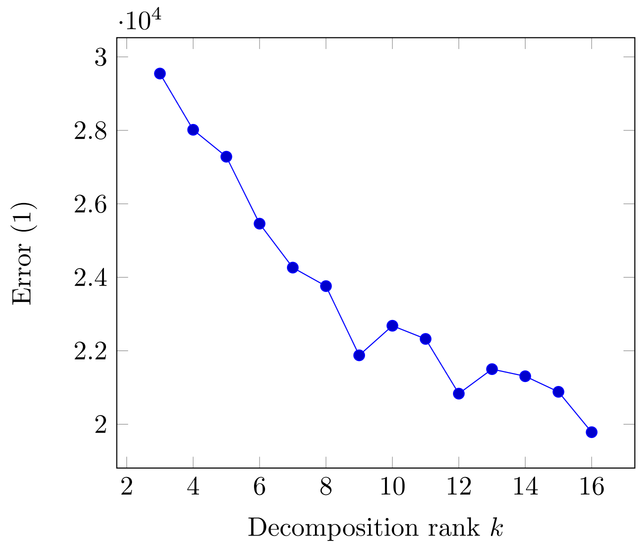

Figure 4: *Model-order selection for IMDB data. The "elbow" point is seen at* $k = 9$*.*