# NEWTON-BASED OPTIMIZATION FOR KULLBACK-LEIBLER NONNEGATIVE TENSOR FACTORIZATIONS 

SAMANTHA HANSEN, TODD PLANTENGA, TAMARA G. KOLDA


#### Abstract

Tensor factorizations with nonnegative constraints have found application in analyzing data from cyber traffic, social networks, and other areas. We consider application data best described as being generated by a Poisson process (e.g., count data), which leads to sparse tensors that can be modeled by sparse factor matrices. In this paper we investigate efficient techniques for computing an appropriate canonical polyadic tensor factorization based on the Kullback-Leibler divergence function. We propose novel subproblem solvers within the standard alternating block variable approach. Our new methods exploit structure and reformulate the optimization problem as small independent subproblems. We employ bound-constrained Newton and quasi-Newton methods. We compare our algorithms against other codes, demonstrating superior speed for high accuracy results and the ability to quickly find sparse solutions.


1. Introduction. Multilinear models have proved useful in analyzing data in a variety of fields. We focus on data that derives from a Poisson process, such as the number of packets sent from one IP address to another on a specific port [36], the number of papers published by an author at a given conference [15], or the count of emails between users in a given time period 3. Data in these applications is nonnegative and often quite sparse, i.e., most tensor elements have a count of zero. The tensor factorization model corresponding to such sparse count data is computed from a nonlinear optimization problem that minimizes the Kullback-Leibler (K-L) divergence function and contains nonnegativity constraints on all variables.

In this paper we show how to make second-order optimization methods suitable for Poisson-based tensor models of large sparse count data. Multiplicative update is one of the most widely implemented methods for this model, but it suffers from slow convergence and inaccuracy in discovering the underlying sparsity. In large sparse tensors, the application of nonlinear optimization techniques requires consideration of sparsity and problem structure to get better performance. We show that, by exploiting the partial separability of the subproblems, we can successfully apply second-order methods. We develop algorithms that scale to large sparse tensor applications and are quick in identifying sparsity in the factors of the model.

There is a need for second-order methods because computing factor matrices to high accuracy, as measured by satisfaction of the first-order KKT conditions, is effective in revealing sparsity. By contrast, multiplicative update methods can make elements small but are slow to reach the variable bound at zero, forcing the user to guess when "small" means zero. We demonstrate that guessing a threshold is inherently difficult, making the high accuracy obtained with second-order methods desirable.

We start from a standard Gauss-Seidel alternating block framework and show that each block subproblem is further separable into a set of independent functions, each of which depends on only a subset of variables. We optimize each subset of variables independently, an obvious idea which has nevertheless not previously appeared in the setting of sparse tensors. We call this a row subproblem formulation because the subset of variables corresponds to one row of a factor matrix. Each row subproblem amounts to minimizing a strictly convex function with nonnegativity constraints, which we solve using two-metric gradient projection techniques and exact or approximate second-order information.

The importance of the row subproblem formulation is demonstrated in Section 4.1,
where we show that applying a second-order method directly to the block subproblem is highly inefficient. We provide evidence that a more effective way to apply secondorder methods is through the use of the row subproblem formulation.

Our contributions in this paper are as follows:

1. A new formulation for nonnegative tensor factorization based on the KullbackLeibler divergence objective that allows for the effective use of second-order optimization methods. The optimization problem is separated into row subproblems containing $R$ variables, where $R$ is the number of factors in the model. The formulation makes row subproblems independent, suggesting a parallel method, although we do not explore parallelism in this paper.
2. Two Matlab algorithms for computing factorizations of sparse nonnegative tensors: one using second derivatives and the other using limited-memory quasi-Newton approximations. The algorithms are made robust with an Armijo line search, damping modifications when the Hessian is ill conditioned, and projections to the bound of zero based on two-metric gradient projection ideas in [6]. The two algorithms have different computational costs: the second derivative method is preferred when $R$ is small, and the quasi-Newton when $R$ is large.
3. Test results that compare the performance of our two new algorithms with the best available multiplicative update method and a related quasi-Newton algorithm that does not formulate using row subproblems. The most significant test results are reported in this paper; detailed results of all experiments are available in the supplementary material (Appendix B).
4. Test results showing the ability of our methods to quickly and accurately determine which elements of the factorization model are zero without using problem-specific thresholds.
The paper is outlined as follows: the remainder of Section 1 surveys related work and provides a review of basic tensor properties. Section 2 formalizes the Poisson nonnegative tensor factorization optimization problem, shows how the Gauss-Seidel alternating block framework can be applied, and converts the block subproblem into independent row subproblems. Section 3 outlines two algorithms for solving the row subproblem, one based on the damped Hessian (PDN-R for projected damped Newton), and one based on a limited-memory approximation (PQN-R for projected quasiNewton). Section 4 details numerical results on synthetic and real data sets and quantifies the accuracy of finding a truly sparse factorization. Additional test results are available in the supplementary material. Section 5 contains a summary of the paper and concluding remarks.
1.1. Related Work. In this paper, we specifically consider nonnegative tensor factorization (NTF) in the case of the canonical polyadic (also known as CANDECOMP/PARAFAC) tensor decomposition. Our focus is on the K-L divergence objective function, but we also mention related work for the least squares (LS) case. Additionally, we consider related work for nonnegative matrix factorization (NMF) for both K-L and LS. Note that there is much more work in the LS case, but the K-L objective function is different enough that it deserves its own attention. We do not discuss other decompositions such as Tucker.

NMF in the LS case was first proposed by Paatero and Tapper 32 and also studied by Bro [7, p. 169]. Lee and Seung later consider the problem for both LS and K-L formulations and introduce multiplicative updates based on the convex subproblems [25, 26]. Their work is extended to tensors by Welling and Weber 39]. Many other
works have been published on the LS versions of NMF [27, 21, 22, 32, 44] and NTF [8, 12, 16, 23, 40.

Lee and Seung's multiplicative update method [25, 26, 39] is the basis for most NTF algorithms that minimize the K-L divergence function. Chi and Kolda provide an improved multiplicative update scheme for K-L that addresses performance and convergence issues as elements approach zero [11; we compare to their method in Section 4. By interpreting the K-L divergence as an alternative Csiszar-Tusnady procedure, Zafeiriou and Petrou [41] provide a probabilistic interpretation of NTF along with a new multiplicative update scheme. The multiplicative update is equivalent to a scaled steepest-descent step [26, so it is a first-order optimization method. Since our method uses second-order information, it allows for convergence to higher accuracy and a better determination of sparsity in the factorization.

Second-order information has been used before in connection with the K-L objective. Zdunek and Cichocki [42, 43] propose a hybrid method for blind source separation applications via NMF that uses a damped Hessian method similar to ours. They recognize that the Hessian of the K-L objective has a block diagonal structure but do not reformulate the optimization problem further as we do. Consequently, their Hessian matrix is large, and they switch to the LS objective function for the larger mode of the matrix because their Newton method cannot scale up. Mixing objective functions in this manner is undesirable because it combines two different underlying models. As a point of comparison, a problem in 43] of size $200 \times 1000$ is considered too large for their Newton method, but our algorithms can factor a data set of this size with $R=50$ components to high accuracy in less than ten minutes (see the supplementary material). The Hessian-based method in 43 has most of the advanced optimization features that we use (though details differ), including an Armijo line search, active set identification, and an adjustable Hessian damping factor. We also note that Zheng and Zhang [44 compute a damped Hessian search direction and find an iterate with a backtracking line search, though this work is for the LS objective in NMF.

Recently, Hsiel and Dhillon 19 reported algorithms for NMF with both LS and K-L objectives. Their method updates one variable at a time, solving a nonlinear scalar function using Newton's method with a constant step size. They achieve good performance for the LS objective by taking the variables in a particular order based on gradient information; however, for the more complex K-L objective, they must cycle through all the variables one by one. Our algorithms solve convex row subproblems with $R$ variables using second-order information; solving these subproblems one variable at a time by coordinate descent will likely have a much slower rate of convergence [30, pp. 230-231].

A row subproblem reformulation similar to ours is noted in earlier papers exploring the LS objective, but it never led to Hessian-based methods that exploit sparsity as ours do. Gonzales and Zhang use the reformulation with a multiplicative update method for NMF [17] but do not generalize to tensors or the K-L objective. Phan et al. [33] note the reformulation is suitable for parallelizing a Hessian-based method for NTF using LS. Kim and Park use the reformulation for NTF with LS [23, deriving small bound-constrained LS subproblems. Their method solves the LS subproblems by exact matrix factorization, without exploiting sparsity, and features a block principal pivoting method for choosing the active set. Other works solve the LS objective by taking advantage of row-by-row or column-by-column subproblem decomposition [12, 28, 34].

Our algorithms are similar in spirit to the work of Kim, Sra and Dhillon [20], which applies a projected quasi-Newton algorithm (called PQN in this paper) to solving NMF with a K-L objective. Like PQN, our algorithms identify active variables, compute a Newton-like direction in the space of free variables, and find a new iterate using a projected backtracking line search. We differ from PQN in reformulating the subproblem and in computing a damped Newton direction; both improvements make a huge difference in performance for large-scale tensor problems. We compare to PQN in Section 4

All-at-once optimization methods, including Hessian-based algorithms, have been applied to NTF with the LS objective function. As an example, Paatero replaces the nonnegativity constraints with a barrier function [31] to yield an unconstrained optimization problem, and Phan, Tichavsky and Cichocki 34 apply a fast damped Gauss-Newton algorithm for minimizing a similar penalized objective. We are not aware of any work on all-at-once methods for the K-L objective in NTF.

Finally, we note that all methods, including ours, find only a locally optimal solution to the NTF problem. Finding the global solution is generally much harder; for instance, Vavasis 38 proves it is NP-hard for an NMF model that fits the data exactly.
1.2. Tensor Review. For a thorough introduction to tensors, see [24] and references therein; we only review concepts that are necessary for understanding this paper. A tensor is a multidimensional array. An $N$-way tensor $\mathcal{X}$ has size $I_{1} \times I_{2} \times \ldots \times I_{N}$. To differentiate between tensors, matrices, vectors, and scalars, we use the following notational convention: $\mathcal{X}$ is a tensor (bold, capitalized, calligraphic), $\mathbf{X}$ is a matrix (bold, capitalized), $\mathbf{x}$ is a vector (bold, lowercase), and $x$ is a scalar (lowercase). Additionally, given a matrix $\mathbf{X}, \mathbf{x}_{j}$ denotes its $j$ th column and $\hat{\mathbf{x}}_{i}$ denotes its $i$ th row.

Just as a matrix can be decomposed into a sum of outer products between two vectors, an $N$-way tensor can be decomposed into a sum of outer products between $N$ vectors. Each of these outer products (called components) yields an $N$-way tensor of rank one. The CP (CANDECOMP/PARAFAC) decomposition [10, 18 represents a tensor as a sum of rank-one tensors (see Figure 1.1):

$$
\begin{equation*}
\boldsymbol{X} \approx \llbracket \boldsymbol{\lambda} ; \mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(N)} \rrbracket=\sum_{r=1}^{R} \lambda_{r} \mathbf{a}_{r}^{(1)} \circ \ldots \circ \mathbf{a}_{r}^{(N)} \tag{1.1}
\end{equation*}
$$

where $\boldsymbol{\lambda}$ is a vector and each $\mathbf{A}^{(n)}$ is an $I_{n} \times R$ factor matrix containing the $R$ vectors contributed to the outer products by mode $n$, i.e.,

$$
\begin{equation*}
\mathbf{A}^{(n)}=\left[\mathbf{a}_{1}^{(n)} \cdots \mathbf{a}_{R}^{(n)}\right] \tag{1.2}
\end{equation*}
$$

Equality holds in (1.1) when $R$ equals the rank of $\mathcal{X}$, but often a tensor is approximated by a smaller number of terms. We let $\mathbf{i}$ denote the multi-index $\left(i_{1}, i_{2}, \ldots, i_{N}\right)$ of an element $x_{\mathbf{i}}$ of $\mathcal{X}$.

We use of the idea of matricization, or unfolding a tensor into a matrix. Specifically, unfolding along mode $n$ yields a matrix of size $I_{n} \cdot J_{n}$, where

$$
J_{n}=I_{1} \cdot I_{2} \cdots I_{n-1} \cdot I_{n+1} \cdots I_{N-1} \cdot I_{N}
$$

We use the notation $\mathbf{X}_{(n)}$ to represent a tensor $\mathcal{X}$ that has been unfolded so that its $n$th mode forms the rows of the matrix, and $x_{i j}^{(n)}$ for its $(i, j)$ element. If a tensor $\mathcal{X}$


Fig. 1.1: CANDECOMP/PARAFAC decomposition of a three-way tensor into $R$ components.
is written in Kruskal form (1.1), then the mode- $n$ matricization is given by

$$
\begin{equation*}
\mathbf{X}_{(n)}=\mathbf{A}^{(n)} \boldsymbol{\Lambda}\left(\mathbf{A}^{(N)} \odot \ldots \odot \mathbf{A}^{(n+1)} \odot \mathbf{A}^{(n-1)} \odot \ldots \odot \mathbf{A}^{(1)}\right)^{T} \tag{1.3}
\end{equation*}
$$

where $\boldsymbol{\Lambda}=\operatorname{diag}(\boldsymbol{\lambda})$ and $\odot$ denotes the Khatri-Rao product [24].
Tensor results are generally easier to interpret when the factors 1.2 are sparse. Moreover, many sparse count applications can reasonably expect sparsity in the factors. For example, the 3-way data considered in [15] counts publications by authors at various conferences over a ten year period. The tensor representation has a sparsity of $0.14 \%$ (only $0.14 \%$ of the data elements are nonzero), and the factors computed by our algorithm with $R=60$ (see Section 4.2 ) have sparsity $9.3 \%, 2.7 \%$, and $77.5 \%$ over the three modes. One meaning of sparsity in the factors is to say that a typical outer product term connects about $9 \%$ of the authors with $3 \%$ of the conferences in 8 of the 10 years. Linking particular authors and conferences is an important outcome of the tensor analysis, requiring clear distinction between zero and nonzero elements in the factors.
2. Poisson Nonnegative Tensor Factorization. In this section we state the optimization problem, examine its structure, and show how to separate it into simpler subproblems.
2.1. Gauss-Seidel Alternating Block Formulation. We seek a tensor model in CP form to approximate data $\mathcal{X}$ :

$$
\boldsymbol{X} \approx \mathcal{M}=\llbracket \boldsymbol{\lambda} ; \mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(N)} \rrbracket=\sum_{r=1}^{R} \lambda_{r} \mathbf{a}_{r}^{(1)} \circ \ldots \circ \mathbf{a}_{r}^{(N)}
$$

The value of $R$ is chosen empirically, and the scaling vector $\boldsymbol{\lambda}$ and factor matrices $\mathbf{A}^{(n)}$ are the model parameters that we compute.

In [11], it is shown that a K-L divergence objective function results when data elements follow Poisson distributions with multilinear parameters. The best-fitting tensor model under this assumption satisfies:

$$
\begin{align*}
\min _{\boldsymbol{\lambda}, \mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(N)}} & f(\mathbf{M})=\sum_{\mathbf{i}} m_{\mathbf{i}}-x_{\mathbf{i}} \log m_{\mathbf{i}} \\
\text { s.t. } & \mathcal{M}=\sum_{r=1}^{R} \lambda_{r} \mathbf{a}_{r}^{(1)} \circ \ldots \circ \mathbf{a}_{r}^{(N)},  \tag{2.1}\\
& \lambda_{r} \geq 0, \mathbf{a}_{r}^{(n)} \geq 0,\left\|\mathbf{a}_{r}^{(n)}\right\|_{1}=1, \quad \forall r \in\{1, \ldots R\}, \forall n \in\{1, \ldots N\},
\end{align*}
$$

where $x_{\mathbf{i}}$ denotes element $\left(i_{1}, \ldots, i_{n}\right)$ of tensor $\mathcal{X}$ and $m_{\mathbf{i}}$ denotes element $\left(i_{1}, \ldots, i_{n}\right)$ of the model $\mathcal{M}$. The model may have terms where $m_{\mathbf{i}}=0$ and $x_{\mathbf{i}}=0$; for this case we define $0 \log 0=0$. Note that for matrix factorization, (2.1) reduces to the K-L divergence used by Lee and Seung [25, 26]. The constraint that normalizes the column sum of the factor matrices serves to remove an inherent scaling ambiguity in the CP factor model.

As in [11], we unfold $\mathcal{X}$ and $\mathcal{M}$ into their $n$th matricized mode, and use (1.3) to express the objective as

$$
f(\mathcal{M})=\mathbf{e}^{T}\left[\mathbf{A}^{(n)} \boldsymbol{\Lambda} \boldsymbol{\Pi}^{(n)}-\mathbf{X}_{(n)} * \log \left(\mathbf{A}^{(n)} \boldsymbol{\Lambda} \boldsymbol{\Pi}^{(n)}\right)\right] \mathbf{e}
$$

where $\mathbf{e}$ is a vector of all ones, the operator $*$ denotes elementwise multiplication, $\log (\cdot)$ is taken elementwise,

$$
\begin{align*}
\boldsymbol{\Pi}^{(n)} & =\left(\mathbf{A}^{(N)} \odot \ldots \odot \mathbf{A}^{(n+1)} \odot \mathbf{A}^{(n-1)} \odot \ldots \odot \mathbf{A}^{(1)}\right)^{T} \in \mathbf{R}^{R \times J_{n}}, \text { and }  \tag{2.2}\\
\boldsymbol{\Lambda} & =\operatorname{diag}(\boldsymbol{\lambda}) \in \mathbf{R}^{R \times R}
\end{align*}
$$

Note that by expanding the Khatri-Rao products in 2.2 and remembering that column vectors $\mathbf{a}_{r}^{(n)}$ are normalized, each row of $\boldsymbol{\Pi}^{(n)}$ conveniently sums to one. This is a consequence of using the $\ell_{1}$ norm in (2.1).

The above representation of the objective motivates the use of an alternating block optimization method where only one factor matrix is optimized at a time. Holding the other factor matrices fixed, the optimization problem for $\mathbf{A}^{(n)}$ and $\boldsymbol{\Lambda}$ is

$$
\begin{align*}
\min _{\boldsymbol{\Lambda}, \mathbf{A}^{(n)}} & f\left(\boldsymbol{\Lambda}, \mathbf{A}^{(n)}\right)=\mathbf{e}^{T}\left[\mathbf{A}^{(n)} \boldsymbol{\Lambda} \mathbf{\Pi}^{(n)}-\mathbf{X}_{(n)} * \log \left(\mathbf{A}^{(n)} \boldsymbol{\Lambda} \mathbf{\Pi}^{(n)}\right)\right] \mathbf{e}  \tag{2.3}\\
\text { s.t. } & \boldsymbol{\Lambda} \geq 0, \mathbf{A}^{(n)} \geq 0, \mathbf{e}^{T} \mathbf{A}^{(n)}=\mathbf{1}
\end{align*}
$$

Problem 2.3 is not convex. However, ignoring the equality constraint and letting $\mathbf{B}^{(n)}=\mathbf{A}^{(n)} \boldsymbol{\Lambda}$, we have

$$
\begin{equation*}
\min _{\mathbf{B}^{(n)} \geq 0} f\left(\mathbf{B}^{(n)}\right)=\mathbf{e}^{T}\left[\mathbf{B}^{(n)} \boldsymbol{\Pi}^{(n)}-\mathbf{X}_{(n)} * \log \left(\mathbf{B}^{(n)} \boldsymbol{\Pi}^{(n)}\right)\right] \mathbf{e} \tag{2.4}
\end{equation*}
$$

which is convex with respect to $\mathbf{B}^{(n)}$. The two formulations are equivalent in that a KKT point of (2.4) can be used to find a KKT point of (2.3). Chi and Kolda show in [11] that (2.4) is strictly convex given certain assumptions on the sparsity pattern of $\mathbf{X}_{(n)}$.

We pause to think about (2.4) when the tensor is two-way. In this case, we solve for two factor matrices by alternating over two block subproblems; for instance, with $n=1$ the subproblem 2.4 finds $\mathbf{B}^{(1)}$ with $\boldsymbol{\Pi}^{(1)}=\left(\mathbf{A}^{(2)}\right)^{T}$. For an $N$-way problem, the only change to 2.4 is $\boldsymbol{\Pi}^{(n)}$, which grows in size exponentially with each additional factor matrix. To efficiently solve the subproblems 2.4 for large sparse tensors we can exploit sparsity to reduce the computational costs. As discussed in the next section, columns of $\boldsymbol{\Pi}^{(n)}$ need to be computed only when the corresponding column in the unfolded tensor $\mathbf{X}_{(n)}$ has a nonzero element. Our row subproblem formulation carries this idea further because each row subproblem generally uses only a fraction of the nonzero elements in $\mathbf{X}_{(n)}$.

At this point we define Algorithm 1, a Gauss-Seidel alternating block method. The algorithm iterates over each mode of the tensor, solving the convex optimization block subproblem. Steps 6 and 7 rescale the factor matrix columns, redistributing the

```
Algorithm 1 Alternating Block Framework
Given data tensor \(\boldsymbol{X}\) of size \(I_{1} \times I_{2} \times \cdots \times I_{N}\), and the number of components \(R\)
Return a model \(\mathcal{M}=\left[\boldsymbol{\lambda} ; \mathbf{A}^{(1)} \ldots \mathbf{A}^{(N)}\right]\)
    Initialize \(\mathbf{A}^{(n)} \in \mathbf{R}^{I_{n} \times R}\) for \(n=1, \ldots, N\)
    repeat
        for \(n=1, \ldots, N\) do
            Let \(\boldsymbol{\Pi}^{(n)}=\left(\mathbf{A}^{(N)} \odot \ldots \odot \mathbf{A}^{(n+1)} \odot \mathbf{A}^{(n-1)} \odot \ldots \odot \mathbf{A}^{(1)}\right)^{T}\)
            Use Algorithm 2 to compute \(\mathbf{B}^{*}\) that minimizes \(f\left(\mathbf{B}^{(n)}\right)\) s.t. \(\mathbf{B}^{(n)} \geq 0\)
            \(\boldsymbol{\lambda} \leftarrow \mathrm{e}^{T} \mathbf{B}^{*}\)
            \(\mathbf{A}^{(n)} \leftarrow \mathbf{B}^{*} \boldsymbol{\Lambda}^{-1}\), where \(\boldsymbol{\Lambda}=\operatorname{diag}(\boldsymbol{\lambda})\)
        end for
    until all mode subproblems have converged
```

weight into $\boldsymbol{\lambda}$. For the moment, we leave the subproblem solution method in Step 5 unspecified. A proof that Algorithm 1 convergences to a local minimum of (2.1) is given in 11.

This outline of Algorithm 1 corresponds exactly with the method proposed in [11]; where we differ is in how to solve subproblem (2.4) in Step 5. Note also that this algorithm is the same as for the least squares objective (references were given in Section 1.1); there the subproblem in Step 5 is replaced by a linear least squares subproblem. We now proceed to describe our method for solving (2.4).
2.2. Row Subproblem Reformulation. We examine the objective function $f\left(\mathbf{B}^{(n)}\right)$ in 2.4 and show that it can be reformulated into independent functions. As mentioned in the previous section, rows of $\Pi^{(n)}$ sum to one if the columns of factor matrices are nonnegative and sum to one. When $\Pi^{(n)}$ is formed at Step 4 of Algorithm 1 the factor matrices satisfy these conditions by virtue of Steps 6 and 7 hence, the first term of $f\left(\mathbf{B}^{(n)}\right)$ is

$$
\mathbf{e}^{T} \mathbf{B}^{(n)} \boldsymbol{\Pi}^{(n)} \mathbf{e}=\mathbf{e}^{T} \mathbf{B}^{(n)} \mathbf{e}=\sum_{i=1}^{I_{n}} \sum_{r=1}^{R} b_{i r}^{(n)} .
$$

The second term of $f\left(\mathbf{B}^{(n)}\right)$ is a sum of elements from the $I_{n} \times J_{n}$ matrix $\mathbf{X}_{(n)} *$ $\log \left(\mathbf{B}^{(n)} \boldsymbol{\Pi}^{(n)}\right)$. Recall that the operations in this expression are elementwise, so the scalar matrix element $(i, j)$ of the term can be written as

$$
x_{i j}^{(n)} \log \left(\sum_{r=1}^{R} b_{i r}^{(n)} \pi_{r j}^{(n)}\right) .
$$

Adding all the elements and combining with the first term gives

$$
\begin{aligned}
f\left(\mathbf{B}^{(n)}\right) & =\sum_{i=1}^{I_{n}} \sum_{r=1}^{R} b_{i r}^{(n)}-\sum_{i=1}^{I_{n}} \sum_{j=1}^{J_{n}} x_{i j}^{(n)} \log \left(\sum_{r=1}^{R} b_{i r}^{(n)} \pi_{r j}^{(n)}\right) \\
& =\sum_{i=1}^{I_{n}} f_{\mathrm{row}}\left(\hat{\mathbf{b}}_{i}^{(n)}, \hat{\mathbf{x}}_{i}^{(n)}, \boldsymbol{\Pi}^{(n)}\right) .
\end{aligned}
$$

where $\hat{\mathbf{b}}_{i}$ and $\hat{\mathbf{x}}_{i}$ are the $i$ th row vectors of their corresponding matrices, and

$$
\begin{equation*}
f_{\text {row }}(\hat{\mathbf{b}}, \hat{\mathbf{x}}, \boldsymbol{\Pi})=\sum_{r=1}^{R} \hat{b}_{r}-\sum_{j=1}^{J_{n}} \hat{x}_{j} \log \left(\sum_{r=1}^{R} \hat{b}_{r} \pi_{r j}\right) \tag{2.5}
\end{equation*}
$$

Problem 2.4 can now be rewritten as

$$
\begin{equation*}
\min _{\hat{\mathbf{b}}_{1}, \ldots, \hat{\mathbf{b}}_{I_{n}} \geq 0} \sum_{i=1}^{I_{n}} f_{\text {row }}\left(\hat{\mathbf{b}}_{i}^{(n)}, \hat{\mathbf{x}}_{i}^{(n)}, \boldsymbol{\Pi}^{(n)}\right) . \tag{2.6}
\end{equation*}
$$

This is a completely separable set of $I_{n}$ row subproblems, each one a convex nonlinear optimization problem containing $R$ variables. The relatively small number of variables makes second-order optimization methods tractable, and that is the direction we pursue in this paper. Algorithm 2 describes how the reformulation fits into Algorithm 1.

```
Algorithm 2 Row Subproblem Framework for Solving (2.6)
Given \(\mathbf{X}_{(n)}\) of size \(I_{n} \times J_{n}\), and \(\boldsymbol{\Pi}^{(n)}\) of size \(R \times J_{n}\)
Return a solution \(\mathbf{B}^{*}\) consisting of row vectors \(\hat{\mathbf{b}}_{1}^{*}, \ldots, \hat{\mathbf{b}}_{I_{N}}^{*}\)
    for \(i=1, \ldots, I_{n}\) do
        Select row \(\hat{\mathbf{x}}_{i}\) of \(\mathbf{X}_{(n)}\)
        Generate one column of \(\boldsymbol{\Pi}^{(n)}\) for each nonzero in \(\hat{\mathbf{x}}_{i}\)
        Use Algorithm 3 or 4 to compute \(\hat{\mathbf{b}}_{i}^{*}\) that solves
            \(\min f_{\text {row }}\left(\hat{\mathbf{b}}_{i}^{(n)}, \hat{\mathbf{x}}_{i}^{(n)}, \boldsymbol{\Pi}^{(n)}\right)\) subject to \(\hat{\mathbf{b}}_{i} \geq 0\)
    end for
```

The independence of row subproblems is crucial for handling large tensors. For example, if a three-way tensor of size $1000 \times 1000 \times 1000$ is factored into $R=100$ components, then $\boldsymbol{\Pi}^{(n)}$ is a $100 \times 10^{6}$ matrix. However, elements of $\boldsymbol{\Pi}^{(n)}$ appear in the optimization objective only where the matricized tensor $\mathbf{X}_{(n)}$ has nonzero elements, so in a sparse tensor many columns of $\boldsymbol{\Pi}^{(n)}$ can be ignored; this point was first published in 11. Algorithm 2 exploits this fact in Step 3

Algorithm 2 also points the way to a parallel implementation of the CP tensor factorization. We note, as did 33, that each row subproblem can be run in parallel and storage costs are determined by the sparsity of the data. In a distributed computing architecture, an algorithm could identify the nonzero elements of each row subproblem at the beginning of execution and collect only the data needed to form appropriate columns of $\boldsymbol{\Pi}^{(n)}$ at a given processing element. We do not implement a parallel version of the algorithm in this paper.
3. Solving the Row Subproblem. In this section we show how to solve the row subproblem (2.6) using second-order information. We describe two algorithms, one applying second derivatives in the form of a damped Hessian matrix, and the other using a quasi-Newton approximation of the Hessian. Both algorithms use projection, but the details differ.

Each row subproblem consists of minimizing a strictly convex function of $R$ variables with nonnegativity constraints. One of the most effective methods for solving
bound-constrained problems is second-order gradient projection; see [35]. We employ a form of two-metric gradient projection from Bertsekas [6]. Each variable is marked in one of three states based on its gradient and current location: fixed at its bound of zero, allowed to move in the direction of steepest-descent, or free to move along a Newton or quasi-Newton search direction. Details are in Section 3.1.

An alternative to Bertsekas is to use methods that employ gradient projection searches to determine the active variables (those set to zero). Examples include the generalized Cauchy point 13 and gradient projection along the steepest-descent direction with a line search. We experimented with using the generalized Cauchy point to determine the active variables, but preliminary results indicated that this approach sets too many variables to be at their bound, leading to more iterations and poor overall performance. Gradient projection steps with a line search calls for an extra function evaluation, which is computationally expensive. Given a more efficient method for evaluating the function, this might be a better approach since, under mild conditions, gradient projection methods find the active set in a finite number of iterations [5].

For notational convenience, in this section we use $\mathbf{b}$ for the column vector representation of row vector $\hat{\mathbf{b}}_{i}$; that is, $\mathbf{b}=\hat{\mathbf{b}}_{i}^{T}$. Iterations are denoted with superscript $k$, and $\nabla_{r}$ represents the derivative with respect to the $r$ th variable. Let $P_{+}[\mathbf{v}]$ be the projection operator that restricts each element of vector $\mathbf{v}$ to be nonnegative. We make use of the first and second derivatives of $f_{\text {row }}$, given by

$$
\begin{align*}
\nabla_{r} f_{\text {row }}(\mathbf{b}) & =\frac{\partial f_{\text {row }}(\mathbf{b}, \mathbf{x}, \boldsymbol{\Pi})}{\partial b_{r}}=1-\sum_{j=1}^{J_{n}} \frac{x_{j} \pi_{r j}}{\sum_{i=1}^{R} b_{i} \pi_{i j}}  \tag{3.1}\\
\nabla_{r s}^{2} f_{\text {row }}(\mathbf{b}) & =\frac{\partial^{2} f_{\text {row }}(\mathbf{b}, \mathbf{x}, \boldsymbol{\Pi})}{\partial b_{r} \partial b_{s}}=\sum_{j=1}^{J_{n}} \frac{x_{j} \pi_{r j} \pi_{s j}}{\left(\sum_{i=1}^{R} b_{i} \pi_{i j}\right)^{2}}
\end{align*}
$$

3.1. Two-Metric Projection. At each iteration $k$ we must choose a set of variables to update such that progress is made in decreasing the objective. Bertsekas demonstrated in [6] that iterative updates of the form

$$
\mathbf{b}^{k+1}=P_{+}\left[\mathbf{b}^{k}-\alpha \mathbf{M}^{k} \nabla f_{\text {row }}\left(\mathbf{b}^{k}\right)\right]
$$

are not guaranteed to decrease the objective function unless $\mathbf{M}^{k}$ is a positive diagonal matrix. Instead, it is necessary to predict the variables that have the potential to make progress in decreasing the objective and then update just those variables using a positive definite matrix. We present the two-metric technique of Bertsekas as it is executed in our algorithm, which differs superficially from the presentation in [6].

A variable's potential effect on the objective is determined by how close it is to zero and by its direction of steepest-descent. If a variable is close to zero and its steepest-descent direction points towards the negative orthant, then the next update will likely project the variable to zero and its small displacement will have little effect on the objective. A closeness threshold $\epsilon_{k}$ is computed from a user-defined parameter $\epsilon>0$ as

$$
\begin{equation*}
\epsilon_{k}=\min \left(w_{k}, \epsilon\right), \quad w_{k}=\left\|\mathbf{b}_{9}^{k}-P_{+}\left[\mathbf{b}^{k}-\nabla f_{\mathrm{row}}\left(\mathbf{b}^{k}\right)\right]\right\|_{2} \tag{3.2}
\end{equation*}
$$

We then define index sets

$$
\begin{align*}
& \mathcal{A}\left(\mathbf{b}^{k}\right)=\left\{r \mid b_{r}^{k}=0, \nabla_{r} f_{\text {row }}\left(\mathbf{b}^{k}\right)>0\right\} \\
& \mathcal{G}\left(\mathbf{b}^{k}\right)=\left\{r \mid 0<b_{r}^{k} \leq \epsilon_{k}, \nabla_{r} f_{\text {row }}\left(\mathbf{b}^{k}\right)>0\right\}  \tag{3.3}\\
& \mathcal{F}\left(\mathbf{b}^{k}\right)=\left(\mathcal{A}\left(\mathbf{b}^{k}\right) \cup \mathcal{G}\left(\mathbf{b}^{k}\right)\right)^{c}
\end{align*}
$$

where superscript $c$ denotes the set complement. Variables in the set $\mathcal{A}$ are fixed at zero, variables in $\mathcal{G}$ move in the direction of the negative gradient, and variables in $\mathcal{F}$ are free to move according to second-order information. Note that if $\epsilon=0$ then $\epsilon_{k}=0, \mathcal{G}$ is empty, and the method reduces to defining an active set of variables by instantaneous line search [2].
3.1.1. Damped Newton Step. The damped Newton direction is taken with respect to only the variables in the set $\mathcal{F}$ from (3.3). Let

$$
\mathbf{g}_{F}^{k}=\left[\nabla f_{\text {row }}\left(\mathbf{b}^{k}\right)\right]_{\mathcal{F}}, \mathbf{H}_{F}^{k}=\left[\nabla^{2} f_{\text {row }}\left(\mathbf{b}^{k}\right)\right]_{\mathcal{F}}, \mathbf{b}_{F}^{k}=\left[\mathbf{b}^{k}\right]_{\mathcal{F}}
$$

where $[\mathbf{v}]_{\mathcal{F}}$ chooses the elements of vector $\mathbf{v}$ corresponding to variables in the set $\mathcal{F}$. Since the row subproblems are strictly convex, the full Hessian and $\mathbf{H}_{F}^{k}$ are positive definite.

The damped Hessian has its roots in trust region methods. At every iteration we form a quadratic approximation $m_{k}$ of the objective plus a quadratic penalty. The penalty serves to ensure that the next iterate does not move too far away from the current iterate, which is important when the Hessian is ill conditioned. The quadratic model plus penalty expanded about $\mathbf{b}^{k}$ for variables $\mathbf{d}_{F} \in \mathbb{R}^{|\mathcal{F}|}$ is

$$
\begin{equation*}
m_{k}\left(\mathbf{d}_{F} ; \mu_{k}\right)=f_{\text {row }}\left(\mathbf{b}^{k}\right)+\mathbf{d}_{F}^{T} \mathbf{g}_{F}^{k}+\frac{1}{2} \mathbf{d}_{F}^{T} \mathbf{H}_{F}^{k} \mathbf{d}_{F}+\frac{\mu_{k}}{2}\left\|\mathbf{d}_{F}\right\|_{2}^{2} \tag{3.4}
\end{equation*}
$$

The unique minimum of $m_{k}(\cdot)$ is

$$
\mathbf{d}_{F}^{k}=-\left(\mathbf{H}_{F}^{k}+\mu_{k} \mathbf{I}\right)^{-1} \mathbf{g}_{F}^{k},
$$

where $\mathbf{H}_{F}^{k}+\mu_{k} \mathbf{I}$ is known as the damped Hessian. Adding a multiple of the identity to $\mathbf{H}_{F}^{k}$ increases each of its eigenvalues by $\mu_{k}$, which has the effect of diminishing the length of $\mathbf{d}_{F}^{k}$, similar to the action of a trust region. The step $\mathbf{d}_{F}^{k}$ is computed using a Cholesky factorization of the damped Hessian, and the full space search direction $\mathbf{d}^{k} \in \mathbb{R}^{R}$ is then given by

$$
d_{r}^{k}= \begin{cases}\left(E \mathbf{d}_{F}^{k}\right)_{r} & r \in \mathcal{F}  \tag{3.5}\\ -\nabla_{r} f_{\text {row }}\left(\mathbf{b}^{k}\right) & r \in \mathcal{G} \\ 0 & r \in \mathcal{A}\end{cases}
$$

where index sets $\mathcal{A}, \mathcal{G}$, and $\mathcal{F}$ are defined in (3.3), and $E$ is an $R \times|\mathcal{F}|$ matrix that elongates the vector $\mathbf{d}_{F}^{k}$ to length $R$. Specifically, $e_{i j}=1$ if $i$ is the row subproblem variable corresponding to the $j$ th variable in $\mathcal{F}$, and zero otherwise.

The damping parameter $\mu_{k}$ is adjusted by a Levenberg-Marquardt strategy [30. First define the ratio of actual reduction over predicted reduction,

$$
\begin{equation*}
\rho=\frac{f_{\text {row }}\left(\mathbf{b}^{k}+\mathbf{d}^{k}\right)-f_{\text {row }}\left(\mathbf{b}^{k}\right)}{m_{k}\left(\mathbf{d}_{F}^{k} ; 0\right)-m_{k}(0 ; 0)} \tag{3.6}
\end{equation*}
$$

where $m_{k}(\cdot)$ is defined by (3.4). Note the numerator of (3.6) calculates $f_{\text {row }}$ using all variables, while the denominator calculates $m_{k}(\cdot)$ using only the variable in $\mathcal{F}$. The damping parameter is updated by the following rule

$$
\mu_{k+1}= \begin{cases}\frac{7}{2} \mu_{k} & \text { if } \rho<\frac{1}{4}  \tag{3.7}\\ \frac{2}{7} \mu_{k} & \text { if } \rho>\frac{3}{4} \\ \mu_{k} & \text { otherwise }\end{cases}
$$

Since $\mathbf{d}_{F}^{k}$ is the minimum of 3.4 , the denominator of 3.6 is always negative. If the search direction $\mathbf{d}^{k}$ increases the objective function, then the numerator of (3.6) will be positive; hence $\rho<0$ and the damping parameter will be increased for the next iteration. On the other hand, if the search direction $\mathbf{d}^{k}$ decreases the objective function, then the numerator will be negative; hence $\rho>0$ and the relative sizes of the actual reduction and predicted reduction will determine how the damping parameter is adjusted.
3.1.2. Line Search. After computing the search direction $\mathbf{d}^{k}$, we ensure the next iterate decreases the objective by using a projected backtracking line search that satisfies the Armijo condition [30]. Given scalars $0<\beta$ and $\sigma<1$, we find the smallest nonnegative integer $t$ that satisfies the inequality

$$
\begin{equation*}
f_{\text {row }}\left(P_{+}\left[\mathbf{b}^{k}+\beta^{t} \mathbf{d}^{k}\right]\right)-f_{\text {row }}\left(\mathbf{b}^{k}\right) \leq \sigma\left(P_{+}\left[\mathbf{b}^{k}+\beta^{t} \mathbf{d}^{k}\right]-\mathbf{b}^{k}\right)^{T} \nabla f_{\text {row }}\left(\mathbf{b}^{k}\right) \tag{3.8}
\end{equation*}
$$

We set $\alpha_{k}=\beta^{t}$ and the next iterate is given by

$$
\mathbf{b}^{k+1}=P_{+}\left[\mathbf{b}^{k}+\alpha_{k} \mathbf{d}^{k}\right]
$$

3.2. Projected Quasi-Newton Step. As an alternative to the damped Hessian step, we adapt the projected quasi-Newton step from [20]. Their work employs a limited-memory BFGS (L-BFGS) approximation [29] in a framework suitable for any convex, bound-constrained problem.

L-BFGS estimates Hessian properties based on the most recent $M$ update pairs $\left\{\mathbf{s}^{i}, \mathbf{y}^{i}\right\}, i \in[\max \{1, k-M\}, k]$, where

$$
\begin{equation*}
\mathbf{s}^{i}=\mathbf{b}^{i+1}-\mathbf{b}^{i}, \quad \mathbf{y}^{i}=\nabla f_{\text {row }}\left(\mathbf{b}^{i+1}\right)-\nabla f_{\text {row }}\left(\mathbf{b}^{i}\right) \tag{3.9}
\end{equation*}
$$

L-BFGS uses a two-loop recursion through the stored pairs to efficiently compute a vector $\mathbf{p}^{k}=\tilde{\mathbf{B}}^{k} \mathbf{g}^{k}$, where $\tilde{\mathbf{B}}^{k}$ approximates the inverse of the Hessian $\left[\mathbf{H}^{k}\right]^{-1}$ using the pairs $\left\{\mathbf{s}^{i}, \mathbf{y}^{i}\right\}$. Storage is set to $M=3$ pairs in all experiments. On the first iterate when $k=0$, we use a multiple of the identity matrix so that $\mathbf{p}^{0}$ is in the direction of the gradient. L-BFGS updates require the quantity $1 /\left(\left(\mathbf{s}^{i}\right)^{T} \mathbf{y}^{i}\right)$ to be positive. We check this condition and skip the update pair if it is violated. This can happen if all row variables are at their bound of zero, or from numerical roundoff at a point near a minimizer. See [30, Chapter 7] for further detail.

The projected quasi-Newton search direction $\mathbf{d}^{k}$, analogous to (3.5), is

$$
d_{r}^{k}= \begin{cases}-\left(E p^{k}\right)_{r} & r \in \mathcal{F}  \tag{3.10}\\ -\nabla_{r} f_{\text {row }}\left(\mathbf{b}^{k}\right) & r \in \mathcal{G} \\ 0 & r \in \mathcal{A}\end{cases}
$$

where $\mathcal{F}$ and $\mathcal{A}$ are determined from (3.3), and $E$ is the elongation matrix defined in 3.5.

The step $p^{k}$ is computed from an L-BFGS approximation over all variables in the row subproblem; in contrast, the step $\mathbf{d}_{F}^{k}$ computed from the damped Hessian in Section 3.1.1 is derived from the second derivatives of only the free variables in $\mathcal{F}$. We could build an L-BFGS model over just the free variables as is done in (9], but the computational cost is higher. Our L-BFGS step is therefore influenced by second-order information from variables not in $\mathcal{F}$. This information is irrelevant to the step, but we find that algorithm performance is still good. We now express the influence in terms of the reduced Hessian and inverse of the reduced Hessian.

Let $\mathbf{H}$ and $\mathbf{B}$ denote the true Hessian and inverse Hessian matrices over all variables in a row subproblem. Suppose the variables in $\mathcal{F}$ are the first $|\mathcal{F}|$ variables, and the remaining variables are in $\mathcal{N}=\mathcal{A} \cup \mathcal{G}$. Then we can write $\mathbf{H}$ in block form as

$$
\mathbf{H}=\left[\begin{array}{ll}
\mathbf{H}_{F F} & \mathbf{H}_{N F}^{T} \\
\mathbf{H}_{N F} & \mathbf{H}_{N N}
\end{array}\right]
$$

with $\mathbf{H}_{F F} \in \mathbb{R}^{|\mathcal{F}| \times|\mathcal{F}|}, \mathbf{H}_{N F} \in \mathbb{R}^{|\mathcal{N}| \times|\mathcal{F}|}$, and $\mathbf{H}_{N N} \in \mathbb{R}^{|\mathcal{N}| \times|\mathcal{N}|}$. The damped Hessian search direction in $\sqrt{3.5}$ is computed from the inverse of the reduced Hessian; that is, $\mathbf{B}_{F}=\mathbf{H}_{\tilde{F} F}^{-1}$.

Let $\tilde{\mathbf{H}}$ and $\tilde{\mathbf{B}}$ denote the L-BFGS approximation to the true and inverse Hessian. To obtain the step $p^{k}$ we use the inverse approximation $\tilde{\mathbf{B}}$, then extract just the free variables for use in 3.10 ; hence, we compute the search direction using the approximation $\tilde{\mathbf{B}}_{F}$. Assuming the Schur complement exists, this matrix is

$$
\tilde{\mathbf{B}}_{F}=\left(\tilde{\mathbf{H}}_{F F}-\tilde{\mathbf{H}}_{N F}^{T} \tilde{\mathbf{H}}_{N N}^{-1} \tilde{\mathbf{H}}_{N F}\right)^{-1}
$$

Comparing with the true reduced Hessian, we see the extra term $\tilde{\mathbf{H}}_{N F}^{T} \tilde{\mathbf{H}}_{N N}^{-1} \tilde{\mathbf{H}}_{N F}$, a matrix of $\operatorname{rank}|\mathcal{N}|$. This is the influence in the L-BFGS approximation of variables not in $\mathcal{F}$; we are effectively using the L-BFGS approximation of the reduced inverse Hessian to compute the step. Note that a small value of the tuning parameter $\epsilon$ in (3.2) can help reduce the size of $|\mathcal{N}|$, lessening the influence.
3.3. Stopping Criterion. Since the row subproblems are convex, any point satisfying the first-order KKT conditions is the optimal solution. Specifically, $\mathbf{b}^{*}$ is a KKT point of 2.6 if it satisfies

$$
\nabla f_{\text {row }}\left(\mathbf{b}^{*}\right)-\boldsymbol{v}^{*}=\mathbf{0}, \quad\left(\mathbf{b}^{*}\right)^{T} \boldsymbol{v}^{*}=0, \quad \mathbf{b}^{*} \geq \mathbf{0}, \quad \boldsymbol{v}^{*} \geq \mathbf{0}
$$

where $\boldsymbol{v}^{*}$ is the vector of dual variables associated with the nonnegativity constraints. Knowing the algorithm keeps all iterates $\mathbf{b}^{k}$ nonnegative, we can express the KKT condition for component $r$ as

$$
\left|\min \left\{b_{r}^{k}, \nabla_{r} f_{\text {row }}\left(\mathbf{b}^{k}\right)\right\}\right|=0
$$

A suitable stopping criterion is to approximately satisfy the KKT conditions to a tolerance $\tau>0$. We achieve this by requiring that all row subproblems satisfy

$$
\begin{equation*}
\mathrm{kkt}_{\mathrm{viol}}=\max _{r}\left\{\left|\min \left\{b_{r}^{k}, \nabla_{r} f_{\text {row }}\left(\mathbf{b}^{k}\right)\right\}\right|\right\} \leq \tau \tag{3.11}
\end{equation*}
$$

The full algorithm solves to an overall tolerance $\tau$ when the $\mathrm{kkt}_{\text {viol }}$ of every row subproblem satisfies (3.11). This condition is enforced for all the row subproblems (Step 4 of Algorithm 2) generated from all the tensor modes (Step 5 of Algorithm 1). Note that enforcement requires examination of $\mathrm{kkt}_{\text {viol }}$ for all row subproblems whenever the solution of any subproblem mode is updated, because the solution modifies the $\boldsymbol{\Pi}^{(n)}$ matrices of other modes.

```
Algorithm 3 Projected Newton-Based Solver for the Row Subproblem (PDN-R)
Given data \(\hat{\mathbf{x}}\) and \(\boldsymbol{\Pi}\), constants \(\mu_{0}, \sigma, \beta, K_{\max }\), stop tolerance \(\tau\), and initial values \(\mathbf{b}^{0}\)
Return a solution \(\mathbf{b}^{*}\) to Step 4 of Algorithm 2
    for \(k=0,1, \ldots, K_{\text {max }}\) do
        Compute the gradient, \(\mathbf{g}^{k}=\nabla f_{\text {row }}\left(\mathbf{b}^{k}\right)\), using \(\hat{\mathbf{x}}\) and \(\boldsymbol{\Pi}\) in 3.1)
        Compute the first-order KKT violation
            \(\mathrm{kkt}_{\text {viol }}=\max _{r}\left\{\left|\min \left\{b_{r}^{k}, g_{r}^{k}\right\}\right|\right\}\)
        if \(\mathrm{kkt}_{\text {viol }} \leq \tau\) then
        return \(\mathbf{b}^{*}=\mathbf{b}^{k} \quad \triangleright\) Converged to tolerance.
        end if
        Find the indices of free variables from (3.3) with \(\epsilon=10^{-3}\) in (3.2
        Calculate the Hessian for free variables
\[
\mathbf{H}_{F}^{k}=\left[\nabla^{2} f_{\mathrm{row}}\left(\mathbf{b}^{k}\right)\right]_{\mathcal{F}}
\]
Compute the damped Newton direction \(\mathbf{d}_{F}^{k}=-\left(\mathbf{H}_{F}^{k}+\mu_{k} I\right)^{-1} \mathbf{g}_{F}^{k}\)
Construct search direction \(\mathbf{d}^{k}\) over all variables using \(\mathbf{d}_{F}^{k}\) and \(\mathbf{g}^{k}\) in (3.5)
Perform the projected line search (3.8) using \(\sigma\) and \(\beta\) to find step length \(\alpha_{k}\) Update the current iterate
\[
\mathbf{b}^{k+1}=P_{+}\left[\mathbf{b}^{k}+\alpha_{k} \mathbf{d}^{k}\right]
\]
Update the damping parameter \(\mu_{k+1}\) according to (3.6)-(3.7)
end for
return \(\mathbf{b}^{*}=\mathbf{b}^{k} \quad \triangleright\) Iteration limit reached.
```

3.4. Row Subproblem Algorithms. Having described the ingredients, we pull everything together into complete algorithms for solving the row subproblem in Step 4 of Algorithm 2. We present two methods in Algorithm 3 and Algorithm 4 PDN-R uses a damped Hessian matrix, and PQN-R uses a quasi-Newton Hessian approximation (the '-R' designates a row subproblem formulation). Both algorithms employ a two-metric projection framework for handling bound constraints and a line search satisfying the Armijo condition.

As mentioned, PQN-R is related to [20. Specifically, we note

1. The free variables chosen in Step 7 of $\mathrm{PQN}-\mathrm{R}$ are found with $\epsilon=10^{-8}$ in (3.2), while [20 effectively uses $\epsilon=0$ for an instantaneous line search. As noted in Section 3.1 convergence is guaranteed when $\epsilon>0$. Step 7 of PDN-R uses the default value $\epsilon=10^{-3}$ because we find it generally leads to faster convergence.
2. The line search in Step 9 of PQN-R and Step 11 of PDN-R satisfies the Armijo condition. This differs from [20, which used $\sigma \alpha\left(\mathbf{d}^{k}\right)^{T} \nabla f_{\text {row }}\left(\mathbf{b}^{k}\right)$ on the righthand side of (3.8). We use (3.8) because it correctly measures predicted progress. In particular, it is easier to satisfy when $\left(\mathbf{d}^{k}\right)^{T} \nabla f_{\text {row }}\left(\mathbf{b}^{k}\right)$ is large and many variables hit their bound for small $\alpha$.
3. Updates to the L-BFGS approximation in Step 11 of PQN-R are unchanged from [20]. Information is included from all row subproblem variables, whether active or free.
```
Algorithm 4 Projected Quasi-Newton Solver for the Row Subproblem (PQN-R)
Given data \(\hat{\mathbf{x}}\) and \(\boldsymbol{\Pi}\), constants \(\mu_{0}, \sigma, \beta, K_{\max }\), stop tolerance \(\tau\), and initial values \(\mathbf{b}^{0}\)
Return a solution \(\mathbf{b}^{*}\) to Step 4 of Algorithm 2
    for \(k=0,1, \ldots, K_{\max }\) do
        Compute the gradient, \(\mathbf{g}^{k}=\nabla f_{\text {row }}\left(\mathbf{b}^{k}\right)\), using \(\hat{\mathbf{x}}\) and \(\boldsymbol{\Pi}\) in 3.1)
        Compute the first-order KKT violation
            \(\mathrm{kkt}_{\mathrm{viol}}=\max _{r}\left\{\left|\min \left\{b_{r}^{k}, g_{r}^{k}\right\}\right|\right\}\)
        if \(\mathrm{kkt}_{\text {viol }} \leq \tau\) then
            return \(\mathbf{b}^{*}=\mathbf{b}^{k} \quad \triangleright\) Converged to tolerance.
        end if
        Find the indices of free variables from (3.3) with \(\epsilon=10^{-8}\) in 3.2
        Construct search direction \(\mathbf{d}^{k}\) using \(\mathbf{g}^{k}\) in (3.10)
        Perform the projected line search (3.8 using \(\sigma\) and \(\beta\) to find step length \(\alpha_{k}\)
        Update the current iterate
            \(\mathbf{b}^{k+1}=P_{+}\left[\mathbf{b}^{k}+\alpha_{k} \mathbf{d}^{k}\right]\)
        Update the L-BFGS approximation with \(\mathbf{b}^{k+1}\) and \(\mathbf{g}^{k+1}\)
    end for
    return \(\mathbf{b}^{*}=\mathbf{b}^{k} \quad \triangleright\) Iteration limit reached.
```

We express the computational cost of PDN-R and PQN-R in terms of the cost per iteration of Algorithm 1 that is, the cost of executing Steps 3 through 8 . The matrix $\Pi^{(n)}$ is formed for the row subproblems of every mode, with the cost for each mode proportional to the number of nonzeros in the data tensor, nnz $(\boldsymbol{X})$. This should dominate the cost of reweighting factor matrices in Steps 6 and 7. The $n$th mode solves $I_{n}$ convex row subproblems, each with $R$ unknowns, using Algorithm 3 (PDN-R) or Algorithm 4 (PQN-R). Row subproblems execute over at most $K_{\max }$ inner iterations. Near a local minimum we expect PDN-R to take fewer inner iterations than PQN-R because the damped Newton method converges asymptotically at a quadratic rate, while L-BFGS convergence is at best R -linear. However, the cost estimate will assume the worst case of $K_{\max }$ iterations for all row subproblems. The dominant cost of Algorithm 3 is solution of the damped Newton direction in Step 9 which costs $O\left(R^{3}\right)$ operations to solve the $R \times R$ dense linear system. Hence, the cost per iteration of PDN-R is

$$
\begin{equation*}
N \cdot O(\mathrm{nnz}(\mathcal{X}))+K_{\max } \cdot O\left(R^{3}\right) \cdot \sum_{n=1}^{N} I_{n} \tag{3.12}
\end{equation*}
$$

The dominant costs of Algorithm 4 are computation of the search direction and updating the L-BFGS matrix, both $O(R)$ operations. Hence, the cost per iteration of PQN-R is

$$
\begin{equation*}
N \cdot O(\mathrm{nnz}(\mathcal{X}))+K_{\max } \cdot O(R) \cdot \sum_{n=1}^{N} I_{n} \tag{3.13}
\end{equation*}
$$

4. Experiments. This section characterizes the performance of our algorithms, comparing them with multiplicative update [11] and second-order methods that do not use the row subproblem formulation. All algorithms fit in the alternating block framework of Algorithm 1, differing in how they solve (2.4) in Step 5.

Our two algorithms are the projected damped Hessian method (PDN-R) and the projected quasi-Newton method (PQN-R), from Algorithms 3 and 4 respectively. Recall that '-R' means the row subproblem formulation is applied. In this paper we do not tune the algorithms to each test case, but instead chose a single set of parameter values: $\mu_{0}=10^{-5}, \sigma=10^{-4}$, and $\beta=1 / 2$. The bound constraint threshold in PDN-R from 3.2 was set to $\epsilon=10^{-3}$ for PDN-R and $\epsilon=10^{-8}$ for PQN-R, values that are observed to give best algorithm performance. The L-BFGS approximations in PQN-R stored the $M=3$ most recent update pairs 3.9 .

The multiplicative update (MU) algorithm that we compare with is that of Chi and Kolda [11], available as function cp_apr in the Matlab Tensor Toolbox [4]. It builds on tensor generalizations of the Lee and Seung method, specifically treating inadmissible zeros (their term for factor elements that are active but close to zero) to improve the convergence rate. Algorithm MU can be tuned by selecting the number of inner iterations for approximately solving the subproblem at Step 5 of Algorithm 1 . We found that ten inner iterations worked well in all experiments.

We also compare to a projected quasi-Newton (PQN) algorithm adopted from Kim et al. 20]. PQN is similar to PQN-R but solves (2.4) without reformulating the block subproblem into row subproblems. PQN identifies the active set using $\epsilon=0$ in $(3.2)$ and maintains a limited-memory BFGS approximation of the Hessian. However, $\mathrm{PQ} N$ uses one L-BFGS matrix for the entire subproblem, storing the three most recent update pairs. We used Matlab code from the authors of 20, embedding it in the alternating framework of Algorithm 1, with the modifications described in Section 3.2.

Additionally, we compare PDN-R to a projected damped Hessian (PDN) method that uses one matrix for the block subproblem instead of a matrix for every row subproblem. PDN exploits the block diagonal nature of the Hessian to construct a search direction for the same computational cost as PDN-R; i.e., one search direction of PDN takes the same effort as computing one search direction for all row subproblems in PDN-R. Similar remarks apply to computation of the objective function for the subproblem (2.4). However, PDN applies a single damping parameter $\mu_{k}$ to the block subproblem Hessian and updates all variables in the block subproblem from a single line search along the search direction.

All algorithms were coded in Matlab using the sparse tensor objects of the Tensor Toolbox [4. All experiments were performed on a Linux workstation with 12GB memory. Data sets were large enough to be demanding but small enough to fit in machine memory; hence, performance results are not biased by disk access issues.

The experiments that follow show three important results, as follows.

1. The row subproblem formulation is better suited to second-order methods than the block subproblem formulation because it controls the number of iterations for each row subproblem independently, and because its convergence is more robust.
2. PDN-R and PQN-R are faster than the other algorithms in terms of in reducing the $\mathrm{kkt}_{\text {viol }}$, especially when solving to high accuracy. This holds for any number of components. PQN-R becomes faster than PDN-R as the number of components increases.
3. PDN-R and PQN-R reach good solutions with high sparsity more quickly than the other algorithms, a desirable feature when the factor matrices are expected to be sparse.
In Section 4.1 we report only performance in solving a single block subproblem (2.4) since the time is representative of the total time it will take to solve the full tensor factorization problem (2.1). In Section 4.2 we report results from solving the full problem within the alternating block framework (Algorithm 1).
4.1. Solving the Convex Block Subproblem. We begin by examining algorithm performance on the convex subproblem 2.4 of the alternating block framework. Here we look at a single representative subproblem. Our goal is to characterize the relative behavior of algorithms on the representative block subproblem.

Appendix A describes our method for generating synthetic test problems with reasonable sparsity. We investigate a three-way tensor of size $200 \times 300 \times 400$, generating $S=500,000$ data samples. The number of components, $R$, is varied over the set $\{20,40,60,80,100\}$. For each value of $R$, the procedure generates a sparse multilinear model $\mathcal{M}=\llbracket \boldsymbol{\lambda} ; \mathbf{A}^{(1)}, \mathbf{A}^{(2)}, \mathbf{A}^{(3)} \rrbracket$ and data tensor $\boldsymbol{X}$. Table 4.1 lists the number of nonzero elements found in the data tensor $\boldsymbol{X}$ that results from 500,000 data samples, averaged over ten random seeds. The number of nonzeros, a key determiner of algorithm cost in equations $(3.12)$ and $(3.13)$, is approximately the same for all values of $R$.

Table 4.1: Subproblem sparsity for number of components $R$

| $R$ | Number Nonzeros | Density |
| ---: | :---: | :---: |
| 20 | 413,460 | $1.72 \%$ |
| 40 | 450,760 | $1.88 \%$ |
| 60 | 464,440 | $1.94 \%$ |
| 80 | 470,950 | $1.96 \%$ |
| 100 | 475,450 | $1.98 \%$ |

We consider just the subproblem obtained by unfolding along mode 1 ; hence, the test case contains 200 row subproblems of the form 2.6 . To solve just the mode- 1 subproblem, the for loop at Step 3 of Algorithm 1 is changed to $n=1$.

We run several trials of the subproblem solver from different initial guesses of the unknowns, holding $\mathbf{A}^{(2)}$ and $\mathbf{A}^{(3)}$ from $\mathcal{M}$ constant. The initial guess draws each element of $\mathbf{A}^{(1)}$ from a uniform distribution on $[0,1)$ and sets each element of $\boldsymbol{\lambda}$ to one. To satisfy constraints in 2.1, the columns of $\mathbf{A}^{(1)}$ are normalized and the normalization factor is absorbed into $\boldsymbol{\lambda}$. The mode-1 subproblem (2.4) is now defined with $\boldsymbol{\Pi}=\left(\mathbf{A}^{(3)} \odot \mathbf{A}^{(2)}\right)^{T}, \mathbf{X}=\mathbf{X}_{(1)}$, and $\mathbf{B}=\mathbf{A}^{(1)} \boldsymbol{\Lambda}$, with unknowns $\mathbf{B}$ initialized using the initial guess for $\boldsymbol{\lambda}$ and $\mathbf{A}^{(1)}$.
4.1.1. PDN-R and PDN on the Convex Subproblem. We first characterize the behavior of our Newton-based algorithm, PDN-R, and compare it with PDN. Row subproblems are solved using Algorithm 3 with stop tolerance $\tau=10^{-8}$ and the parameter values mentioned at the beginning of Section 4. The value of $K_{\max }$ in Algorithm 3 is large enough that the $\mathrm{kkt}_{\text {viol }}$ converges to $\tau$ before $K_{\max }$ is reached.

Figures 4.1a-4.1c show how KKT violations decrease with iteration for three different values of $R$. The subproblem was solved ten times from different randomly chosen start points. (Since the subproblem is strictly convex, there is a single unique


Fig. 4.1: Convergence behavior of PDN-R (Algorithm 3) and PDN over ten runs with different start points, for three values of $R$. The upper graphs plot $\log \left(\mathrm{kkt}_{\mathrm{viol}}\right)$, showing how the maximum violation over all row subproblems varies as the number of iterations increases. Solid lines are PDN-R, dashed lines are PDN. The lower graphs plot the number of rows violating the KKT-based stop tolerance (ten runs of only PDN-R).
minimum that is reached from every start point.) Each solid line plots the maximum $\mathrm{kkt}_{\text {viol }}$ over all 200 row subproblems for one of the ten PDN-R runs. Each dashed line plots the $\mathrm{kkt}_{\text {viol }}$ of the block subproblem for one of the ten PDN runs. Note the $y$-axis is the $\log _{10}$ of $\mathrm{kkt}_{\text {viol }}$. The figure demonstrates that after some initial slow progress, both algorithms exhibit the fast quadratic convergence rate typical of Newton methods. PDN-R clearly takes fewer iterations to compute a factorization with small $\mathrm{kkt}_{\text {viol }}$.

Figures 4.1d - 4.1 f show the number of row subproblems in PDN-R that satisfy the KKT-based stop tolerance after a given number of iterations. Remember that all row subproblems must satisfy the KKT tolerance before the algorithm declares a solution.

Figure 4.2 shows additional features of the convergence, for just the case of $R=$ 100 components (behavior is similar for other values of $R$ ). In Figure 4.2a we see the number of elements of $\mathbf{A}^{(1)}$ exactly equal to zero. Data for this experiment was generated stochastically from sparse factor matrices (see Appendix A); hence, we expect a sparse solution. The plot indicates that sparsity can be achieved after


Fig. 4.2: Additional convergence behavior over ten runs with different start points. Solid lines are PDN-R, dashed lines are PDN. The red horizontal line shows the true number of elements exactly equal to zero.
reducing $\mathrm{kkt}_{\text {viol }}$ to a moderately small tolerance (around $10^{-2}$ in this example). We return to sparsity of the solution in the sections below.

In Figure 4.2 b we see that execution time per iteration decreases when variables are closer to a solution. PDN-R execution time becomes very small because only a few row subproblems need to satisfy the convergence tolerance, and only these are updated. PDN takes more time per iteration because it computes a single search direction for updating all variables in the block subproblem, even when most of the variables are near an optimal value. These experiments show that PDN-R and PDN behave similarly for the convex subproblem and that PDN-R is a little faster; much larger differences appear when the full factorization is computed in Section 4.2.1.
4.1.2. PQN-R and PQN on the Convex Subproblem. In this section we demonstrate the importance of the row subproblem formulation by comparing PQN-R with PQN, showing the huge speedup achieved with our row subproblem formulation. We compare the algorithms on the mode-1 subproblem described above, from the same ten random initial guesses for $\mathbf{A}^{(1)}$. Table 4.2 lists the average CPU times over ten runs. PQN-R was executed until the $\mathrm{kkt}_{\text {viol }}$ was less than $\tau=10^{-8}$. PQN was unable to achieve this level of accuracy, so execution was stopped at a tolerance of $10^{-3}$. Results in the table show that PQN-R is much faster at decreasing the KKT violation. We note that a KKT violation of $10^{-8}$ is approximately the square root of machine epsilon, the smallest practical value that can be attained.

The two algorithms also differ in how they discover the number of elements in $\mathbf{A}^{(1)}$ equal to zero. Both eventually agree on the number of zero elements, but PQNR is much faster. Figure 4.3 shows the progress made by the two algorithms; the behavior of PQN for this quantity is erratic and slow to converge.

Algorithm PQN might be relatively more competitive for tensor subproblems with a small number of rows. Nevertheless, it is apparent that applying L-BFGS to the block subproblem does not work as well as applying separate instances of L-BFGS to the row subproblems. This is not surprising since the first method ignores the block diagonal structure of the true Hessian. We see no advantages to using PQN and do not consider it further.

Table 4.2: Convergence comparison in solving the subproblem

|  | Algorithm PQN |  | Algorithm PQN-R |  |
| ---: | ---: | ---: | ---: | ---: |
| $R$ | $\mathrm{kkt}_{\text {viol }}<10^{-1}$ | $\mathrm{kkt}_{\text {viol }}<10^{-3}$ | $\mathrm{kkt}_{\text {viol }}<10^{-1}$ | $\mathrm{kkt}_{\text {viol }}<10^{-8}$ |
| 20 | 625 secs | 690 secs | 12.4 secs | 17.1 secs |
| 40 | 755 secs | 846 secs | 10.9 secs | 16.4 secs |
| 60 | 822 secs | 920 secs | 11.3 secs | 16.8 secs |
| 80 | 1022 secs | 1141 secs | 13.7 secs | 19.5 secs |
| 100 | 993 secs | 1125 secs | 13.1 secs | 20.2 secs |



Fig. 4.3: Number of elements equal to zero in $\mathbf{A}^{(1)}$ found by PQN-R (solid lines, forming a short segment in the upper left) and PQN (dashed lines) as a function of compute time, for a subproblem with different values of $R$. The same subproblem was solved from six different start points, corresponding to the different colors. The red horizontal line shows the true value.
4.1.3. PDN-R, PQN-R, and MU on the Convex Subproblem. Next we compare our new row-based algorithms, PDN-R and PQN-R, with the multiplicative update method MU [11]. Again we use the mode-1 subproblem of Section 4.1, from the same ten random initial guesses.

As described in Section 4, MU is a state-of-the-art representative of the most common algorithm for nonnegative tensor factorization. It is a form of scaled steepestdescent with bound constraints [26], and therefore is expected to converge more slowly than Newton or quasi-Newton methods. We see this clearly in Table 4.3 for two different stop tolerances. The MU algorithm was executed with a time limit of 1800 seconds per problem, and failed to reach $\mathrm{kkt}_{\mathrm{viol}}<10^{-3}$ before this limit when $R$ was 60 or larger.

Of course, the disparity in convergence time is more pronounced when a smaller KKT error is demanded. Figure 4.4 shows the decrease in KKT violation as a function of compute time. Here we see that MU makes a faster initial reduction in KKT violation than PDN-R or PQN-R, but then it slows to a linear rate of convergence. Notice the gap from time zero for PDN-R and PQN-R, which reflects setup cost before the first iteration result is computed. For PQN-R the setup time is fairly constant with $R$ (about 3.8 seconds), while PDN-R has a setup time that increases with $R$ (11.5 seconds for $R=100$ ). Unlike MU, both algorithms must construct software

Table 4.3: Time to reach stop tolerance for three algorithms (averaged over ten runs)

|  | $\mathrm{kkt}_{\text {viol }}=10^{-2}$ |  |  | $\mathrm{kkt}_{\text {viol }}=10^{-3}$ |  |  |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| $R$ | PDN-R | PQN-R | MU | PDN-R | PQN-R | MU |
| 20 | 8.1 secs | 14.5 secs | 97.7 secs | 8.1 secs | 15.6 secs | 161.3 secs |
| 40 | 25.1 secs | 13.1 secs | 239.2 secs | 25.2 secs | 14.6 secs | 485.9 secs |
| 60 | 53.6 secs | 13.8 secs | 469.2 secs | 53.7 secs | 15.6 secs | $>1800$ secs |
| 80 | 92.8 secs | 16.3 secs | 455.4 secs | 92.9 secs | 18.1 secs | $>1800$ secs |
| 100 | 139.8 secs | 16.0 secs | 730.7 secs | 140.0 secs | 18.3 secs | $>1800$ secs |

structures for all row subproblems before a first iteration result appears. Figure 4.4 also reveals that PDN-R is slower relative to PQN-R as the number of components $R$ increases. This is because the cost of solving a Newton-based Hessian is $O\left(R^{3}\right)$, while the limited-memory BFGS Hessian cost is $O(R)$.


Fig. 4.4: Convergence behavior comparison on subproblem for different values of $R$ (ten runs each). Algorithm MU (blue) makes fast initial progress in reducing the violation, but slows dramatically after reaching a violation of about 0.1. PDN-R (black) and PQN-R (green) reduce the violation much further, with PQN-R being faster for higher values of $R$.

Figure 4.4 indicates that algorithm MU is preferred if a relatively large $\mathrm{kkt}_{\text {viol }}$ is acceptable. We contend that this is not a good choice if the goal is to find a sparse solution. Figure 4.5 plots the number of elements that equal zero as a function of CPU time. It shows that PDN-R and PQN-R both converge to the correct number of zeros much faster than MU.

On closer inspection we see that MU is actually making factor elements small, and is just very slow at making them exactly zero. If we choose a small positive threshold instead of zero, then MU might arguably do well at finding a sparse solution. Figure 4.6 summarizes an investigation of this idea. Three different thresholds are shown: $10^{-3}, 10^{-4}$, and $10^{-5}$. The first threshold is clearly too large, declaring elements to be "zero" when they never converge to such a value. A threshold of $10^{-4}$ is also too large for $R=20$, though possibly acceptable for $R=40$ and $R=60$. The choice of $10^{-5}$ correctly identifies elements converging to zero, but PDN-R and PQN$R$ identifies them much faster. We conclude that PDN-R and PQN-R are significantly


Fig. 4.5: Effectiveness of the three algorithms in finding a sparse solution for different values of $R$. In each case the number of elements in $\mathbf{A}^{(1)}$ equal to zero is plotted against execution time. The PDN-R (black) and PQN-R (green) algorithms are much faster than MU (blue) at finding the zeros. The horizontal red line shows the true value.


Fig. 4.6: Comparison of PDN-R, PQN-R and MU in finding elements of $\mathbf{A}^{(1)}$ equal to zero for a sample run. MU rarely finds exact zeros; therefore, we show results of applying various thresholds. Some experimentation may be needed to find the best threshold; regardless, it is slower than the methods proposed here.
better at finding a true sparse solution than $M \mathrm{M}$, in terms of robustness (no need to choose an ad-hoc threshold) and computation time (assuming a suitable threshold for MU is known).
4.2. Solving the Full Problem. In this section we move from a convex subproblem to solving the full factorization 2.1 . We generate the same $200 \times 300 \times 400$ tensor data as in Section 4.1 and now treat all modes as optimization variables. An initial guess is constructed for all three modes in the same manner that $\mathbf{A}^{(1)}$ was initialized in Section 4.1. We generate ten different tensors by changing the random seed used in Algorithm 5 and solve each from a single initial guess. All tensors are factorized from the same initial guess; however, since the full factorization is a nonconvex
optimization problem, algorithms may converge to different local solutions.
We expect our local solutions to be reasonably close to the multilinear model $\mathcal{M}=\llbracket \boldsymbol{\lambda} ; \mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(N)} \rrbracket$ that generated the synthetic tensor data. We compared computed solutions with the original model using the Tensor Toolbox function score with option greedy. This function implements the congruence test described in 37 and [1]. The comparison considers angular differences between corresponding vectors of the factor matrices, producing a number between zero (poor match) and one (exact match). Solutions computed with any algorithm to a tolerance of $\tau=10^{-4}$ scored above 0.84 (see the supplementary material for a detailed breakdown). Perfect scores cannot be seen because the tensor data is generated from the model as a noisy Poisson process. Scores of less than 0.01 resulted when comparing the solution to other models generated with a different random seed. These results show that an accurate factorization can yield good approximations to the original factors for our test problems; however, our focus is on behavior of the algorithms in computing a solution.
4.2.1. Comparing PDN-R with PDN. We first compare the two Newtonbased methods: PDN-R, which solves row subproblems for each tensor mode, and PDN, which instead solves the block subproblem as a single matrix. In Section 4.1.1 we saw that the two methods behaved similarly for the convex subproblem of a single tensor mode (except that PDN-R was faster). However, on the full factorization PDN is often unable to make progress from a start point where the KKT violation is large. Sometimes the search direction does not satisfy the sufficient decrease condition of the Armijo line search, even after ten backtracking iterations. More frequently, the line search puts too many variables at the bound of zero, causing the objective function to become undefined in equation 2.4 because $\mathbf{B}^{(n)} \boldsymbol{\Pi}^{(n)}$ is zero for elements where $\mathbf{X}_{(n)}$ is nonzero.

If the line search fails in a subproblem, then we compute a multiplicative update step for that iteration to make progress. This allows PDN to reach points where the KKT error is smaller, and we find that subsequent damped Newton steps are successful until convergence. Table 4.4 quantifies the number of line search failures over the first 20 iterations, beginning from a random start point where $\mathrm{kkt}_{\text {viol }}$ is typically larger than $10^{3}$. Columns in the table correspond to different values for the initial damping parameter $\mu_{0}$. We expect larger values of $\mu_{k}$ to improve robustness by effectively shortening the step length and hopefully avoiding the mistake of setting too many variables to zero. However, a serious drawback to increasing $\mu_{k}$ is that it damps out Hessian information, which can hinder the convergence rate. The table shows that improvement in robustness is made; however, PDN still suffers from some line search failures. In contrast, PDN-R does not have any line search failures for the same test problems and start points, using the default $\mu_{0}=10^{-5}$.

Table 4.4: Line search failures by PDN in the first 20 iterations, averaged over five runs. There were up to 900 possible line searches in each case (a maximum of 15 inner iterations per mode, over 20 outer iterations).

| $R$ | $\mu_{0}=10^{1}$ | $\mu_{0}=10^{-2}$ | $\mu_{0}=10^{-5}$ |
| ---: | :---: | :---: | :---: |
| 20 | 57.8 | 88.4 | 142.2 |
| 40 | 76.2 | 87.4 | 164.6 |
| 60 | 59.0 | 90.8 | 201.0 |
| 80 | 41.4 | 82.8 | 184.6 |
| 100 | 28.8 | 62.2 | 168.0 |

22

Table 4.5 shows that PDN-R is significantly faster than PDN even in the region where PDN operates robustly. These runs begin at a start point where $\mathrm{kkt}_{\text {viol }}<0.1$ and use $\mu_{0}=10^{-2}$ (PDN-R uses its default of $\mu_{0}=10^{-5}$ ) so that PDN does not suffer any line search failures. Five runs are made for each of the five values of $R$, and the method stops when the algorithm reduces $\mathrm{kkt}_{\text {viol }}$ below a given threshold (rows of Table 4.5). PDN does not always reach a threshold value in the three-hour-computation-time limit, but PDN-R always succeeds. The third column shows that the number of outer iterations needed to reach a threshold is very similar between PDN-R and PDN. The fourth column shows that PDN-R executes much faster.

Table 4.5: Comparison of PDN-R and PDN execution times for various stop tolerances. 25 experiments were run, the algorithms compared for each experiment that PDN completed, and the average value reported. The third column is computed as $\left|\operatorname{its}_{P D N}-\operatorname{its}_{P D N-R}\right| / \max \left\{\operatorname{its}_{P D N}\right.$, itsPDN-R $\}$. The fourth column shows that PDN-R executes from 8 to 9 times faster than PDN (the column reports average and standard deviation). The last column shows average execution time of PDN-R.

| kkt $_{\text {viol }}$ | PDN failures | avg diff in outer its | PDN-R speedup | PDN-R time |
| :---: | :---: | :---: | :---: | :---: |
| $10^{-2}$ | 2 | $2.48 \%$ | $9.1 \pm 1.4$ | 463.3 secs |
| $10^{-3}$ | 5 | $3.02 \%$ | $8.7 \pm 1.4$ | 595.0 secs |
| $10^{-4}$ | 9 | $2.68 \%$ | $8.5 \pm 1.7$ | 609.7 secs |
| $10^{-5}$ | 13 | $7.93 \%$ | $9.5 \pm 2.4$ | 573.1 secs |

Iterations of PDN-R run faster because each row subproblem has an individualized step size and damping parameter (this was discussed previously in Section 4.1.1). Given the large disparity in execution time and the lack of robustness when far from a solution, we find no advantages to using PDN and do not consider it further.
4.2.2. Comparing PDN-R, PQN-R, and MU. Table 4.6 summarizes the time to reach a KKT threshold of $10^{-3}$ for each algorithm over the synthetic data tensors. Like the convex subproblem tested in Section 4.1.3, the PDN-R and PQN-R methods converge to this relatively high accuracy much faster than MU, again showing the value of second-order information. As in the subproblem, we see that PQN-R is faster relative to PDN-R as the number of components, $R$, increases. Figure 4.7 shows convergence behavior of the full factorization problem in the same way that Figure 4.4 showed behavior of the convex subproblem. The KKT error of the full problem does not reach the quadratic rate of decrease seen in the subproblem. This is due to nonconvexity of the full factorization problem, and the alternation between solutions of each mode.

As with the subproblem, we also observe better convergence by our methods to a sparse solution. Figure 4.8 shows PDN-R and PQN-R reaching the final count of zero elements much faster than MU. As in Section 4.1.3, we argue that PDN-R and PQN-R are superior when the task is to find a solution with correct sparsity.

We performed similar experiments on tensors of the same size but different sparsity. Results are in the supplementary material. They lead to the same conclusions as the data in Table 4.6 namely, that PDN-R and PQN-R are faster than MU, and PQN-R becomes faster than PDN-R as the number of components increases.

The supplementary material also describes a simple experiment with sparse tensors whose factor matrices have a high degree of collinearity between column vectors. Such problems sometimes lead to poor algorithm performance (e.g., the "swamps" in

Table 4.6: Time to reach stop tolerance $10^{-3}$ on full problem (over ten runs). Mean and standard deviation are reported. Some runs of MU failed to reach the tolerance in three hours of execution. Results for different stop tolerances are in the supplementary material.

| $R$ | PDN-R | PQN-R | MU |  |
| ---: | ---: | ---: | ---: | ---: |
| 20 | $229 \pm 57$ secs | $397 \pm 123$ secs | $3355 \pm 1933$ secs | (0 failures) |
| 40 | $493 \pm 151$ secs | $818 \pm 185$ secs | $8101 \pm 2045$ secs | (2 failures) |
| 60 | $1003 \pm 349$ secs | $966 \pm 286$ secs | $9628 \pm 978$ secs | (5 failures) |
| 80 | $1682 \pm 642$ secs | $1639 \pm 390$ secs | no successes | (10 failures) |
| 100 | $2707 \pm 773$ secs | $1995 \pm 743$ secs | no successes | (10 failures) |



Fig. 4.7: Convergence behavior of the PDN-R (black lines) and PQN-R (green lines) algorithms in computing a full three-way solution. Each algorithm was run on ten different tensors.
[31). Performance of PDN-R and PQN-R was much better than algorithm MU in this experiment as well.
4.2.3. Comparing with DBLP Data. We also compare the same three algorithms on the sparse three-way tensor of DBLP data [14] examined in 15]. The data counts the number of papers published by author $i_{1}$ at conference $i_{2}$ in year $i_{3}$, with dimensions $7108 \times 1103 \times 10$. The tensor contains 112,730 nonzero elements, a density of $0.14 \%$. The data was factorized for $R$ between 10 and 100 in [15] (using a least squares objective function), so we use $R \in\{20,60,100\}$ in our experiments. Behavior of the algorithms on the DBLP data was similar to behavior on our synthetic data. Figure 4.9 shows how the count of elements equal to zero changes as algorithms progress, for ten runs that start from different random initial guesses. Again we see that PDN-R and PQN-R reach a sparse solution faster than MU.

Factorizations of the DBLP data computed with PDN-R and PQN-R were quite sparse, making them easier to interpret. The fraction of elements exactly equal to zero in the computed conference factor matrix was $98.1 \%$. The author factor matrix was also very sparse, with $95.4 \%$ of the elements exactly zero. These results were for a factorization with $R=100$, stopped after 800 seconds of execution with the KKT violation reduced to around $5 \times 10^{-4}$. Figure 4.10 shows a component that


Fig. 4.8: Effectiveness of the algorithms in finding a sparse solution for a full three-way solution. In each case the total number of elements in $\mathbf{A}^{(1)}, \mathbf{A}^{(2)}$, and $\mathbf{A}^{(3)}$ equal to zero is plotted against execution time. The PDN-R (black lines) and PQN-R (green) algorithms are much faster than MU (blue). Each algorithm was run on ten different tensors, so the final number of zero elements has ten different values.


Fig. 4.9: Effectiveness of the algorithms in finding a sparse solution for the DBLP tensor. In each case the total number of elements in $\mathbf{A}^{(1)}, \mathbf{A}^{(2)}$, and $\mathbf{A}^{(3)}$ equal to zero is plotted against execution time. The PDN-R (black lines) and PQN-R (green) algorithms are much faster than MU (blue).
detects related conferences that took place only in even years. The two dominant conferences are the same as those reported in Figure 7 of [15]. Figure 4.11 shows another component that groups conferences that took place only in odd years. In both components, the sparsity is striking, especially for the conference factor.
5. Summary. In this paper we consider the problem of nonnegative tensor factorization using a K-L objective function, and we derive a row subproblem formulation that allows efficient use of second order information. We present two new algorithms that exploit the row subproblem reformulation: PDN-R uses second derivatives in the optimization, while PQN-R uses a quasi-Newton approximation. We show that using the same second order information in a block subproblem formulation is less robust


Fig. 4.10: Computed factors from DBLP data for component 26 (i.e., the 26th largest component by weight). The two dominant conferences, ECAI and KR, occurred only in even years, except for KR in 1991. Factors are extremely sparse: $91 \%$ (6456) of elements in the author factor are exactly zero, as are $98 \%$ (1084) of the conference elements.


Fig. 4.11: Computed factors from DBLP data for component 41. The three dominant conferences (ICDAR, ICIAP, and CAIP) occurred only in odd years. In this component, $93 \%$ (6626) of elements in the author factor are exactly zero, as are $98 \%$ (1083) conference elements.
and more expensive computationally than a row subproblem formulation. We show that both PDN-R and PQN-R are much faster than the best multiplicative update method,especially when high accuracy solutions are desired. We further show that high accuracy is needed to identify zeros and compute sparse factors without resorting to the use of ad-hoc thresholds. This is important because sparse count data is likely to have sparsity in the factors, and sparse factors are always easier to interpret.

Our Matlab algorithms will appear in Version 2.6 of the Tensor Toolbox [4]. We mention in section 2.2 that row subproblems can be solved in parallel, and we anticipate developing other versions of the algorithms for shared and distributed memory machines.

Acknowledgments. We thank the authors of [20] for sharing Matlab code that we used in the experiments. We thank the anonymous referees for clarifying some of our arguments, and for suggesting additional experiments.

This work was partially funded by the Laboratory Directed Research \& Development (LDRD) program at Sandia National Laboratories. Sandia National Laboratories is a multiprogram laboratory operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC0494AL85000.

## REFERENCES

[1] E. Acar, D. M. Dunlavy, and T. G. Kolda, A scalable optimization approach for fitting canonical tensor decompositions, Journal of Chemometrics, 25 (2011), pp. 67-86.
[2] G. Andrew and J. Gao, Scalable training of $l_{1}$-regularized log-linear models, in Proceedings of the 24th International Conference on Machine Learning, ACM, 2007, pp. 33-40.
[3] B. W. Bader, M. W. Berry, and M. Browne, Discussion tracking in Enron email using PARAFAC, in Survey of Text Mining: Clustering, Classification, and Retrieval, M. W. Berry and M. Castellanos, eds., Springer, 2nd ed., 2007, ch. 8, pp. 147-162.
[4] B. W. Bader, T. G. Kolda, et al., MATLAB Tensor Toolbox version 2.5. Available online, January 2012. http://www.sandia.gov/~tgkolda/TensorToolbox/.
[5] D. Bertsekas, On the Goldstein-Levitin-Polyak gradient projection method, IEEE Transactions on Automatic Control, 21 (1976), pp. 174-184.
[6] D. Bertsekas, Projected newton methods for optimization problems with simple constraints, SIAM Journal on Control and Optimization, 20 (1982), pp. 221-246.
[7] R. Bro, Multi-way Analysis in the Food Industry: Models, Algorithms, and Applications, PhD thesis, Universiteit van Amsterdam, 1998.
[8] R. Bro and S. D. Jong, A fast non-negativity-constrained least squares algorithm, Journal of Chemometrics, 11 (1997), pp. 393-401.
[9] R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu, A limited memory algorithm for bound constrained optimization, SIAM Journal on Scientific Computing, 16 (1995), pp. 1190-1208.
[10] J. D. Carroll and J. J. Chang, Analysis of individual differences in multidimensional scaling via an N-way generalization of 'Eckart-Young' decomposition, Psychometrika, 35 (1970), pp. 283-319.
[11] E. C. Chi and T. G. Kolda, On tensors, sparsity, and nonnegative factorizations, SIAM Journal on Matrix Analysis and Applications, 33 (2012), pp. 1272-1299.
[12] A. Cichocki and A.-H. Phan, Fast local algorithms for large scale nonnegative matrix and tensor factorizations, IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, 92 (2009), pp. 708-721.
[13] A. Conn, N. Gould, and P. Toint, Global convergence of a class of trust region algorithms for optimization with simple bounds, SIAM Journal on Numerical Analysis, 25 (1988), pp. 433-460.
[14] DBLP data, 2011. http://www.informatik.uni-trier.de/~ley/db/.
[15] D. M. Dunlavy, T. G. Kolda, and E. Acar, Temporal link prediction using tensor and matrix factorizations, ACM Transactions on Knowledge Discovery from Data, 5 (2011).
[16] M. P. Friedlander and K. Hatz, Computing nonnegative tensor factorizations, Computational Optimization and Applications, 23 (2008), pp. 631-647.
[17] E. Gonzalez and Y. Zhang, Accelerating the Lee-Seung algorithm for non-negative matrix factorization, Tech. Report TR-05-02, Department of Computational and Applied Mathematics, Rice University, Houston, TX, 2005.
[18] R. A. Harshman, Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multi-modal factor analysis, UCLA Working Papers in Phonetics, 16 (1970). Available online, http://publish.uwo.ca/~harshman/wpppfac0.pdf
[19] C.-J. Hsieh and I. S. Dhillon, Fast coordinate descent methods with variable selection for non-negative matrix factorization, in Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2011, pp. 1064-1072.
[20] D. Kim, S. Sra, and I. Dhillon, Tackling box-constrained optimization via a new projected quasi-Newton approach, SIAM Journal on Scientific Computing, 32 (2010), pp. 3548-3563.
[21] D. Kim, S. Sra, and I. S. Dhillon, Fast projection-based methods for the least squares nonnegative matrix approximation problem, Statistical Analysis and Data Mining, 1 (2008), pp. 38-51.
[22] H. Kim and H. Park, Nonnegative matrix factorization based on alternating nonnegativity constrained least squares and active set method, SIAM Journal on Matrix Analysis and Applications, 30 (2008), pp. 713-730.
[23] J. Kim and H. Park, Fast nonnegative tensor factorization with an active-set-like method, in High-Performance Scientific Computing, Algorithms and Applications, M. W. Berry, K. A. Gallivan, E. Gallopoulos, A. Grama, B. Philippe, Y. Saad, and F. Saied, eds., Springer, 2012, pp. 311-326.
[24] T. G. Kolda and B. W. Bader, Tensor decompositions and applications, SIAM Review, 51 (2009), pp. 455-500.
[25] D. D. Lee and H. S. Seung, Learning the parts of objects by non-negative matrix factorization, Nature, 401 (1999), pp. 788-791.
[26] ——, Algorithms for non-negative matrix factorization, Advances in Neural Information Processing Systems, 13 (2001), pp. 556-562.
[27] C. Lin, Projected gradient methods for nonnegative matrix factorization, Neural computation, 19 (2007), pp. 2756-2779.
[28] J. Liu, J. Liu, P. Wonka, and J. Ye, Sparse non-negative tensor factorization using columnwise coordinate descent, Pattern Recognition, 45 (2012), pp. 649-656.
[29] J. Nocedal, Updating quasi-Newton matrices with limited storage, Mathematics of Computation, 35 (1980), pp. 773-782.
[30] J. Nocedal and S. J. Wright, Numerical Optimization, Springer, 2nd ed., 2006.
[31] P. PaAtero, A weighted non-negative least squares algorithm for three-way "PARAFAC" factor analysis, Chemometrics and Intelligent Laboratory Systems, 38 (1997), pp. 223-242.
[32] P. Paatero and U. Tapper, Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values, Environmetrics, 5 (1994), pp. 111-126.
[33] A. Phan, A. Cichocki, R. Zdunek, and T. Dinh, Novel alternating least squares algorithm for nonnegative matrix and tensor factorizations, in Neural Information Processing. Theory and Algorithms, K. W. Wong, B. S. U. Mendis, and A. Bouzerdoum, eds., vol. 6443 of Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2010, pp. 262-269.
[34] A. H. Phan, P. Tichavsky, and A. Cichocki, Fast damped Gauss-Newton algorithm for sparse and nonnegative tensor factorization, in Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on, IEEE, 2011, pp. 1988-1991.
[35] M. Schmidt, D. Kim, and S. Sra, Projected Newton-type methods in machine learning, in Optimization for Machine Learning, S. Sra, S. Nowozin, and S. J. Wright, eds., MIT Press, 2011, pp. 305-330.
[36] J. Sun, D. TaO, and C. Faloutsos, Beyond streams and graphs: Dynamic tensor analysis, in KDD '06, Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2006, pp. 374-383.
[37] G. Tomasi and R. Bro, A comparison of algorithms for fitting the PARAFAC model, Computational Statistics and Data Analysis, 50 (2006), pp. 1700-1734.
[38] S. Vavasis, On the complexity of nonnegative matrix factorization, SIAM Journal on Optimization, 20 (2009), pp. 1364-1377.
[39] M. Welling and M. Weber, Positive tensor factorization, Pattern Recognition Letters, 22 (2001), pp. 1255-1261.
[40] Y. XU and W. Yin, A block coordinate descent method for multi-convex optimization with applications to nonnegative tensor factorization and completion, Tech. Report 12-15, CAAM Rice University, Houston, TX, 2012.
[41] S. Zafeiriou and M. Petrou, Nonnegative tensor factorization as an alternative CsiszarTusnady procedure: algorithms, convergence, probabilistic interpretations and novel probabilistic tensor latent variable analysis algorithms, Data Mining and Knowledge Discovery, 22 (2011), pp. 419-460.
[42] R. Zdunek and A. Cichocki, Non-negative matrix factorization with quasi-Newton optimization, in Eighth International Conference on Artificial Intelligence and Soft Computing, ICAISC, Springer, 2006, pp. 870-879.
[43] ——, Nonnegative matrix factorization with constrained second-order optimization, Signal Processing, 87 (2007), pp. 1904-1916.
[44] Y. Zheng and Q. Zhang, Damped Newton based iterative non-negative matrix factorization for intelligent wood defects detection, Journal of Software, 5 (2010), pp. 899-906.
Appendix A. Generating Synthetic Test Data.
The goal is to create artificial nonnegative factor matrices and use these to compute a data tensor whose elements follow a Poisson distribution with multilinear parameters. Factorizing the data tensor should yield quantities that are close to the original factor matrices. The procedure is based on the work of [11.

The data tensor should be sparse, reflecting Poisson distributions whose probability of zero is not negligible. Each entry is a count of the number of samples assigned to this cell, out of a given total number of samples $S$. We generate factor matrices in each tensor mode and treat them as stochastic quantities to draw the $S$ samples that provide data tensor counts.

Our generation procedure utilizes the function create_problem from Tensor Toolbox for Matlab [4, supplying a custom function for the Factor_Generator parameter (available as Matlab code from the authors). We create a multilinear model, $\mathcal{M}=\left[\boldsymbol{\lambda} ; \mathbf{A}^{(1)} \ldots \mathbf{A}^{(N)}\right]$, where $\mathbf{A}^{(n)} \in \mathcal{R}^{I_{n} \times R}$ and $\boldsymbol{\lambda} \in \mathcal{R}^{R}$, and sizes $I_{n}$ and $R$ are given. The model is generated by the following procedure:

```
Algorithm 5 Generation of Synthetic Sparse Poisson Tensor Data
\(\overline{\text { Given tensor size, } I_{1} \times \cdots \times I_{N} \text {, number of components, } R \text {, and number of samples, }}\)
\(S\).
Return a model \(\mathcal{M}=\left[\boldsymbol{\lambda} ; \mathbf{A}^{(1)} \ldots \mathbf{A}^{(N)}\right]\) and corresponding sparse data tensor \(\boldsymbol{X}\).
1: In each column of \(\mathbf{A}^{(n)}\), choose \(20 \%\) of the elements at random and set their value to \(1+10 R x\), where \(x\) is a random value from a uniform distribution on \((0,1)\). Set the other elements equal to the small constant 0.1.
Choose random values for elements in \(\boldsymbol{\lambda}\) from a uniform distribution on \((0,1)\).
Rescale each column of \(\mathbf{A}^{(n)}\) so entries sum to 1 , absorbing the scale factor into the corresponding element of \(\boldsymbol{\lambda}\).
Rescale the vector \(\boldsymbol{\lambda}\) so entries sum to 1 .
for \(s=1, \ldots, S\) do
Treating \(\boldsymbol{\lambda}\) as a distribution, choose a component \(r\) at random.
Treating the \(r\) th column of \(\mathbf{A}^{(1)}\) as a distribution, choose an index \(i_{1}\) with probability proportional to \(\mathbf{a}_{r}^{(1)}\). Do the same for indices \(i_{2}, \ldots, i_{N}\), resulting in the index \(\mathbf{i}\) chosen with probability
```

$$
P(\mathbf{i})=a_{i_{1} r}^{(1)} a_{i_{2} r}^{(2)} \ldots a_{i_{N} r}^{(N)}
$$

8: $\quad$ Increment the $\mathbf{i t h}$ entry of $\mathcal{X}$ by one.
end for
Rescale $\boldsymbol{\lambda} \leftarrow S \boldsymbol{\lambda}$ so that $\|\boldsymbol{\lambda}\|_{1}=S . \quad \triangleright$ Recall Step 4 sets $\|\boldsymbol{\lambda}\|_{1}=1$.

Step 1 defines a strong preference for certain values of each index. As $R$ increases,
the relative probability of these indices is also increased so that they continue to stand out as strong preferences.

Step 10 rescales $\boldsymbol{\lambda}$ so that the $\ell_{1}$ norms of the generated data tensor $\boldsymbol{X}$ and K-tensor are the same in any mode- $n$ unfolding. For example:

$$
\begin{aligned}
\left\|\mathbf{X}_{(1)}\right\|_{1} & =\left\|\mathbf{A}^{(1)} \mathbf{\Lambda}\left(\mathbf{A}^{(N)} \odot \ldots \odot \mathbf{A}^{(2)}\right)^{T}\right\|_{1} \\
& =\sum_{i=1}^{I_{1}} \sum_{r=1}^{R} \lambda_{r} a_{r i}^{(1)} \\
& =\sum_{r=1}^{R} \lambda_{r}
\end{aligned}
$$

The second equality uses the fact that rows of the Khatri-Rao product sum to one when columns of $\mathbf{A}^{(n)}$ sum to one (see the comments after equation 2.2).

## Appendix B. Supplementary Material.

This supplementary section provides detailed results for some experiments mentioned in the paper. Refer to Section 4 of the paper for a description of default parameters used for the algorithms and characteristics of the workstation used for testing.
B.1. Performance on $200 \times 1000$ Data Set. Section 1.1 of the paper states that a 2-way tensor of size $200 \times 1000$ was too large to factorize by the method of Zdunek and Cichocki [42] but not difficult for our algorithms. Here we support our claim with experimental evidence. Reference [42] does not specify tensor details except that the desired factorization has $R=10$ components. We generated a synthetic data tensor according to the procedure in Appendix A of the paper, modifying Step 1 to boost $9 \%$ of the elements to $1+4 R x$, where $x$ is a random number chosen from a uniform distribution on $(0,1)$. We tested $R=10$ and $R=50$ components. We compare algorithms PDN-R and PQN-R using default parameters (see Section 4 for values) over ten instances of synthetic data. Table B.1 shows characteristics of the data, and results in Table B.2 demonstrate that the problems are solved to high accuracy in under ten minutes.

Table B.1: Sparsity of synthetic tensors versus the number of components $R$, for tensors of size $200 \times 1000$, generated by boosting $9 \%$ of the elements to $1+4 R x$. Number of nonzeros is the average over ten tensors.

| $R$ | Number Nonzeros | Density |
| ---: | :---: | :---: |
| 10 | 50,144 | $25.1 \%$ |
| 50 | 61,826 | $30.9 \%$ |

Table B.2: Time to reach various stop tolerances for tensors of size $200 \times 1000$. Mean and standard deviation are reported over ten runs. Both methods reach the tolerance of $\tau=10^{-4}$ in well under ten minutes for $R=50$ components.

|  | $R$ | PDN-R | PQN-R |
| :--- | ---: | :---: | :---: |
| $\tau=10^{-2}$ | 10 | $63 \pm 26$ secs | $155 \pm 84$ secs |
|  | 50 | $170 \pm 31$ secs | $311 \pm 59$ secs |
| $\tau=10^{-3}$ | $R$ | PDN-R | PQN-R |
|  | 10 | $80 \pm 25$ secs | $184 \pm 100$ secs |
|  | 50 | $191 \pm 26$ secs | $357 \pm 86$ secs |
| $\tau=10^{-4}$ | $R$ | PDN-R | PQN-R |
|  | 10 | $91 \pm 25$ secs | $208 \pm 116$ secs |
|  | 50 | $231 \pm 59$ secs | $385 \pm 84$ secs |

B.2. Detail for Results in Section 4.2.2 . Section 4.2 .2 of the paper contains a table showing the performance of algorithms PDN-R, PQN-R, and MU on a tensor of size $200 \times 300 \times 400$ for different values of $R$. The table reports execution time for the algorithms to reach a stop tolerance of $\tau=10^{-3}$. Supplementary Tables B.3 B. 6 contain additional results from the same tests. Table B. 3 shows that the relative advantage of the algorithms holds for all tolerances; in particular, PDN-R and PQN-R are faster than MU at all tolerances, and PQN-R becomes faster than PDN-R for a sufficiently large number of components. Table B. 4 shows that the final values of the objective function attained by the three algorithms are nearly identical.

Tables B.5 and B. 6 report the agreement between the models computed by each algorithm and the factor matrices from which tensor data was generated (the "true model"). Agreement is measured by the Tensor Toolbox score function, which implements the method described in [1]. The score considers the angles between all factor matrix column vectors. Let the two $R$-component models for a three-way tensor have weight vectors and factor matrices $\left\{\boldsymbol{\lambda}^{A}, \mathbf{A}^{(1)}, \mathbf{A}^{(2)}, \mathbf{A}^{(3)}\right\}$ and $\left\{\boldsymbol{\lambda}^{B}, \mathbf{B}^{(1)}, \mathbf{B}^{(2)}, \mathbf{B}^{(3)}\right\}$. Assume columns of factor matrices are normalized to one in the $\ell_{2}$ norm, adjusting the weights as needed. The comparison is computed from

$$
\begin{equation*}
s=\frac{1}{R} \sum_{r=1}^{R}\left[\left(\mathbf{a}_{r}^{(1)}\right)^{T} \mathbf{b}_{r}^{(1)}\right]\left[\left(\mathbf{a}_{r}^{(2)}\right)^{T} \mathbf{b}_{r}^{(2)}\right]\left[\left(\mathbf{a}_{r}^{(3)}\right)^{T} \mathbf{b}_{r}^{(3)}\right] . \tag{B.1}
\end{equation*}
$$

The score is the largest $s$ over column permutations of the models. Ideally, $s$ is computed for all possible permutations, but this is not feasible when $R$ is large. Instead, we use the greedy option to limit the number of permutations considered. The score is a number between zero and one, with one indicating a perfect match between the two models.

Table B.5 shows that all three algorithms produce solutions with scores in the range of $[0.84,0.92]$, with no clear superiority for any algorithm. To give a sense of score magnitudes, the table also reports the score between computed solutions and factor matrices used to generate other models. The generation procedure in Appendix A is based on random values, so we expect poor scores between a solution and random factor matrices. This is indeed the case, with no scores greater than 0.01. Table B. 6 reports the score for computed solutions when the algorithm is initialized

Table B.3: Time to reach various stop tolerances $\left(\tau=10^{-3}\right.$ repeats Table 6 in Section 4.2.2 of the paper). Mean and standard deviation are reported over ten runs. The last column for MU is the number of runs that failed to reach the stop tolerance after three hours ( 10,800 seconds) of execution.

|  | $R$ | PDN-R | PQN-R | MU |  |
| :---: | :---: | :---: | :---: | :---: | :---: |
|  | 20 | $170 \pm 37 \mathrm{secs}$ | $317 \pm 127$ secs | $997 \pm 347$ secs | (0 failures) |
| $\tau=10^{-2}$ | 40 | $339 \pm 71$ secs | $573 \pm 268$ secs | $3275 \pm 1357$ secs | (0 failures) |
|  | 60 | $611 \pm 126$ secs | $740 \pm 235$ secs | $5758 \pm 1275$ secs | (0 failures) |
|  | 80 | $1145 \pm 358$ secs | $1340 \pm 358$ secs | $9134 \pm 1030$ secs | (1 failure) |
|  | 100 | $1739 \pm 359$ secs | $1623 \pm 583$ secs | $9687 \pm 372$ secs | (6 failures) |
| $\tau=10^{-3}$ | $R$ | PDN-R | PQN-R | MU |  |
|  | 20 | $229 \pm 57 \mathrm{secs}$ | $397 \pm 123$ secs | $3355 \pm 1933$ secs | (0 failures) |
|  | 40 | $493 \pm 151 \mathrm{secs}$ | $818 \pm 185$ secs | $8101 \pm 2045$ secs | (2 failures) |
|  | 60 | $1003 \pm 349$ secs | $966 \pm 286$ secs | $9628 \pm 978$ secs | (5 failures) |
|  | 80 | $1682 \pm 642$ secs | $1639 \pm 390$ secs | - | (10 failures) |
|  | 100 | $2707 \pm 773$ secs | $1995 \pm 743$ secs | - | (10 failures) |
| $\tau=10^{-4}$ | $R$ | PDN-R | PQN-R | MU |  |
|  | 20 | $270 \pm 67 \mathrm{secs}$ | $451 \pm 108$ secs | $5207 \pm 1975$ secs | (1 failure) |
|  | 40 | $756 \pm 383$ secs | $905 \pm 171$ secs | - | (10 failures) |
|  | 60 | $1166 \pm 390$ secs | $1102 \pm 305$ secs | - | (10 failures) |
|  | 80 | $1890 \pm 617$ secs | $1859 \pm 413$ secs | - | (10 failures) |
|  | 100 | $2834 \pm 741$ secs | $2280 \pm 808$ secs | - | (10 failures) |

Table B.4: Final objective function values reached by the algorithms. Algorithms minimize K-L divergence, but the negative of K-L is shown because this equals the (unnormalized) maximum likelihood. Mean and maximum (most optimal) values are reported over ten runs. Values are in units of $10^{6}$; for example, the average unnormalized log likelihood reached by algorithm PDN-R for $R=20$ was $1.146 \times 10^{6}$. The likelihood increases with the number of components because this allows for a better fit to the data.

|  | PDN-R |  | PQN-R |  | MU |  |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| $R$ | MEAN | MAX | MEAN | MAX | MEAN | MAX |
| 20 | 1.146 | 1.205 | 1.144 | 1.193 | 1.145 | 1.193 |
| 40 | 1.448 | 1.475 | 1.448 | 1.475 | 1.447 | 1.473 |
| 60 | 1.610 | 1.643 | 1.611 | 1.647 | 1.610 | 1.640 |
| 80 | 1.716 | 1.735 | 1.715 | 1.734 | 1.715 | 1.738 |
| 100 | 1.794 | 1.811 | 1.794 | 1.810 | 1.794 | 1.811 |

with the factor matrices of the true model. The solution should be a local minimum close to the true model, representing an easily found solution that gives one of the best possible fits to the data. Since tensor data is generated with statistical noise, a perfect fit is not seen. These "ideal" scores are slightly higher than the scores from algorithm solutions in Table B.5.

We notice that scores in Table B. 6 decrease as $R$ increases. This is because the tensor data has about the same number of samples in each test case (see Table B.7),
but a model with more components has more variables. As the ratio of data samples to model variables decreases, the variables are affected more strongly by noise in the data, leading to solutions with a lower score.

Table B.5: Comparison of final computed solutions with the true model from which data was generated. The comparison uses the Tensor Toolbox score function, executed with option greedy. The score is a number between zero (poor match) and one (exact match). The TRUE MODEL column shows the lowest (worst) score obtained over ten runs when the computed solution is scored against the true model. For comparison, the OTHERS column shows the highest (best) score between the computed solution and any of the nine other models, again averaged over ten runs.

|  | PDN-R |  | PQN-R |  | MU |  |
| ---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $R$ | TRUE MODEL | OTHERS | TRUE MODEL | OTHERS | TRUE MODEL | OTHERS |
| 20 | 0.919 | 0.009 | 0.845 | 0.009 | 0.854 | 0.009 |
| 40 | 0.880 | 0.010 | 0.888 | 0.010 | 0.892 | 0.010 |
| 60 | 0.846 | 0.010 | 0.873 | 0.010 | 0.844 | 0.010 |
| 80 | 0.876 | 0.010 | 0.873 | 0.010 | 0.889 | 0.010 |
| 100 | 0.865 | 0.010 | 0.864 | 0.010 | 0.847 | 0.010 |

Table B.6: Comparison of final computed solutions with the true model from which data was generated using the Tensor Toolbox score function. The PDN-R algorithm was initialized with the factor matrices of Appendix A that generated the test data. Columns in the table show the highest (best) and lowest (worst) scores obtained over ten runs.

|  | PDN-R |  |
| ---: | :---: | :---: |
| $R$ | BEST SCORE | WORST SCORE |
| 20 | 0.996 | 0.961 |
| 40 | 0.989 | 0.964 |
| 60 | 0.982 | 0.913 |
| 80 | 0.965 | 0.937 |
| 100 | 0.952 | 0.907 |

B.3. Section 4.2.2 Repeated with Different Sparsity. The data tensor of Section 4.2.2 was generated by the method described in Step 1 of Appendix A with $20 \%$ of element values boosted to $1+10 R x$. In this section, we report on convergence of the algorithms for tensors of the same size but different sparsity. Sparsity was modified by boosting different numbers of elements. The boosting procedure provides tensor data with appropriate Poisson distributions but does not give exact control over the number of nonzeros. We select boost parameters that yield approximately the same number of nonzeros for $R \in\{20,40,60,80,100\}$ so we can more easily compare algorithm performance as a function of $R$. Tables B.7-B.9 describe the sparsity of the different test tensors.

Table B.7: Sparsity of synthetic tensors versus the number of components $R$, for tensors of size $200 \times 300 \times 400$, generated by boosting $20 \%$ of the elements to $1+10 R x$. Number of nonzeros is the average over ten tensors. (This table repeats Table 1 from the paper.)

| $R$ | Number Nonzeros | Density |
| ---: | :---: | :---: |
| 20 | 413,458 | $1.72 \%$ |
| 40 | 450,756 | $1.88 \%$ |
| 60 | 464,443 | $1.94 \%$ |
| 80 | 470,950 | $1.96 \%$ |
| 100 | 475,455 | $1.98 \%$ |

Table B.8: Sparsity of synthetic tensors versus the number of components $R$, for tensors of size $200 \times 300 \times 400$, generated by boosting $5 \%$ of the elements to $1+2 R x$. Number of nonzeros is the average over ten tensors.

| $R$ | Number Nonzeros | Density |
| ---: | :---: | :---: |
| 20 | 158,616 | $0.66 \%$ |
| 40 | 141,778 | $0.59 \%$ |
| 60 | 148,273 | $0.62 \%$ |
| 80 | 161,212 | $0.67 \%$ |
| 100 | 177,060 | $0.74 \%$ |

Table B.9: Sparsity of synthetic tensors versus the number of components $R$, for tensors of size $200 \times 300 \times 400$, generated by boosting $3 \%$ of the elements to $1+10 R x$. Number of nonzeros is the average over ten tensors.

| $R$ | Number Nonzeros | Density |
| ---: | :---: | :---: |
| 20 | 55,471 | $0.23 \%$ |
| 40 | 44,862 | $0.19 \%$ |
| 60 | 47,171 | $0.20 \%$ |
| 80 | 51,827 | $0.22 \%$ |
| 100 | 57,700 | $0.24 \%$ |

The experiments in Section 4.2.2 were performed for each of these tensors. Supplementary Table B.3 in the previous section shows algorithm performance for the
sparse tensor of Table B.7 ( $20 \%$ boost). The tables below show performance for the $5 \%$ and $3 \%$ boosted tensors. We see that the conclusions from Section 4.2 .2 hold for these cases as well: PDN-R and PQN-R are faster than MU in nearly every case, and PQN-R becomes faster than PDN-R for sufficiently large $R$. Comparing the three tables with each other, we see that performance improves when the number of nonzeros in the data decreases.

Table B.10: Time to reach various stop tolerances $\tau$ on the tensor with the sparsity in supplementary Table B. 8 resulting from $5 \%$ boost. Mean and standard deviation are reported over ten runs. The last column for MU is the number of runs that failed to reach the stop tolerance after three hours of execution.

|  | $R$ | PDN-R | PQN-R | MU |
| :---: | :---: | :---: | :---: | :---: |
|  | 20 | $165 \pm 40$ secs | $353 \pm 137$ secs | $345 \pm 214$ secs |
| $\tau=10^{-2}$ | 40 | $393 \pm 125$ secs | $454 \pm 148$ secs | $541 \pm 129$ secs |
|  | 60 | $516 \pm 124$ secs | $394 \pm 103$ secs | $865 \pm 186$ secs |
|  | 80 | $558 \pm 48$ secs | $386 \pm 53$ secs | $1235 \pm 210$ secs |
|  | 100 | $658 \pm 59$ secs | $343 \pm 44$ secs | $1633 \pm 228$ secs |
|  | $R$ | PDN-R | PQN-R | MU |
|  | 20 | $270 \pm 139$ secs | $580 \pm 323$ secs | no successes |
| $=10^{-3}$ | 40 | $444 \pm 145$ secs | $509 \pm 147$ secs | no successes |
|  | 60 | $623 \pm 128$ secs | $475 \pm 135$ secs | no successes |
|  | 80 | $726 \pm 130$ secs | $487 \pm 46$ secs | no successes |
|  | 100 | $893 \pm 150$ secs | $442 \pm 86$ secs | no successes |
|  | $R$ | PDN-R | PQN-R | MU |
|  | 20 | $287 \pm 139$ secs | $585 \pm 320$ secs | no successes |
| $=10^{-4}$ | 40 | $456 \pm 142$ secs | $526 \pm 146$ secs | no successes |
|  | 60 | $700 \pm 199$ secs | $566 \pm 193$ secs | no successes |
|  | 80 | $810 \pm 214$ secs | $662 \pm 119$ secs | no successes |
|  | 100 | $964 \pm 173$ secs | $631 \pm 187$ secs | no successes |

Table B.11: Time to reach various stop tolerances $\tau$ on the tensor with with the sparsity in supplementary Table B. 9 resulting from $3 \%$ boost. Mean and standard deviation are reported over ten runs. The last column for MU is the number of runs that failed to reach the stop tolerance after three hours of execution.

|  | $R$ | PDN-R | PQN-R | MU |  |
| :---: | :---: | :---: | :---: | :---: | :---: |
|  | 20 | $143 \pm 75$ secs | $182 \pm 73$ secs | $108 \pm 34$ secs | (0 failures) |
| $\tau=10^{-2}$ | 40 | $141 \pm 19$ secs | $134 \pm 53$ secs | $143 \pm 49$ secs | (0 failures) |
|  | 60 | $174 \pm 29$ secs | $114 \pm 23$ secs | $218 \pm 46$ secs | (0 failures) |
|  | 80 | $219 \pm 30 \mathrm{secs}$ | $114 \pm 20$ secs | $347 \pm 69$ secs | (0 failures) |
|  | 100 | $230 \pm 35 \mathrm{secs}$ | $166 \pm 37 \mathrm{secs}$ | $481 \pm 86$ secs | (0 failures) |
| $\tau=10^{-3}$ | $R$ | PDN-R | PQN-R | MU |  |
|  | 20 | $172 \pm 80 \mathrm{secs}$ | $208 \pm 77 \mathrm{secs}$ | $278 \pm 135$ secs | (0 failures) |
|  | 40 | $169 \pm 40 \mathrm{secs}$ | $151 \pm 41$ secs | $363 \pm 226$ secs | (0 failures) |
|  | 60 | $328 \pm 122$ secs | $139 \pm 37$ secs | $552 \pm 229$ secs | (0 failures) |
|  | 80 | $290 \pm 70 \mathrm{secs}$ | $135 \pm 21$ secs | $897 \pm 302$ secs | (0 failures) |
|  | 100 | $302 \pm 84$ secs | $190 \pm 55$ secs | $1320 \pm 397$ secs | (0 failures) |
| $\tau=10^{-4}$ | $R$ | PDN-R | PQN-R | MU |  |
|  | 20 | $178 \pm 80 \mathrm{secs}$ | $213 \pm 76$ secs | - | (10 failures) |
|  | 40 | $237 \pm 188$ secs | $159 \pm 39$ secs | - | (10 failures) |
|  | 60 | $342 \pm 127 \mathrm{secs}$ | $176 \pm 65$ secs | - | (10 failures) |
|  | 80 | $290 \pm 70$ secs | $189 \pm 94$ secs | - | (10 failures) |
|  | 100 | $321 \pm 98$ secs | $259 \pm 129$ secs | - | (10 failures) |

B.4. Collinear Factor Matrices. Tensor data in these experiments was designed to reflect underlying factor matrices that have nearly collinear columns. Algorithms PDN-R and PQN-R do much better than MU in these experiments. The basic idea, proposed in Phan et al. [33], is to generate random factor matrices and then modify column vectors according to

$$
\begin{equation*}
\mathbf{a}_{r}^{(n)}=\mathbf{a}_{1}^{(n)}+\alpha \mathbf{a}_{r}^{(n)} \quad \text { for } r \in 2, \ldots, R, \tag{B.2}
\end{equation*}
$$

with $\alpha=0.5$. We generated synthetic data according to Appendix A but added the collinearity modification after Step 1.

The modification significantly impacts the sparsity of the generated tensor, adding many nonzero elements because the boosted elements in column $r=1$ become boosted elements in all other columns. This makes it difficult to compare performance "before" and "after" the collinearity modification. Instead, we made two experiments with different values of $\alpha$ and view them as two different results.

Collinearity is measured as the cosine between pairs of column vectors. For realvalued $N$-vectors $\mathbf{x}$ and $\mathbf{y}$,

$$
\cos (\mathbf{x}, \mathbf{y})=\frac{\mathbf{x}^{T} \mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|}
$$

If elements of $\mathbf{x}$ and $\mathbf{y}$ are independent random variables chosen from a uniform distribution on $(0,1)$, then the expected value of their cosine is 0.75 . However, the sparse columns generated by Appendix A are not uniform. For instance, average collinearity among column pairs in the experiments of supplementary Section B. 2 is close to 0.10 .

Tables B.12-B.15 show tensor characteristics and computational performance for two experiments. The tensors in the experiments have the same size but use a different value of $\alpha$ in equation ( $\bar{B} .2$ ) to modify collinearity, which also changes their sparsity; hence, the two experiments should not be compared with each other. What they each show is that algorithms PDN-R and PQN-R are much faster than MU, especially when high accuracy is desired.

Table B.12: Sparsity and collinearity of synthetic tensors versus the number of components $R$, for tensors of size $50 \times 50 \times 50$, generated by boosting $10 \%$ of the elements to $1+10 R x$, and modifying collinearity using $\alpha=0.5$ in equation B.2. All Pairs shows the average collinearity between all pairs of column vectors within each factor matrix. Data under $\mathbf{a}_{1}$ Pairs shows the average collinearity between the $\mathbf{a}_{1}^{(n)}$ column vector and all others within each factor matrix. All measurements are an average over ten tensors.

|  | Sparsity |  | Collinearity |  |
| :---: | :---: | :---: | :---: | :---: |
| $R$ | Number Nonzeros | Density | All Pairs | $\mathbf{a}_{1}$ Pairs |
| 10 | 16,442 | $13.2 \%$ | 0.839 | 0.900 |
| 20 | 14,908 | $11.9 \%$ | 0.828 | 0.897 |
| 30 | 15,572 | $12.5 \%$ | 0.830 | 0.900 |
| 40 | 16,892 | $13.5 \%$ | 0.800 | 0.882 |

Table B.13: Time to reach various stop tolerances $\tau$ on tensors of size $50 \times 50 \times 50$, generated by boosting $10 \%$ of the elements to $1+10 R x$, and modifying collinearity using $\alpha=0.5$ in equation (B.2). Mean and standard deviation are reported over ten runs. The last column for MU is the number of runs that failed to reach the stop tolerance after 1000 seconds of execution.


Table B.14: Sparsity and collinearity of synthetic tensors versus the number of components $R$, for tensors of size $50 \times 50 \times 50$, generated by boosting $10 \%$ of the elements to $1+10 R x$, and modifying collinearity using $\alpha=0.1$ in equation B.2. All Pairs shows the average collinearity between all pairs of column vectors within each factor matrix. Data under $\mathbf{a}_{1}$ Pairs shows the average collinearity between the $\mathbf{a}_{1}^{(n)}$ column vector and all others within each factor matrix. All measurements are an average over ten tensors.

|  | Sparsity |  | Collinearity |  |
| :---: | :---: | :---: | :---: | :---: |
| $R$ | Number Nonzeros | Density | All Pairs | $\mathbf{a}_{1}$ Pairs |
| 10 | 9,432 | $7.55 \%$ | 0.991 | 0.995 |
| 20 | 8,572 | $6.86 \%$ | 0.990 | 0.995 |
| 30 | 8,828 | $7.06 \%$ | 0.991 | 0.995 |
| 40 | 9,419 | $7.54 \%$ | 0.988 | 0.993 |

Table B.15: Time to reach various stop tolerances $\tau$ on tensors of size $50 \times 50 \times 50$, generated by boosting $10 \%$ of the elements to $1+10 R x$, and modifying collinearity using $\alpha=0.1$ in equation (B.2). Mean and standard deviation are reported over ten runs. The last column for MU is the number of runs that failed to reach the stop tolerance after 1000 seconds of execution.

|  | $R$ | PDN-R | PQN-R | MU |  |
| :---: | :---: | :---: | :---: | :---: | :---: |
|  | 10 | $11 \pm 4$ secs | $16 \pm 4$ secs | $33 \pm 8$ secs | (0 failures) |
| $\tau=10^{-2}$ | 20 | $12 \pm 3$ secs | $23 \pm 8$ secs | $99 \pm 31$ secs | (0 failures) |
|  | 30 | $14 \pm 3$ secs | $30 \pm 6$ secs | $226 \pm 66$ secs | (0 failures) |
|  | 40 | $17 \pm 4$ secs | $30 \pm 7$ secs | $340 \pm 114$ secs | (0 failures) |
| $\tau=10^{-3}$ | $R$ | PDN-R | PQN-R | MU |  |
|  | 10 | $19 \pm 9$ secs | $26 \pm 6$ secs | $181 \pm 62$ secs | (0 failures) |
|  | 20 | $22 \pm 9$ secs | $32 \pm 7$ secs | $512 \pm 200$ secs | (2 failures) |
|  | 30 | $26 \pm 4$ secs | $41 \pm 6$ secs | $787 \pm 174$ secs | (7 failures) |
|  | 40 | $30 \pm 7$ secs | $43 \pm 8$ secs | $827 \pm 0$ secs | (9 failures) |
| $\tau=10^{-4}$ | $R$ | PDN-R | PQN-R | MU |  |
|  | 10 | $24 \pm 14$ secs | $28 \pm 6$ secs | $545 \pm 228$ secs | (2 failures) |
|  | 20 | $25 \pm 8$ secs | $36 \pm 11$ secs | - | (10 failures) |
|  | 30 | $30 \pm 5$ secs | $43 \pm 7$ secs | - | (10 failures) |
|  | 40 | $33 \pm 6$ secs | $45 \pm 8$ secs | - | (10 failures) |

