On Confidence-Constrained Rank Recovery in Topic Models

Behrouz Behmardi and Raviv Raich
School of EECS, Oregon State University, Corvallis, OR, 97331-5501
{behmardb,raich}@eecs.oregonstate.edu

***Abstract*—Topic models have been proposed to model a collection of data such as text documents and images in which each object (e.g., a document) contains a set of instances (e.g., words). In many topic models, the dimension of the latent topic space (the number of topics) is assumed to be a deterministic unknown. The number of topics significantly affects the prediction performance and interpretability of the estimated topics. In this paper, we propose a confidence-constrained rank minimization (CRM) to recover the exact number of topics in topic models with theoretical guarantees on recovery probability and mean squared error of the estimation. We provide a computationally-efficient optimization algorithm for the problem to further the applicability of the proposed framework to large real world datasets. Numerical evaluations are used to verify our theoretical results. Additionally, to illustrate the applicability of the proposed framework to practical problems, we provide results in image classification for two real world datasets and text classification for three real world datasets.**

***Index Terms*—Topic models, Low-rank matrix recovery, Nuclear norm minimization, Confidence constraints, Rank estimation.**

I. Introduction

In many applications of machine learning, such as text classification, image processing, and web classification, a multi-instance representation of objects is commonly used [1], [2]. In multi-instance datasets, an object is represented as a set of instances or bag of instances instead of a single instance. For example, in a corpus of documents, a document $(\textit{object})$ comprises of words $(\textit{instances})$. Often, distributions can be considered to represent multi-instance data. For example, in a multi-instance discrete dataset such as documents, the bag-of-words is a representation of a histogram over a given vocabulary. Due to the high dimensional nature of objects in multi-instance datasets (e.g., a usual vocabulary size in a corpus of documents can be about $20,000$), it is beneficial to simplify the representation of objects in multi-instance datasets by exploring the inner structure of such datasets. The framework of topic models introduces a low dimensional structure by associating documents with a low dimensional distributions over a small set of topics. In the generative approach to topic models, a subset of topics is first selected and the document is generated based on selecting words from the assigned topics. Some of the early well-known topic models are latent semantic indexing (LSI) [3], probabilistic latent semantic indexing (pLSI) [4], and latent Dirichlet allocation (LDA) [5]. We refer the reader to [6] for review on more recent developed topic models.

The number of topics (dimension of the latent space) has a significant effect on the quality of the model and interpretability of the estimated topics [5]. Heuristically, this problem is addressed in the literature by scanning through a range of numbers of topics and comparing performance measures such as perplexity on a held-out dataset or classification accuracy across the range [4], [5], [7]. In [8], it is mentioned that overestimating the number of topics can be remedied by ranking the topics and removing those which are not related to the theme of the data. Bayesian nonparametric topic models [9]–[11] provide a solution using Hierarchical Dirichlet Processes (HDP). The associated Bayesian inference is often regarded as a computationally complex approach [12]. A cross validation approach for selecting the number of topics in topic models is proposed in [13]. While this approach seems to be efficient in number of topics selection, different choices of held-out patterns and sizes have significant impact on the results. Term-by-document matrix is commonly used for data representation in topic models. The number of topics is the rank of such a matrix. Our interest is in devising a provable and computationally efficient method to jointly determine the rank and recover the term-by-document probability matrix from its noisy observation.

Constrained rank recovery of an unknown matrix has been studied vastly in the literature in the communities of signal processing, control system, and machine learning [14]–[16] in problems such as matrix completion [17] and matrix decomposition [18]. While for simple cases singular value decomposition (SVD) has been a common tool, in the constrained setting rank minimization presents additional challenges. One of the main challenges is the non-convex nature of the rank operator. Rank minimization is heuristically replaced with a nuclear norm minimization [19]–[23]. Nuclear norm minimization can be formulated as a semidefinite programming (SDP) and solved via general SDP solvers such as SDPT3 and SeDuMi. Although the convergence of these solvers is guaranteed, they can not be applied for a large scale problem due to the high computational complexity of Newton direction [24]–[26]. Due to the problem of computational complexity of SDP, several economical approaches have been developed. Most of these approaches are based on the idea of proximal point approximation (Moreau-Yosida regularization [27]) resulting in a closed-form solution for nuclear norm minimization [24]–[27]. An Augmented Lagrange multiplier (ALM) [28] is an

This work was partially supported by the Air Force Office of Scientific Research grant number FA9550-09-1-0471.

alternative which proposes to minimize the nuclear norm of the low-rank component plus l_1 norm of the sparse component with augmented Lagrange approach. These methods have been promising in terms of computational complexity. For example, in [28] robust PCA is implemented using only 20 iterations of a highly economical version of SVD. The conditions under which the low-rank matrix with missing entries can be estimated with high probability are proposed in [18], [21]. These methods have been applied to video surveillance and image recovery. We are interested in using rank recovery methods to determine the number of topics in topic models. However, we are faced with the following challenges. First, the observed term-by-document matrix is contaminated by a multinomial sampling noise as opposed to Gaussian noise [29], [30] or sparse noise [18]. Our problem includes a specific set of constraints such as positivity and sum-to-one which restrict the search space in the optimization problem.

In this paper, we present a framework and algorithms for a provable rank recovery in topic models. Specifically, our contributions in this paper are as follows: $1)$ We propose sufficient conditions for exact rank recovery in topic models as a rank minimization problem. $2)$ We provide a new framework of parameter free confidence-constrained convex optimization as an alternative to rank minimization problem, which can overcome the issues of Bayesian inferences such as $i)$ computational complexity associated with sampling methods, $ii)$ approximation associated with variational Bayes approach [31], and $iii)$ computational complexity associated with hyper-parameter tuning [32]. $3)$ We provide an analytical evaluation of the sufficient conditions for exact recovery of the number of topics in topic models. Moreover, we provide a bound on the sum of squared errors in terms of the model parameters such as number of documents, vocabulary size, and number of words in each document. $4)$ We provide an accelerated algorithm to solve the proposed convex optimization problem. We reformulate the problem in the dual form. By evaluating the duality gap, we are able to provide accuracy guarantees for the algorithm. $5)$ We evaluate our theoretical results on synthetic datasets. $6)$ Finally, we apply the proposed method on two image datasets and three real world text datasets to illustrate how the method can be applied to perform dimension reduction.

The rest of the paper is organized as follows. In Section II, the exact rank recovery in topic models is formulated. Section III introduces the method of confidence-constrained rank recovery in topic models. Section IV provides the theoretical guarantees for the proposed confidence-constrained rank minimization. In Section V, an accelerated gradient projection method for solving the dual form of confidence-constrained nuclear norm minimization is proposed. In Section VI, the evaluation of our theoretical results against the simulation is presented. Section VII illustrates how our method can be applied to image and text datasets. Finally, we summarize the paper in Section VIII along with the ideas for the future work.

II. Problem formulation

In this section, we present the problem of determining the number of topics in probabilistic topic models. We start with the generative process associated with the probabilistic topic model and then proceed with the formulation of identifying the number of topics in topic models. The theoretical framework for exact rank recovery proposed in this paper can be applied to topic models with the following properties: (i) The generative process involves a multinomial sampling from a probability matrix and (ii) the probability matrix can be decomposed as a product of two probability matrices. We carry out our derivation on the well-known LDA model.

A. Probabilistic topic models

Probabilistic topic models are generative models. Topic probabilities provide an explicit representation of documents in probabilistic topic models. The sampling process from this model can be explained as follows (for a list of notation, we refer the reader to Table I).

TABLE I
NOTATION USED IN THIS PAPER

Ψ	Term-by-document matrix
$\hat{\Psi}$	Sample term-by-document matrix
Ψ_0	Rank minimizing term-by-document matrix
M	Number of documents
L	Vocabulary size
T	Number of topics (Rank(Ψ))
n_d	Number of words in document d
σ_T	Smallest non-zero singular value of Ψ
θ_d	Per-document topic proportion
Φ	Topics matrix
z_{dj}	Per-word per-document topic assignment
α	Dirichlet prior parameter for topic proportion
β	Dirichlet prior for Topics matrix
λ	Lagrangian multiplier
n	$\min(n_d), d = 1, \ldots, M$

Each document is drawn in an i.i.d. fashion. For the dth document, $d = \{1, \ldots, M\}$, a random distribution of topics $p(z_{dj} = t|\theta) \triangleq \theta_d(t)$, $t \in \{1, \ldots, T\}$ is drawn. In LDA, $\theta_d \sim \text{Dir}(\alpha)$. Then, for jth word in document d, $j = \{1, \ldots, n_d\}$, a topic assignment z_{dj} is drawn, based on the topic distribution $\theta_d(t)$. Finally, word w_{dj} is drawn based on the conditional distribution $p(w_{dj} = l|z_{dj} = t, \Phi) \triangleq \Phi_{lt}, l = \{1, \ldots, L\}$. Note that Φ is a topics matrix where columns corresponds to topics $\{1, \ldots, T\}$ and rows correspond to vocabulary words. The graphical representation of LDA is shown in Fig. 1 and the precise sampling process for LDA is described in Algorithm 1. A key observation in topic models is that the probability

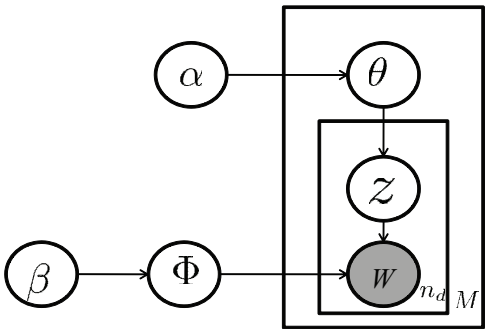


Fig. 1. The graphical model for LDA [33].

distribution of word w_{dj} can be obtained by marginalizing

the joint word-topic distribution over the topic:

$$p(w_{dj} = l|\theta_d) = \sum_{t=1}^{T} p(w_{dj} = l|z_{dj} = t, \Phi)p(z_{dj} = t|\theta_d). \quad (1)$$

To simplify the notation, we represent (1) in a matrix format,

$$\Psi = \Phi\theta, \quad (2)$$

where $\Psi_{ld} \triangleq p(w_{dj} = l|\theta_d)$, $\Psi \in \mathbb{R}^{L\times M}, \Phi \in \mathbb{R}^{L\times T}$, and $\theta \in \mathbb{R}^{T\times M}$. In other words, the vocabulary term-by-document matrix Ψ can be decomposed into the product of Φ and θ where Φ is the vocabulary probability per topic (topic matrix) and θ is the topic proportion per document. Note that the model in (2) is also applicable to pLSI. Columns of these matrices are probability vectors satisfying non-negativity and sum-to-one property. The introduction of latent topic variables allows for reduced dimension representation of the term-by-document matrix Ψ. The rank of the matrix Ψ is the number of topics T. We define the sample term-by-document matrix

Algorithm 1 Generative process for LDA

for $t = 1$ to T **do**
  Draw $\Phi_t \sim$ Dirichlet(β)
end for
for $d = 1$ to M **do**
  Draw $\theta_d \sim$ Dirichlet(α)
  for $j = 1$ to n_d **do**
    Draw $z_{dj} \sim$ Discrete(θ_d)
    Draw $w_{dj} \sim$ Discrete($\phi_{z_{dj}}$)
  end for
end for

$\hat{\Psi}$ as follows:

$$\hat{\Psi}_{ld} = \frac{1}{n_d}\sum_{j=1}^{n_d} I(w_{dj} = l). \quad (3)$$

Therefore, $n_d\hat{\Psi}_{\cdot d} \sim$ multinomial$(\Psi_{\cdot d}, n_d)$ which for notational ease we denote $\hat{\Psi} \sim$ norm-multinomial$(\Psi, \mathbf{n})$, where $\mathbf{n} = [n_1, \ldots, n_d]$.

B. Topics number recovery

Assume an unknown low-rank term-by-document matrix Ψ is obtained through the process explained in Section II-A. We observe matrix $\hat{\Psi} \sim$ norm-multinomial$(\Psi, \mathbf{n})$. Since $\hat{\Psi}$ could be full-rank due to the presence of noise in the sampling process, a straightforward examination of its singular values may not provide an immediate indication on the rank of Ψ. Furthermore, even if rank of the matrix Ψ is available, identifying a low-rank matrix Ψ which is similar to $\hat{\Psi}$ is a nontrivial problem. Specifically, we are interested in: 1) Estimating the term-by-document matrix Ψ from its noisy observations matrix $\hat{\Psi}$. 2) Quantifying the accuracy of the estimator of Ψ in two aspects: (i) Understanding the conditions under which the exact rank of the true matrix Ψ can be recovered. (ii) Characterizing the estimation error of the matrix Ψ associated with the matrix reconstruction. Note that we propose the estimation of the matrix Ψ rather than the decomposition of Ψ into the product of two probability matrices Φ and θ. While the connection is obvious, the problem of decomposing the estimated low-rank Ψ into the products of two probability matrices presents additional challenges which we reserve for future work.

III. Confidence-constrained rank recovery

In this section, we introduce the framework of confidence-constrained rank recovery. We start by describing the maximum likelihood (ML) solution for estimating matrix Ψ from its noisy observation $\hat{\Psi}$. Then, we introduce the regularized ML to address the problem of rank recovery. Finally, we conclude this section with the introduction of confidence-constrained rank minimization approach.

A. Unconstrained maximum likelihood

The log-likelihood for the probabilistic topic model in (1) can be written as follows [4]:

$$\mathcal{L} = \sum_{d=1}^{M}\sum_{l=1}^{L} n_{ld}\log\Psi_{ld}. \quad (4)$$

Using the fact that $n_{ld} = n_d\hat{\Psi}_{ld}$, we can rewrite the negative log-likelihood function as follows:

$$\sum_{d=1}^{M} n_d D_{kl}(\hat{\Psi}_{\cdot d}\|\Psi_{\cdot d}) = -\mathcal{L} + \Upsilon, \quad (5)$$

where $\Upsilon = \sum_{d=1}^{M} n_d \sum_{l=1}^{L} \hat{\Psi}_{ld}\log\hat{\Psi}_{ld}$ is a constant and $D_{kl}(p\|q) = \sum_k p_k \log\frac{p_k}{q_k}$. Hence, the unconstrained ML estimate of Ψ can be obtained using the following optimization

$$\begin{aligned}\hat{\Psi}_{ML} \quad = \quad & \arg\min_{\tilde{\Psi}} \sum_{d=1}^{M} n_d D_{kl}(\hat{\Psi}_{\cdot d}\|\tilde{\Psi}_{\cdot d}),\\ & \text{subject to} \qquad \tilde{\Psi} \geq 0,\\ & \qquad\qquad\quad 1^T\tilde{\Psi} = 1. \end{aligned} \quad (6)$$

Since the ML formulation does not incorporate information on rank of the matrix Ψ, its solution is the trivial $\hat{\Psi}_{ML} = \hat{\Psi}$ solution. In other words, even though the nonnegative $\sum_{d=1}^{M} n_d D_{kl}(\hat{\Psi}_{\cdot d}\|\tilde{\Psi}_{\cdot d})$ can be made zero by setting $\tilde{\Psi} = \hat{\Psi}$, the rank difference $|\text{Rank}(\tilde{\Psi}) - \text{Rank}(\Psi)|$ may be large. The ML approach in its unconstrained formulation advocates the potentially full rank matrix $\hat{\Psi}$ as an estimate for Ψ. In the following, we show how the ML approach can be modified to account for rank constraints using a regularization/penalty term.

B. Penalized Maximum Likelihood

In this section, we introduce regularized ML, constrained ML, and model order selection (MOS) that potentially can be used to address the problem of rank recovery associated with ML solution. For each framework, we start with the formulation and then proceed with the corresponding challenges. In contrast to confidence-constrained rank minimization approach which we introduce in the following section, there are no guarantees for exact rank recovery in topic models using

penalized ML. Analogous to the use of l_1-regularizer for sparsity, we consider the use of the nuclear norm to enforce the rank constraint in the matrix setting. The heuristic replacement of rank with nuclear norm has been proposed in the literature for matrix completion [20], [29], collaborative filtering [34], and multi-task learning [35].

In regularized ML, a regularized nuclear norm is added to the objective function in (6) yielding:

$$\begin{aligned} \text{minimize} \quad & \sum_{d=1}^{M} n_d D_{kl}(\hat{\Psi}_{\cdot d} \| \tilde{\Psi}_{\cdot d}) + \eta \|\tilde{\Psi}\|_*, \\ \text{subject to} \quad & \tilde{\Psi} \geq 0, \\ & 1^T \tilde{\Psi} = 1. \end{aligned} \tag{7}$$

The regularization parameter η weighs the nuclear norm. The regularized ML can be viewed as maximum a posteriori (MAP) criterion using a prior distribution over matrix $\tilde{\Psi}$ of the form $Ce^{-\eta\|\tilde{\Psi}\|}$. This is similar to the interpretation of l_1-regularization for sparse recovery as MAP with a Laplacian prior. Since one can apply the Lagrange multipliers framework to replace a constraint with a regularization term, (7) can be formulated as constrained ML. The constrained ML formula considers incorporating the nuclear norm as an additional constraint to (6):

$$\begin{aligned} \text{minimize} \quad & \sum_{d=1}^{M} n_d D_{kl}(\hat{\Psi}_{\cdot d} \| \tilde{\Psi}_{\cdot d}), \\ \text{subject to} \quad & \|\tilde{\Psi}\|_* \leq \nu, \\ & \tilde{\Psi} \geq 0, \\ & 1^T \tilde{\Psi} = 1, \end{aligned} \tag{8}$$

where $\nu \geq 0$ is a tuning parameter. For each value of η in (8) there is a value of ν in (7) which produces the same solution [36]. As an alternative to (7) and (8), MOS can be applied to rank estimation of a matrix [37], [38]. MOS offers a way to evaluate the classical trade-off between goodness of fit and model complexity. For $r = 1, 2, \ldots, \min(L, M)$, a sequence of optimization problems in the form of (6) subject to rank $= r$ is solved to obtain $\tilde{\Psi}^{*(r)}$. Then for each rank r, a cost function including negative log-likelihood at $\tilde{\Psi}^{*(r)}$ plus a penalty term pen(r) is evaluated. The penalty term corresponds to the complexity of the model and is measured based on an information criterion such as Akaike Information Criterion (AIC) or Minimal Description Length (MDL) [37], [38]. Note that in AIC the penalty term corresponds to the number of free parameters in the model. In MDL, each model candidate is assigned with a code length and minimum code length is used for model selection. In some implementations of MDL, each model is assigned with a prior probability and the model that yields the maximum posterior probability is selected. The use of rank minimization for model order selection in system identification is proposed in [39], [40]. Furthermore in [39], the authors proposed the heuristic replacement of the rank with the nuclear norm and showed that it makes the selection of an appropriate model order easier. In the following discussion, we illustrate some of the challenges associated with regularized ML, constrained ML, and MOS proposed in this section.

Discussion One of the challenges associated with the regularized and constrained ML is the choice of the regularization parameters (η and ν, respectively). Often, a criterion for selecting a value for the regularization parameters that guarantees exact rank recovery of matrix Ψ is unavailable. For the problem of low-rank matrix estimation in the noisy setting, asymptotic relationship between the regularization parameter and estimation accuracy is proposed in [41], [42]. Such results cannot be applied directly to our problem for the following reason. Counter to the sampling process in Section II-A, the sampling process proposed in [42] follows an $i.i.d.$ model without the positivity and sum-to-one. In MOS approach, solving the sequence of an optimization problem with rank constraint and evaluating the cost function for different value of rank ($r = 1, 2, \ldots, \min(L, M)$) is computationally complex. While in the unconstrained setting SVD provides a one-shot solution [37], in the constrained setting rank minimization is NP-hard [43]. The heuristic replacement of rank with nuclear norm in MOS proposed in [39], [40] suggests a regularization parameter framework. However, no recipe is provided for selecting the regularization parameter to guarantee rank recovery. In the following, we define the confidence-constrained rank minimization and show how our formulation of the problem can address the issues associated with parameter tuning in regularized ML and constrained ML and exhaustive rank search for MOS stated in this section.

C. Confidence-constrained rank minimization

We consider the concept of the confidence-constrained rank minimization for rank recovery in topic models. Using the statistical formulation of the problem proposed in Section II, an in-probability bound on the objective function in (6) can be obtained. The probability bound on data fit criterion allows us to define a confidence set. Confidence set is a high-dimensional generalization of the confidence interval and restricts the search space of the problem. Search inside the confidence set guarantees a low-rank solution. Hence, in this approach the roles of ML objective and rank constrained are replaced. We consider rank minimization subject to ML objective constraint. The confidence-constrained rank minimization is given by:

$$\begin{aligned} \text{minimize} \quad & \text{Rank}(\tilde{\Psi}) \\ \text{subject to} \quad & \sum_{d=1}^{M} n_d D_{KL}(\hat{\Psi}_{d\cdot} \| \tilde{\Psi}_{d\cdot}) \leq \epsilon(\delta), \\ & \tilde{\Psi} \geq 0, \\ & 1^T \tilde{\Psi} = 1, \end{aligned} \tag{9}$$

where $\epsilon(\delta)$ is an in-probability bound for the estimation error. Note in this formulation the tuning parameter $\epsilon(\delta)$ can be obtained by bounding $\sum_{d=1}^{M} n_d D_{KL}(\hat{\Psi}_{d\cdot} \| \tilde{\Psi}_{d\cdot})$. Intuitively the KL confidence-constrained set in (9) includes the matrix Ψ, and hence it is guaranteed (w.p. $1 - \delta$) that the rank of the solution to (9) is less than or equal to the rank of matrix Ψ. The main problem with KL divergence between two matrices is that there is no straightforward way of translating it to the distance between their singular values. Since singular values

are related to the rank of a matrix, it is hard to provide the theoretical guarantees for rank recovery in the KL version of the confidence-constrained set. While the KL confidence-constrained formulation is difficult to handle, the Frobenius-norm confidence-constrained formulation provides a convenient framework for proving rank recovery in topic models. The problem of parameter tuning is elegantly addressed in this framework by obtaining a model based in-probability uniform bound on the confidence set. Moreover, the approach does not require a scan through a range of rank values. In the following, we show that in the Frobenius-norm confidence-constrained rank minimization exact rank recovery can be guaranteed.

IV. Exact rank recovery: theoretical guarantees

In this part, we introduce Frobenius-norm confidence-constrained rank recovery and provide the theoretical guarantees for exact rank recovery in topic models. The KL-divergence confidence-constrained rank recovery in (9) is replaced with Frobenius norm confidence-constrained rank recovery since the theoretical results can be shown for the Frobenius-norm case while such results are unavailable for the KL-divergence.

A. Frobenius-norm confidence-constrained rank minimization (CRM)

For the problem defined in Section II-B, we propose the following confidence-constrained rank minimization:

$$\begin{array}{lll} \text{(CRM):} & \text{minimize} & \text{Rank}(\tilde{\Psi}) \\ & \text{subject to} & \|\tilde{\Psi}-\hat{\Psi}\|_F \le \epsilon(\delta_k), \\ & & \tilde{\Psi} \ge 0, \\ & & 1^T\tilde{\Psi} = 1. \end{array} \tag{10}$$

where

$$\epsilon(\delta_k) = \epsilon^*(\delta_k) \triangleq \sqrt{\frac{1}{n}\left(M + k\sqrt{\frac{M}{2}(1+\frac{3}{n})}\right)},$$
$$\delta_k = \frac{1}{1+k^2}, \tag{11}$$

where $n_d = n$ for all d. In Appendix B, ϵ^* is developed for the general case where document d has n_d words. Here for simplicity, we present the case where $n_d = n$. The parameter $k = \sqrt{\delta_k^{-1} - 1}$ is the number of standard deviation away from the mean, e.g., for $k = 3$, with probability $1-1/(1+k^2) = 0.9$, $\|\tilde{\Psi}-\hat{\Psi}\|_F \le \epsilon(\delta_3)$ where $\epsilon(\delta_3) = \sqrt{\frac{1}{n}(M + 3\sqrt{\frac{M}{2}(1+\frac{3}{n})})}$. Note that (10) is free of tuning parameters for the following reason. Since the samples are governed by a multinomial distribution, an in-probability bound on the estimation error of the form $\|\Psi-\hat{\Psi}\|_F \le \epsilon(\delta_k)$ w.p. $1-\delta$ can be obtained. Moreover, since the true low-rank matrix Ψ satisfies the Frobenius norm inequality constraint w.p. $1-\delta$, then Ψ_0 the solution to (10) is of equal or lower rank to that of Ψ. While this result is straightforward, the following theorem shows that in fact the CRM solution Ψ_0 has the same rank as Ψ. Moreover, theorem provides a bound on the estimation error [44].

Theorem 1: Let Ψ be a γ-distinct rank T matrix and $\hat{\Psi} \sim$ norm-multinomial$(\Psi, \mathbf{n})$. Assume $\gamma > 2\epsilon$, and $\epsilon = \epsilon^*$ defined in (11). Then, with probability at least $1-\delta_k$, Ψ_0 the solution to (10) satisfies:

1) $\Psi_0 \in 2\epsilon$-neighborhood of Ψ,
2) Rank$(\Psi_0) = T$.

Theorem 1 characterizes Ψ_0 the solution to CRM in (10). First, Ψ_0 is at most 2ϵ away from the true matrix Ψ. Theorem 1 is formulated with specific ϵ in (11) which comes from the statistical model presented in Section II. With ϵ in (11), the Frobenius norm of the estimation error $(\Psi_0 - \Psi)$ is $\mathcal{O}(\sqrt{M/n})$. The second property asserts that under the hypothesis of the Theorem 1, it is guaranteed that with probability $1-\delta$ Ψ_0 has the same rank as the rank of the true unknown matrix Ψ. In other words, the exact rank of the true matrix Ψ can be recovered by solving the CRM optimization problem in (10). We now proceed with the proof of Theorem 1. For this, first we provide a detail framework as follows:

Definition 2: Ψ' is a γ-distinct rank r matrix if $\sigma_1(\Psi') \ge \sigma_2(\Psi') \ge \ldots \ge \sigma_r(\Psi') > \gamma > \sigma_{r+1}(\Psi') = \ldots = \sigma_L(\Psi') = 0$, where σ_i is the i^{th} largest singular value of matrix Ψ'. In other words, Ψ' is γ-distinct if all of its non zero singular values are greater than γ.

Definition 3: Matrix Ψ' is in the ζ-neighborhood of matrix Ψ if $\|\Psi - \Psi'\|_F \le \zeta$.

Lemma 4: W.p. $1-\delta$ matrix Ψ satisfies $\|\Psi - \hat{\Psi}\|_F \le \epsilon$, where $\epsilon = \epsilon^*$ is given by (11).

Proof: See Appendix B. ∎

Lemma 4 guarantees that w.p. $1-\delta$ the confidence-constrained set $S(\hat{\Psi}, \epsilon^*) = \{\Psi' \mid \|\hat{\Psi} - \Psi'\|_F \le \epsilon\}$ contains the true low-rank matrix Ψ.

Lemma 5: Let Ω be γ-distinct rank r matrix. Then there exists no matrix in the γ-neighborhood of Ω, with the rank $r_0 < r$.

Proof: Suppose $\exists\Omega'$ in the γ-neighborhood with rank $r_0 < r$, therefore

$$\begin{aligned} \gamma &\ge \|\Omega' - \Omega\|_F \\ &\ge \min_{\text{Rank}(\tilde{\Omega})=r_0} \|\tilde{\Omega} - \Omega\|_F. \end{aligned} \tag{12}$$

By Eckart-Young theorem [45] the closest $\tilde{\Omega}$ with rank r_0 to Ω in the Frobenius norm is $\tilde{\Omega} = U\Sigma^* V^T$, where $\Omega = U\Sigma V^T$ and $\Sigma^* = \text{diag}(\sigma_1, \ldots, \sigma_{r_0}, 0, \ldots, 0)$. For such $\tilde{\Omega}$, $\|\tilde{\Omega} - \Omega\|_F^2 = \sum_{i=r_0+1}^{r} \sigma_i^2$. Thus, $\gamma \ge \sqrt{\sum_{i=r_0+1}^{r} \sigma_i^2} \ge \sigma_r(\Omega)$. By contradiction to the assumption that $\sigma_r(\Omega) > \gamma$, there exists no such Ω' in γ-neighborhood with rank lower than r. ∎

Based on Lemma 5, the γ-distinct property of matrix Ψ assures that all the matrices inside the γ-neighborhood of matrix Ψ have a rank greater than or equal to rank of matrix Ψ . Using Definitions 2 and 3 and Lemmas 4 and 5, we proceed with the proof of Theorem 1.

Proof: 1) Using the triangle inequality, we have

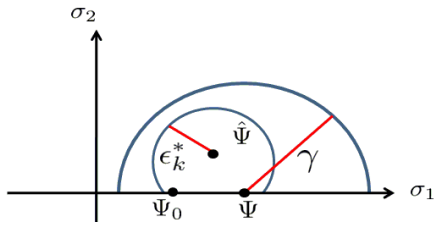
$$\|\Psi_0 - \Psi\|_F \le \|\Psi_0 - \hat{\Psi}\|_F + \|\hat{\Psi} - \Psi\|_F. \tag{13}$$

Note that the first term on the RHS of (13) is less than ϵ with probability 1, since Ψ_0 the solution to (10) satisfies the

confidence-constraint. Thus, $\Psi_0 \in \epsilon$-neighborhood of $\hat{\Psi}$. The second term on the RHS of (13) is a random quantity which can be bounded by ϵ with probability $1-\delta$ by Lemma 4. Therefore $\|\Psi_0 - \Psi\|_F \leq 2\epsilon$ with probability $1-\delta$. ■

Proof: 2) Since Ψ_0 is in the 2ϵ-neighborhood of Ψ and $2\epsilon < \gamma$, then Ψ_0 is also in the γ-neighborhood of Ψ. Hence, based on Lemma 5 $\text{Rank}(\Psi_0) \geq \text{Rank}(\Psi)$. On the other hand, since $\Psi \in \epsilon$-neighborhood of $\hat{\Psi}$ w.p. $1-\delta_k$, and Ψ_0 is the minimum rank solution matrix in ϵ-neighborhood of $\hat{\Psi}$, then $\text{Rank}(\Psi_0) \leq \text{Rank}(\Psi)$. The inequalities can hold only if $\text{Rank}(\Psi_0) = \text{Rank}(\Psi) = T$. ■

Discussion The basic idea of Theorem 1 relies on two main principles. 1) γ-distinct property of matrix Ψ which corresponds to the robustness of Ψ to the sampling noise. If γ is large, the matrix Ψ is robust enough to be rank recoverable given a small sampling noise (for illustration see Fig. 2). 2) The second principle associates with the magnitude of the sampling noise which controls the size of the confidence-constrained set. Since the statistics of the sampling noise is known, it provides the theoretical guarantees for recovering the exact rank of the matrix Ψ.



(a)

Fig. 2. This figure shows two sets: i) ϵ-neighborhood of matrix $\hat{\Psi}$ (confidence-constrained set) which is defined as $\{\Psi | \|\hat{\Psi} - \Psi\|_F \leq \epsilon\}$ and ii) γ-neighborhood of matrix Ψ which is defined as $\{\Psi' | \|\Psi - \Psi'\|_F \leq \gamma\}$. In this figure, matrix Ψ is γ distinct and $\gamma > 2\epsilon_k^*$. Thus, the assumptions of Theorem 1 hold. As a result, Ψ_0 will have the same rank as matrix Ψ.

B. Confidence-constrained nuclear norm minimization (CNM)

In general, rank minimization problems are NP hard [46]. Various algorithms have been proposed to solve the general rank minimization problem locally (e.g., see [43], [47]). A heuristic replacement of the rank minimization with a nuclear norm minimization is commonly proposed [19], [20]. The nuclear norm of a matrix is defined as $\|X\|_* = \sum_i \sigma_i$ where $\sigma_i \geq 0$ are the singular values of matrix X. The nuclear norm is a special class of Schatten norm. The Schatten norm for matrix X is defined as $\|X\|_p = (\sum_i \sigma_i^p)^{\frac{1}{p}}$. When $p = 1, \|X\|_p$ is equal to the nuclear norm, which is the sum of the singular values of matrix X. Similar to the use of l_1-regularization for sparsity, nuclear norm regularization is used to enforce low-rank in the matrix setting. To solve the rank minimization problem proposed in (10), we propose the widely used approach of replacing the rank minimization with the tractable convex optimization problem of nuclear norm minimization. In Section VI, we provide the evaluation of CNM only, due to the prohibitive computation complexity associated with CRM. In the following, confidence-constrained nuclear norm minimization (CNM) is proposed as a convex alternative to (10):

$$\begin{aligned} \text{(CNM):} \quad & \text{minimize} && \|\tilde{\Psi}\|_* \\ & \text{subject to} && \|\tilde{\Psi} - \hat{\Psi}\|_F \leq \epsilon, \\ & && \tilde{\Psi} \geq 0, \\ & && 1^T \tilde{\Psi} = 1. \end{aligned} \tag{14}$$

We denote the solution to (14) by $\tilde{\Psi}^*$. Since the nuclear norm is a convex function, and the set of the inequality and equality constraints construct a convex set, (14) is a convex optimization problem. This formulation targets the problem of exact rank recovery for probability matrices under the sampling process described in Section II-A.

V. Confidence-constrained nuclear norm minimization algorithm (CNMA)

The nuclear norm minimization problem can be reformulated as an SDP [19]. Off-the-shelf SDP solvers such as SDPT3 and SeDuMi are used to solve this problem. Such software packages use the interior point method with Newton direction which is computationally expensive [24]–[26]. The SDP problem of CNM has $(M+L) \times (M+L)$ semidefinite constraints and $(ML + M + 1)$ equality and inequality constraints. The computational complexity is $\mathcal{O}(\min\{M, L\})^6$ and the memory requirement is $\mathcal{O}(\min\{M, L\})^4$. So while the reformulation is theoretically appealing, computational challenges remain. In the following, we provide an accelerated projection gradient algorithm to solve the dual formulation of CNM. We start with the dual formulation of CNM and then solve it with the gradient projection approach [48]. We propose an accelerated version of our algorithm using two point approximation [49] and a highly economical SVD-based implementation.

A. Dual formulation background

We solve (14) through formulating the dual problem. Generally, the dual formulation of a problem in the form of

$$\begin{aligned} & \text{minimize} && f_0(x) \\ & \text{Subject to} && f_1(x) \leq 0 \\ & && h(x) = 0, \end{aligned}$$

can be obtained first by constructing the Lagrangian $\mathcal{L}(x, \lambda_1, \underline{\lambda}_2)$ as follows:

$$\mathcal{L}(x, \lambda_1, \lambda_2) = f_0(x) + \lambda_1^T f_1(x) + \lambda_2^T h(x),$$

where $\lambda_1 \geq 0$ and λ_2 are the Lagrange multipliers for the set of inequality and equality constraints, respectively. The Lagrangian incorporates the constraints into the objective function using the Lagrange multipliers λ_1, and λ_2. The second step is to minimize the Lagrangian $\mathcal{L}(x, \lambda_1, \lambda_2)$ with respect to the primal objective variable x. Define $x^*(\lambda_1, \lambda_2)$ as:

$$x^*(\lambda_1, \lambda_2) = \arg\min_x \mathcal{L}(x, \lambda_1, \lambda_2).$$

By replacing $x^*(\lambda_1, \lambda_2)$ in the Lagrangian, we obtain the dual:

$$g(\lambda_1, \lambda_2) = \mathcal{L}(x^*(\lambda_1, \lambda_2), \lambda_1, \lambda_2).$$

The dual formulation is given by the following optimization

$$\begin{array}{ll} \text{maximize} & g(\lambda_1, \lambda_2) \\ \text{Subject to} & \lambda_1 \geq 0. \end{array}$$

The dual formulation of the optimization problem has several advantages. First, it provides a lower bound for the primal problem. One can show for any feasible point $\tilde{x}$ in the primal problem, $g(\lambda_1, \lambda_2) \leq f(\tilde{x})$. If the primal problem is convex and the set of inequalities is strictly satisfied for some point inside the feasibility set, then based on Slater's condition the strong duality holds [50]. Hence, the duality gap $f(\tilde{x}) - g(\lambda_1, \lambda_2)$ provides means of assessing convergence of the optimization algorithm. Furthermore, the positivity constraint in the dual formulation can be handled using a simple projection onto the positive orthant. Note that in the primal formulation the projection onto the set of equality and inequality constraints could be more complex.

B. Dual formulation of CNM

We follow the steps explained in Section V-A. First, we construct the Lagrangian of (14) to obtain the dual formulation [51]. The Lagrangian $\mathcal{L}(\tilde{\Psi}, \lambda_1, \underline{\lambda}_2, \Lambda_3)$ for problem in (14) can be written as

$$\mathcal{L}(\tilde{\Psi}, \lambda_1, \underline{\lambda}_2, \Lambda_3) = \|\tilde{\Psi}\|_* + \frac{\lambda_1}{2}(\|\tilde{\Psi} - \hat{\Psi}\|_F^2 - \epsilon^2) + \underline{\lambda}_2^T(1 - \tilde{\Psi}^T 1) - \text{tr}(\Lambda_3^T \tilde{\Psi}), \quad (15)$$

where $\lambda_1 \in \mathbb{R}^+$, $\underline{\lambda}_2 \in \mathbb{R}^{M\times 1}$, and $\Lambda_3 \in \mathbb{R}^{+L\times M}$. If we minimize $\mathcal{L}(\tilde{\Psi}, \lambda_1, \underline{\lambda}_2, \Lambda_3)$ with respect to $\tilde{\Psi}$, we obtain $\tilde{\Psi}^*(\lambda_1, \underline{\lambda}_2, \Lambda_3)$. We start by rewriting (15) as follows:

$$\mathcal{L}(\tilde{\Psi}, \lambda_1, \underline{\lambda}_2, \Lambda_3) = \|\tilde{\Psi}\|_* + \frac{\lambda_1}{2}\|\tilde{\Psi} - \Psi'\|_F^2 + C(\lambda_1, \underline{\lambda}_2, \Lambda_3), \quad (16)$$

where $\Psi' = \hat{\Psi} + \frac{1\underline{\lambda}_2^T}{\lambda_1} + \frac{\Lambda_3}{\lambda_1}$, and $C(\lambda_1, \underline{\lambda}_2, \Lambda_3) = -\frac{\lambda_1}{2}\|\Psi'\|_F^2 + \frac{\lambda_1}{2}\|\hat{\Psi}\|_F^2 + \underline{\lambda}_2^T 1 - \frac{\lambda_1}{2}\epsilon^2$. The solution to the minimization of (16) w.r.t. $\tilde{\Psi}$ is

$$\tilde{\Psi}^*(\lambda_1, \underline{\lambda}_2, \Lambda_3) = D_{\frac{1}{\lambda_1}}(\Psi'),$$

where $D_\tau(X)$ is the soft thresholding operator on the singular value of matrix X (for proof see [24]) defined by $D_\tau(X) = U(S - \tau I)_+ V^T$, where $X = USV^T$ is the SVD of X. To obtain the dual, we substitute $\tilde{\Psi}^*(\lambda_1, \underline{\lambda}_2, \Lambda_3)$ back into (16), simplify and obtain

$$f(\lambda_1, \underline{\lambda}_2, \Lambda_3) = -\frac{\lambda_1}{2}\|D_{\frac{1}{\lambda_1}}(\Psi')\|_F^2 + \frac{\lambda_1}{2}\|\hat{\Psi}\|_F^2 + \underline{\lambda}_2^T 1 - \frac{\lambda_1}{2}\epsilon^2.$$

Thus the dual formulation of the CNM problem in (14) is

$$\begin{array}{ll} \text{maximize} & f(\lambda_1, \underline{\lambda}_2, \Lambda_3) \\ \text{subject to} & \lambda_1 \geq 0 \\ & \Lambda_3 \geq 0, \end{array}$$

where $\lambda_1 \in \mathbb{R}$, $\underline{\lambda}_2 \in \mathbb{R}^{M\times 1}$, and $\Lambda_3 \in \mathbb{R}^{L\times M}$. Note that the positivity for matrix Λ_3 is elementwise. Rather than maximizing the concave dual function, we proceed with the convex minimization of the negative dual, $\tilde{f}(\lambda_1, \underline{\lambda}_2, \Lambda_3) = -f(\lambda_1, \underline{\lambda}_2, \Lambda_3)$.

C. Gradient projection algorithm for CNM

The CNM optimization problem is expressed as follows:

$$\begin{array}{ll} \text{minimize} & \tilde{f}(\lambda_1, \underline{\lambda}_2, \Lambda_3) \\ \text{subject to} & \lambda_1 \geq 0 \\ & \Lambda_3 \geq 0, \end{array} \quad (17)$$

where $\tilde{f}(\lambda_1, \underline{\lambda}_2, \Lambda_3) = \frac{\lambda_1}{2}\|D_{\frac{1}{\lambda_1}}(\hat{\Psi} + \frac{1\underline{\lambda}_2^T}{\lambda_1} + \frac{\Lambda_3}{\lambda_1})\|_F^2 - \frac{\lambda_1}{2}\|\hat{\Psi}\|_F^2 - \underline{\lambda}_2^T 1 + \frac{\lambda_1}{2}\epsilon^2$. We consider the gradient projection method to solve (17). The gradient projection method for minimizing a continuous convex function over a closed convex set was proposed in [52]. The modified backtracking approach for the gradient projection method was defined in [48]. Application of the gradient projection method to our problem consists of the following iterations:

$$\begin{aligned} \lambda_1^{k+1} &= [\lambda_1^k - t^k \nabla \tilde{f}_{\lambda_1^k}(\lambda_1, \underline{\lambda}_2, \Lambda_3)]_+ \\ \underline{\lambda}_2^{k+1} &= \underline{\lambda}_2^k - t^k \nabla \tilde{f}_{\underline{\lambda}_2^k}(\lambda_1, \underline{\lambda}_2, \Lambda_3) \\ \Lambda_3^{k+1} &= [\Lambda_3^k - t^k \nabla \tilde{f}_{\Lambda_3^k}(\lambda_1, \underline{\lambda}_2, \Lambda_3)]_+, \end{aligned}$$

where $[x]_+ = x$ for $x \geq 0$, and otherwise is zero, $\nabla \tilde{f}_{\lambda_i}(\lambda_1, \lambda_2, \Lambda_3)$ is the gradient with respect to λ_1, $\underline{\lambda}_2$, Λ_3, and t^k is the step size. Note that since the positivity of λ_1 and Λ_3 can be enforced coordinatewise, the projection is trivial. The gradient of $\tilde{f}(\underline{\lambda})$ with respect to $\lambda_1, \underline{\lambda}_2$, and Λ_3 is respectively,

$$\begin{aligned} &\nabla \tilde{f}_{\lambda_1}(\lambda_1, \underline{\lambda}_2, \Lambda_3) = \frac{1}{2}\|D_{\frac{1}{\lambda_1}}(\Psi')\|_F^2 + \frac{1}{\lambda_1}\|D_{\frac{1}{\lambda_1}}(\Psi')\|_* \\ &-\frac{1}{\lambda_1}\text{tr}((1\underline{\lambda}_2^T + \Lambda_3)^T D_{\frac{1}{\lambda_1}}(\Psi')) - \frac{1}{2}\|\hat{\Psi}\|_F^2 + \frac{\epsilon^2}{2}, \\ &\nabla \tilde{f}_{\underline{\lambda}_2}(\lambda_1, \underline{\lambda}_2, \Lambda_3) = D_{\frac{1}{\lambda_1}}(\Psi')^T 1 - 1, \\ &\nabla \tilde{f}_{\Lambda_3}(\lambda_1, \underline{\lambda}_2, \Lambda_3) = D_{\frac{1}{\lambda_1}}(\Psi'). \end{aligned}$$

The derivative of $\tilde{f}$ with respect to λ_1 is given by $\frac{d}{d\lambda_1}(\frac{\lambda_1}{2}\|D_{\frac{1}{\lambda_1}}(\Psi')\|_F^2) - \frac{1}{2}\|\hat{\Psi}\|_F^2 + \frac{\epsilon^2}{2}$. The derivation of the term $\frac{d}{d\lambda_1}(\frac{\lambda_1}{2}\|D_{\frac{1}{\lambda_1}}(\Psi')\|_F^2)$ which leads to the explicit expression of $\nabla \tilde{f}_{\lambda_1}(\lambda_1, \underline{\lambda}_2, \Lambda_3)$ is provided in Appendix A. Upon convergence of the Lagrange multipliers $[\lambda_1, \underline{\lambda}_2, \Lambda_3]$, one can compute the primal objective parameters using $\tilde{\Psi} = D_{\frac{1}{\lambda_1}}(\hat{\Psi} + \frac{1\underline{\lambda}_2^T}{\lambda_1} + \frac{\Lambda_3}{\lambda_1})$. In the following, we first show how to choose the step size for the gradient method using the backtracking approach. Then, we provide the accelerated gradient projection method.

1) Step size: To choose the step size t^k, we use the backtracking approach for gradient projection [48]. The backtracking line search for gradient projection requires the smallest nonnegative integer m_k such that

$$\tilde{f}\Big(\lambda_1^k(t^k), \underline{\lambda}_2^k(t^k), \Lambda_3^k(t^k)\Big) \leq \tilde{f}(\lambda_1^k, \underline{\lambda}_2^k, \Lambda_3^k) - \gamma\Big(\nabla \tilde{f}_{\lambda_1} \Delta\lambda_1^k + \nabla \tilde{f}_{\underline{\lambda}_2}^T \Delta\underline{\lambda}_2^k + \text{tr}(\nabla \tilde{f}_{\Lambda_3}^T \Delta\Lambda_3^k)\Big),$$

where $\Delta\lambda_1^k = \lambda_1^k - \lambda_1^k(t^k)$, $\Delta\underline{\lambda}_2^k = \underline{\lambda}_2^k - \underline{\lambda}_2^k(t^k)$, $\Delta\Lambda_3^k = \Lambda_3^k - \Lambda_3^k(t^k)$, $t^k = \eta^{m_k} t^0$, $\gamma \in (0, 0.5)$, $t^0 > 0$, and $\eta \in (0, 1)$. The proposed backtracking approach in (18) finds a step size t^k which reduces the objective function sufficiently. However to avoid making a small step in each iteration, we start with a large enough step size t^0 which satisfies the following condition:

$$\tilde{f}\Big(\lambda_1^k(t^0), \underline{\lambda}_2^k(t^0), \Lambda_3^k(t^0)\Big) > \tilde{f}(\lambda_1^k, \underline{\lambda}_2^k, \Lambda_3^k) - \gamma\Big(\nabla\tilde{f}_{\lambda_1}\Delta\lambda_1^k + \nabla\tilde{f}_{\underline{\lambda}_2}^T\Delta\underline{\lambda}_2^k + \mathrm{tr}(\nabla\tilde{f}_{\Lambda_3}^T\Delta\Lambda_3^k)\Big).$$

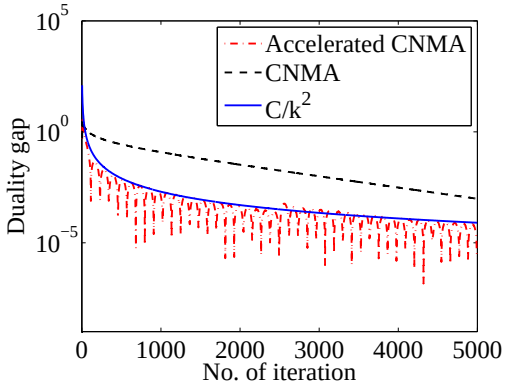
Algorithm 2 Accelerated CNMA for exact rank recovery

Choose $\lambda_1^0 = \lambda_1^1 > 0, \underline{\lambda}_2^0 = \underline{\lambda}_2^1 = 0, \Lambda_3^0 = \Lambda_3^1 = 0$, $a_0 = a_1 = 1$, $\eta \in (0, 1)$, $\gamma \in (0, 0.5)$, $\mu > 1$, $t_0 > 0$, K, υ
for $k = 1$ to K **do**
  $\bar{\lambda}_1^k = \lambda_1^k + \frac{a_{k-1}-1}{a_k}(\lambda_1^k - \lambda_1^{k-1})$, $\bar{\underline{\lambda}}_2^k = \underline{\lambda}_2^k + \frac{a_{k-1}-1}{a_k}(\underline{\lambda}_2^k - \underline{\lambda}_2^{k-1})$, $\bar{\Lambda}_3^k = \Lambda_3^k + \frac{a_{k-1}-1}{a_k}(\Lambda_3^k - \Lambda_3^{k-1})$ {Acceleration}
  $\Psi'^k = \hat{\Psi} + \frac{1\bar{\underline{\lambda}}_2^{kT}}{\bar{\lambda}_1^k} + \frac{\bar{\Lambda}_3^k}{\bar{\lambda}_1^k}$
  $(U, S, V^T) = \mathrm{svd}(\Psi'^k)$
  $\tilde{\Psi}^{k+1} = U(S - 1/\bar{\lambda}_1^k)_+ V^T$ {Soft thresholding}
  while $\tilde{f}\Big(\lambda_1^k(t^0), \underline{\lambda}_2^k(t^0), \Lambda_3^k(t^0)\Big) \leq \tilde{f}(\bar{\lambda}_1^k, \bar{\underline{\lambda}}_2^k, \bar{\Lambda}_3^k) - \gamma\Big(\nabla\tilde{f}_{\bar{\lambda}_1}\Delta\bar{\lambda}_1^k + \nabla\tilde{f}_{\bar{\underline{\lambda}}_2}^T\Delta\bar{\underline{\lambda}}_2^k + \mathrm{tr}(\nabla\tilde{f}_{\bar{\Lambda}_3}^T\Delta\bar{\Lambda}_3^k)\Big)$ **do**
    $t^0 = \mu^{n_k} t_0$ {line search (wolf condition)}
  end while
  while $\tilde{f}\Big(\lambda_1^k(t^k), \underline{\lambda}_2^k(t^k), \Lambda_3^k(t^k)\Big) > \tilde{f}(\bar{\lambda}_1^k, \bar{\underline{\lambda}}_2^k, \bar{\Lambda}_3^k) - \gamma\Big(\nabla\tilde{f}_{\bar{\lambda}_1}\Delta\bar{\lambda}_1^k + \nabla\tilde{f}_{\bar{\underline{\lambda}}_2}^T\Delta\bar{\underline{\lambda}}_2^k + \mathrm{tr}(\nabla\tilde{f}_{\bar{\Lambda}_3}^T\Delta\bar{\Lambda}_3^k)\Big)$ **do**
    $t^k = \eta^{m_k} t^0$ {line search (backtracking condition)}
  end while
  $\lambda_1^{k+1} = [\bar{\lambda}_1^k - t^k\nabla\tilde{f}(\bar{\lambda}_1)]_+$, $\underline{\lambda}_2^{k+1} = \bar{\underline{\lambda}}_2^k - t^k\nabla\tilde{f}(\bar{\underline{\lambda}}_2)$, $\Lambda_3^{k+1} = [\bar{\Lambda}_3^k - t^k\nabla\tilde{f}(\bar{\Lambda}_3)]_+$
  $a_{k+1} = (1 + \sqrt{4a_k^2 + 1})/2$, and $t^0 = t^k$. {updating the dual variables}
  if Duality-Gap $\leq \upsilon$ **then**
    break
  end if
end for

2) Acceleration: The general convergence rate for gradient approach is $\mathcal{O}(\frac{1}{k})$, where k is the iteration number. In [49], it is proved that the extrapolation step makes the convergence faster as much as $\mathcal{O}(\frac{1}{k^2})$. We define the extrapolated solution $\bar{\lambda}^k$ as follows:

$$\bar{\lambda}_1^k = \lambda_1^k + \frac{a_{k-1}-1}{a_k}(\lambda_1^k - \lambda_1^{k-1}),$$
$$\bar{\underline{\lambda}}_2^k = \underline{\lambda}_2^k + \frac{a_{k-1}-1}{a_k}(\underline{\lambda}_2^k - \underline{\lambda}_2^{k-1}),$$
$$\bar{\Lambda}_3^k = \Lambda_3^k + \frac{a_{k-1}-1}{a_k}(\Lambda_3^k - \Lambda_3^{k-1})$$

where $a_k = \frac{1+\sqrt{4a_{k-1}^2+1}}{2}$. For the pseudo code for the proposed CNMA see Algorithm 2. To illustrate that the proposed acceleration improves the convergence from $\mathcal{O}(1/k)$ to $\mathcal{O}(1/k^2)$, we present a plot of the duality gap vs. the number of iterations for the original CNMA and accelerated CNMA in Fig. 3. The evaluation of the SVD in each iteration is



(a)

Fig. 3. Comparison of duality gap for $M = 50$, $L = 80$, $T = 10$, $n = 1000$, $\alpha = 0.1$, and $\beta = 0.01$ for CNMA vs. accelerated CNMA

expensive and is $\mathcal{O}(\min\{M, L\}^3)$. As in [24]–[26], we use the PROPACK package to compute a partial SVD. Because PROPACK can not automatically calculate the singular values which are greater than specific value τ, we use the following procedure. To facilitate the computation of singular value 5 at a time, we set $b_0 = 5$ and update b_{l+1} for $l = 0, 1, \ldots$ as follows:

$$b_{l+1} = \begin{cases} \mathrm{Rank}(\tilde{\Psi}^{k+1}) & \text{if } \mathrm{Rank}(\tilde{\Psi}^{k+1}) < b_k \\ \mathrm{Rank}(\tilde{\Psi}^{k+1}) + 5 & \text{if } \mathrm{Rank}(\tilde{\Psi}^{k+1}) \geq b_k. \end{cases}$$

This procedure stops when $b_{l+1} = b_l$. Partial SVD calculation reduces the cost of the computation significantly, especially in the low-rank setting. The pseudo code for calculating SVD is in Algorithm 3.

Algorithm 3 SVD calculation using PROPACK

Choose $r_0 = 0$, and $i = 5$
in step l
$b_l = r_{k-1} + 1$
repeat
  $[USV]_{b_l} = \mathrm{SVD}(\Psi'^k)$
  $b_l = b_l + i$
until $s_{b_l - i}^k \leq \frac{1}{\lambda_1^k}$
$r_k = \max\{j : s_j^k > \frac{1}{\lambda_1^k}\}$
$\tilde{\Psi}^{k+1} = \sum_{j=1}^{r_k}(s_j^k - \frac{1}{\lambda_1^k})u_j^k v_j^k$

VI. Experimental results

We evaluate both theoretical and computational aspects of the confidence-constrained rank minimization problem. For the theoretical part, we provide the followings: 1) Sensitivity analysis of rank recovery accuracy as a function of ϵ, and 2) Phase diagram analysis applied to a synthetic dataset to show that the exact rank recovery obtained by CNMA is consistent with the sufficient conditions proposed by Theorem 1. For the

computational part, we provide a runtime comparison between CNMA and HDP and show the applicability of CNM for large datasets. For HDP, we use an efficient implementation of the algorithm in Matlab [1] provided by the authors of [9]. Note that in all of our experiments, we fixed the confidence value $1-\delta_k = 0.9$ and consequently set $k = 3$.

A. Sensitivity with respect to ϵ

We would like to illustrate the effect of ϵ on rank recovery. Theorem 1 suggests that by selecting $\epsilon = \epsilon^*$ (11), rank minimization guarantees exact rank recovery with probability $1-\delta$. To examine the effect of varying ϵ on rank recovery accuracy, we consider the following setup. We consider a range of values for $\epsilon = [\epsilon^*/16, \epsilon^*/8, \epsilon^*/4, \epsilon^*/2, \epsilon^*, 2\epsilon^*, 4\epsilon^*, 8\epsilon^*, 16\epsilon^*]$. The value of ϵ^* based on (11) is equal to 0.2550. We generate matrix Ψ with $M = 50$, $L = 50$, $T = 10$, $\alpha = 0.1$, and $\beta = 0.01$ following the model in Section II-A and sample $\hat{\Psi}$ 10 times. For each value of ϵ, we solve CNM in (14) for each of the ten realization of $\hat{\Psi}$ using CVX and CNMA and evaluate the rank of the recovered matrix $\tilde{\Psi}^*$. The rank evaluation is done by counting the number of singular value of matrix $\tilde{\Psi}^*$ exceeding a threshold to avoid miscounting due to numerical errors. The threshold is defined based on the empirical distribution of the smallest nonzero singular values of the true matrix Ψ (i.e., mean minus three times the standard deviation). We compute mean (μ) and standard deviation (σ) of the recovered rank for matrix $\tilde{\Psi}$ and plot the error bar ([mean-std, mean+std]) for both CVX and CNMA. Rank estimates as a function of ϵ for CVX and for CNMA are shown in Figures 4(a) and 4(b), respectively. Figures 4(a) and 4(b) support Theorem 1 by indicating that the choice of $\epsilon = \epsilon^*$ (11) leads to exact rank recovery, since for only $\epsilon = \epsilon^*$ the exact rank is recovered for 10 out of 10 leading to $\mu = 10$ and $\sigma = 0$. In other words, as we deviate from ϵ^* the true rank of matrix Ψ can no longer be recovered. We provide the following explanation. When we increase ϵ, the confidence-constrained set may include low-rank matrices which are not in the γ-neighborhood of matrix Ψ. Hence, rank minimization inside the confidence-constrained set may lead to a recovery of a low-rank matrix. On the other hand, as we decrease ϵ the confidence-constrained set may not include the true matrix Ψ. Therefore, the rank of the recovered matrix $\tilde{\Psi}$ may be higher than the rank of matrix Ψ. By comparing Figures 4(a) and 4(b), we can see that the performance of CNMA is comparable to that of CVX. To assess the effect of the number of CNMA iterations on accuracy, we terminate the algorithm after 200, 500, and 1000 iterations and present the rank recovery results in Figures 4(b). Comparing the graphs in Fig. 4(b), we observe that with an increased number of iterations the results approach that of CVX. Moreover, CNMA with a smaller number of iterations correctly recovers the rank at $\epsilon = \epsilon^*$. This hints at the potential reduction in computational complexity that CNMA can provide by reducing the number of iterations. For the relaxed CNMA graph in Fig. 4(b), we removed the positivity and sum to one constraints to assess the importance of the probability matrix constraints. We observe an increase in variation from the true rank at $\epsilon = \epsilon^*$ (11). This suggests that including the probability constraints can improve the rank recovery accuracy.

[1] http://www.gatsby.ucl.ac.uk/ ywteh/research/software.html

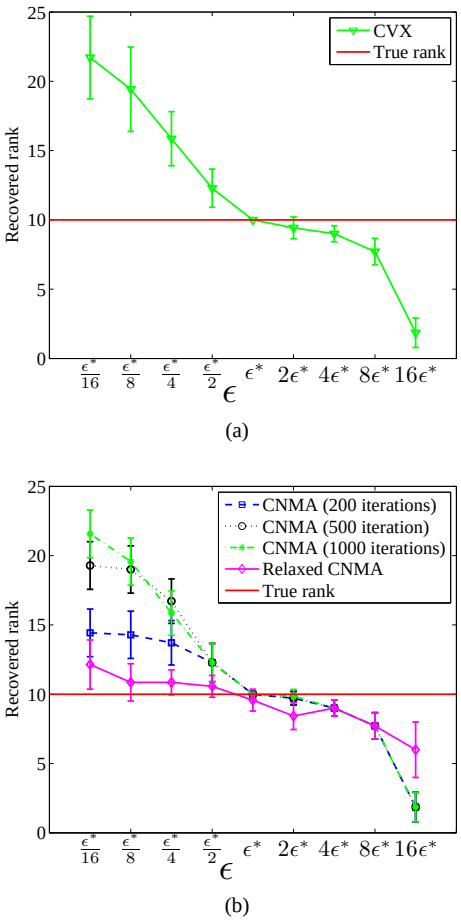
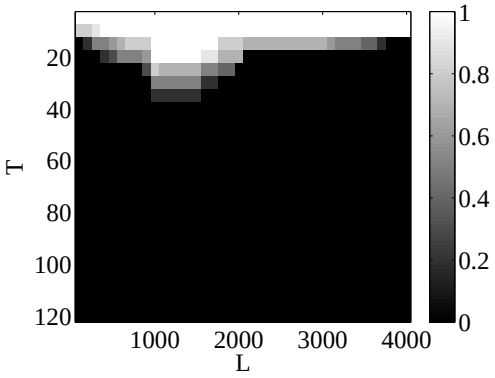


Fig. 4. This figure shows the sensitivity of rank recovery to the value of ϵ. We scan through a range of values of ϵ and plot the mean of the recovered rank including the confidence intervals for (a) CVX and (b) CNMA.

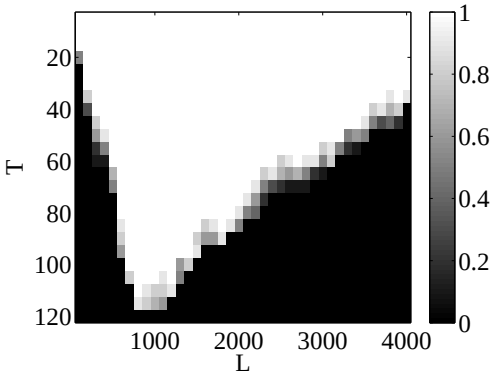
B. Phase diagram analysis

We use the notion of phase diagram as proposed in [53] to evaluate probability of exact rank recovery using CNMA for a wide range of matrices of different dimensions (i.e., vocabulary size terms $\times$ number of documents) and different number of topics and compare it with the sufficient conditions proposed by Theorem 1. We would like to show that the condition proposed in Theorem 1 for rank recovery is still valid when rank minimization is replaced with nuclear norm minimization. We generate $N = 50$ $i.i.d$ realizations of Ψ using the sampling process in Section II-A with $M = 500$, $n = 1000$, $\alpha = 0.01$, $\beta = 0.001$, over a grid of (L, T), with L ranging through 40 equispaced points in the interval $[100, 4000]$, and T ranging through 24 equispaced points in the interval $[5, 120]$. In Fig. 5(a), each pixel intensity corresponds to the empirical estimate of $P(\sigma_T > 2\epsilon)$, i.e., $\sum_{i=1}^{N} I(\sigma_T^{(i)} > 2\epsilon)/N$, where σ_T is the smallest non-zero singular value. To evaluate correct rank recovery probability, for each pixel in phase diagram we produce 20 realization of the pair $(\Psi, \hat{\Psi})$. We run CNMA for

each of the 20 realizations of $\hat{\Psi}$ and compared the rank of the recovered matrix $\tilde{\Psi}^*$ with the true rank of matrix Ψ. The rank of matrix $\tilde{\Psi}^*$ is computed following the procedure described in Section VI-A. In Fig. 5(a), the white area corresponds to success region[2] (the region where the rank recovery is guaranteed with high probability based on Theorem 1). In



(a)



(b)

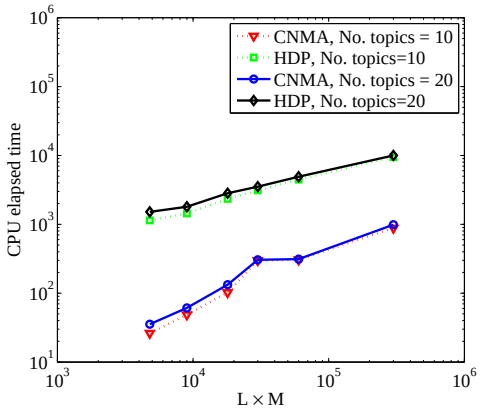
Fig. 5. (a) $P(\sigma_T > 2\epsilon)$ for $M = 1000$, $n = 1000$, $\alpha = 0.01$, and $\beta = 0.001$ (b) $\hat{P}$ (exact rank recovery) obtained by CNMA.

Fig. 5(b), the white area corresponds to exact rank recovery obtained by CNMA. Since the area for exact rank recovery probability obtained by CNMA covers the success region, the sufficient condition proposed by Theorem 1 appear to hold for the heuristic replacement of nuclear norm minimization. Comparing Figures 5(a), and 5(b) suggests that the sufficient condition for exact rank recovery proposed in Theorem 1 can be further improved. This could be attributed to the fact that the proposed sufficient conditions for exact rank recovery involve several bounds.

The LDA model in Section II depends on two hyperparameters α and β. When α is small the effective number of topics per document is small. Similarly, when β is small the effective number of words per topic is small. Intuitively, with small α and β the model is simpler (i.e., fewer topics and fewer words per topic). We are interested in evaluating the impact of α and β on the rank recovery rate. In Fig. 6, the left hand column shows the phase diagram for exact rank recovery obtained by CNMA for different values of α, and β. As we decrease the value of hyperparameters, the wider area for exact rank recovery can be covered by CNMA in phase diagram. The middle and left hand side graphs show the singular value scree plot of matrix $\hat{\Psi}$ for the point indicated by darker and lighter pointer on the phase diagram, respectively. The scree plots illustrate the fact that as we decrease α and β, Ψ becomes more distinct, i.e., the gap between the smallest non zero singular value and the following one is more distinguished. Hence, its rank is easier to recover. Moreover, by comparing the scree plots in the middle and left hand columns, it is clear that when the exact rank cannot be recovered by CNMA, the gap in the singular values of matrix $\hat{\Psi}$ cannot be found easily. We would like to emphasize that although the scree plot can be use to study the rank of a matrix, it does not provide a complete solution to the problem, i.e., it fails to suggest an admissible estimate for Ψ. Without probability constraints, an SVD can be use to obtain a low-rank estimate for Ψ. However, in the presence of probability constraint the problem is NP-hard [43].

C. Computational complexity comparison

We compare the CPU runtime of CNMA with HDP. We consider $(M, L) = [(80, 60)\ (100, 90)\ (150, 120)\ (200, 150)\ (300, 200)\ (600, 500)]$. We compute the CPU runtime using a MATLAB built in function $\{\texttt{cputime}\}$. CNMA and HDP algorithm run on a standard desktop computer with 2.5 GHz CPU (dual core) and 4 GB of memory. Figure 7(a) shows the CPU runtime comparison for CNMA vs. HDP. In Fig. 7(a), the x-axis shows the dimension of the matrix $L \times M$ and the y-axis shows the elapsed CPU time in seconds. Figure 7(a) shows that the runtime of HDP is longer than that of CNMA by at least an order of magnitude. Note that we compared the runtime of CVX (using SDPT3 as an SDP solver) with that of CNMA and observed that the runtime of CVX is longer than that of CNMA by over two orders of magnitude. This suggests that CNMA, i.e., our proposed algorithmic implementing of CNM, provides a fast and feasible solution to practical size problems and diminishes the computational limitations associated with generic solvers.



(a)

Fig. 7. Runtime comparison between CNMA and HDP.

[2]This notation is used in [53]

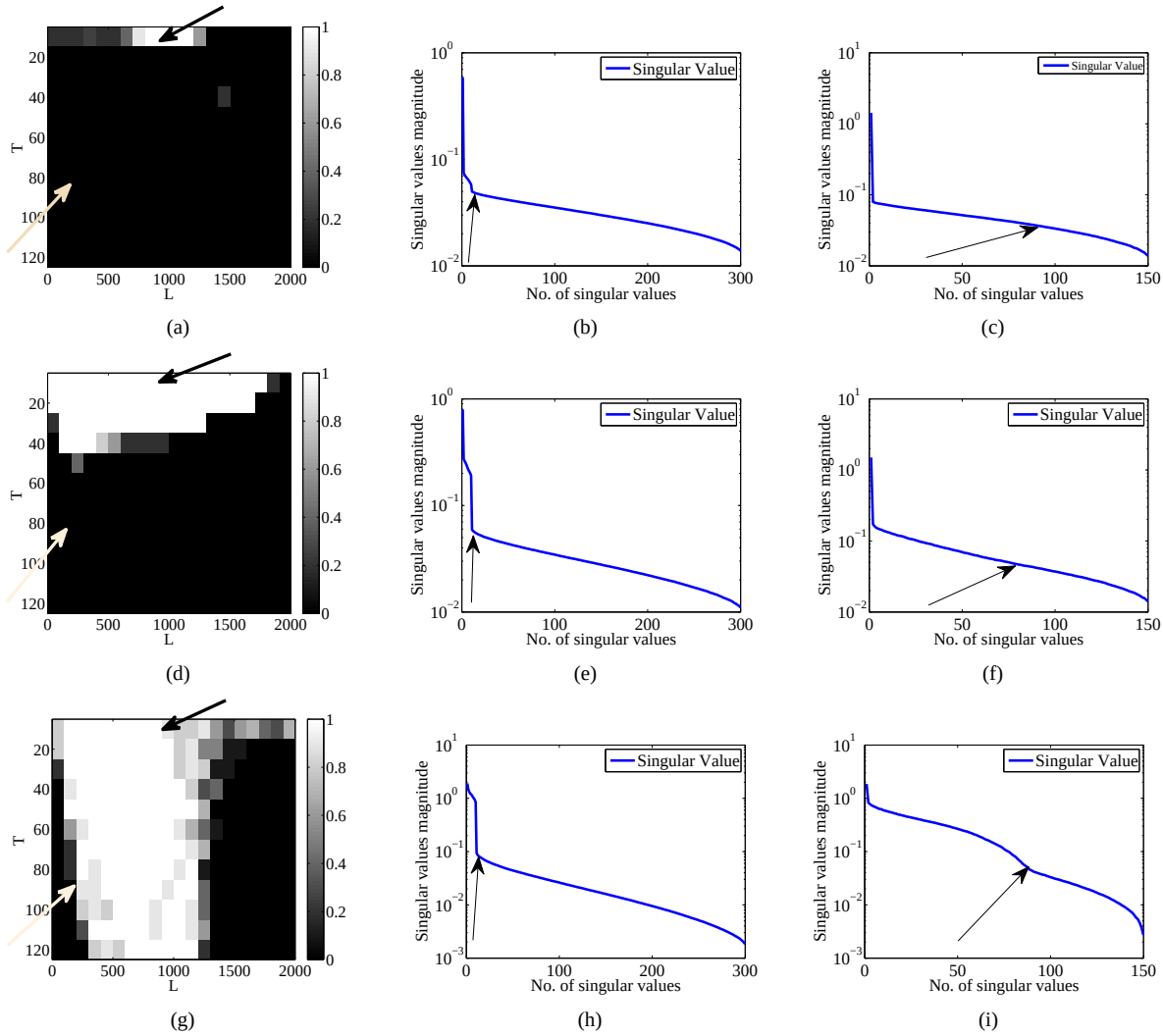


Fig. 6. This figure shows the effect of the value of the hyperparameters α and β on rank recovery rate. The first column is the phase diagram of $P(\sigma_T > 2\epsilon)$ as a function of the number of topics and the vocabulary size. Each row corresponds to a different setup of the hyperparameters α and β. (a) $\alpha = 1$, $\beta = 1$ (d) $\alpha = 0.5$, and $\beta = 0.1$ (g) $\alpha = 0.1$, and $\beta = 0.01$. The second column is the plot of the singular values for the setting indicated by black arrows. The last column is the plot of the singular values indicated by white arrows. Note that the black arrow in the phase diagram corresponds to the success region proposed by Theorem 1 and the white arrow corresponds to the fail region.

VII. Applications

As the previous section suggests, the proposed computationally-efficient algorithmic implementation of CNM can be used to solve problem of realistic dimensions. In this section, we would like to illustrate that the low-rank solution obtained by CNMA provides competitive results to that of LDA, HDP, and the optimal low-rank SVD approximation of matrix $\hat{\Psi}$ in terms of classification accuracy on two real image datasets and three real text datasets.

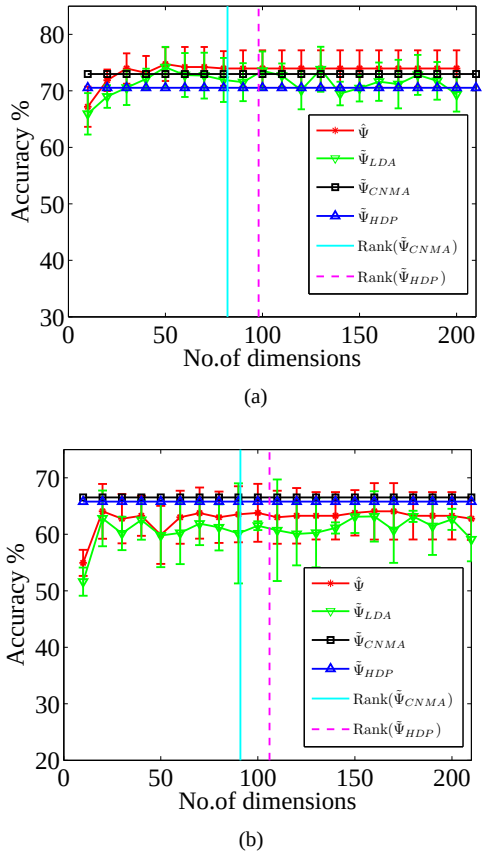
A. Image datasets

We consider two image datasets MSRCv2[3], and Corel1000[4]. MSRCv2 image dataset contains 591 images in 23 object classes. We perform a multiclass classification for MSRCv2 using the 8 row classes: *'book'*, *'grass, cow'*, *'tree, grass, sky'*, *'bike, building'*, *'sign'*, *'water, boat'*, *'aeroplane, grass, sky'*, *'road, building'* resulting in a dataset with 240 images in 8 different classes. Corel1000 image dataset contains 1000 images in 10 different classes each includes 100 images. We consider 7 classes: *'buildings'*, *'buses'*, *'flowers'*, *'elephants'*, *'horses'*, *'food' and 'mountains'* in our simulation. Note that we excluded the classes which contained images with different format of RGB representations. We randomly sampled 50 images in each class resulting in 350 images in 7 classes.

To obtain matrix $\hat{\Psi}$, we take the approach of representing each image as a collection of blocks and mapping each block to a discrete index associated with the closest dictionary template. We separate each image to several $10 \times 10 \times 3$ blocks. To construct the dictionary, we run k-means on the collection of blocks from all images to obtain L cluster centroids. The L centroids are used as the dictionary templates and each block is mapped to the index of the closest dictionary template. We run CNMA, LDA, and HDP to obtain matrix

[3]http://research.microsoft.com/en-us/projects/objectclassrecognition/default.htm
[4]http://wang.ist.psu.edu/docs/related/

$\tilde{\Psi}^*_{CNMA}$, $\tilde{\Psi}^*_{LDA}$, and $\tilde{\Psi}^*_{HDP}$, respectively. To find the optimal low-rank approximation of $\hat{\Psi}$, we project the columns of $\hat{\Psi}$ into its top d-largest left singular vectors where d scans through the dimension of matrix $\hat{\Psi}$. We use multi class SVM with Gaussian kernel for classification [54]. Parameters C and γ of SVM model are learned by k-fold cross validation where $k = 5$.

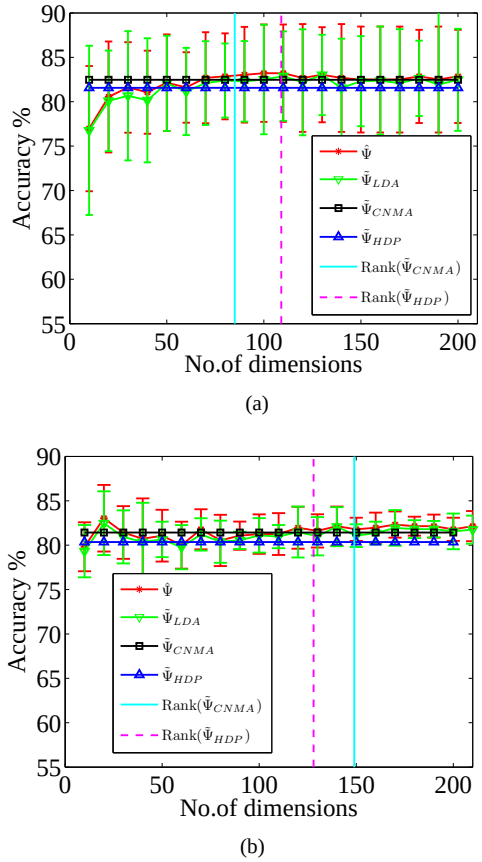


(a) (b)

Fig. 8. Multiclass classification accuracy for MSRCv2 dataset with number of clusters (a) 200 (b) 500.

In Figures 8 and 9, the classification accuracies obtained by running SVM on $\tilde{\Psi}^*_{CNMA}$, $\tilde{\Psi}^*_{LDA}$, and $\tilde{\Psi}^*_{HDP}$ as well as on different low-rank SVD-based approximations of matrix $\hat{\Psi}$ are shown. The classification accuracy provided by matrix $\tilde{\Psi}^*_{CNMA}$ is competitive with that of the others. Since CNMA and HDP determine the number of topics in an automated fashion, the accuracy for each was computed without the need to scan through the different number of topics. The number of dimensions is only relevant for the LDA and SVD approaches, in which the number of topics is an additional input to the algorithm. In both Figures 8 and 9, the vertical line shows the rank of the recovered matrix $\tilde{\Psi}^*$. We observe that the classification accuracy for the SVD based dimension reduced $\tilde{\Psi}$ remains stable for ranks greater than Rank(Ψ^*). This suggests that the number of rank proposed by CNMA can be considered for dimension reduction of matrix $\hat{\Psi}$. Moreover, $\tilde{\Psi}^*_{CNMA}$ produces competitive performance results to that of $\tilde{\Psi}^*_{LDA}$ and $\tilde{\Psi}^*_{HDP}$.

In [55], supervised LDA was run on MSRCv2 dataset. The highest classification accuracy obtained by running variational Bayes on LDA in [55] is 69%, which is 5% percent below the results obtained by CNMA. We have to emphasize that since CNM is an unsupervised approach for dimension reduction, its classification accuracy can be further improved by introducing class label information to CNM. We also ran similar simulations using the SIFT representation of the features proposed by [56] instead of blocks. The sparsity of matrix $\hat{\Psi}$ obtained by SIFT representation is lower than the sparsity of $\hat{\Psi}$ obtained using a block representation. The theory we present in this paper and the numerical evaluations in Section VI-B suggest that when α and β are large (lower sparsity), the rank recovery success region is diminished. This is consistent with the decrease in performance we observed.



(a) (b)

Fig. 9. Multiclass classification accuracy for Corel1000 dataset with number of clusters (a) 200 (b) 500.

B. Text datasets

We evaluate the classification accuracy of the proposed CNMA approach with HDP, LDA and SVD approaches on TDT2[5], Reuters[6], and 20Newsgroup[7] datasets. The TDT2 corpus consists of data collected during the first half of 1998 and taken from 6 sources including 2 newswires (APW, NYT), 2 radio programs (VOA, PRI), and 2 television programs (CNN, ABC), total 11201 documents in 96 different categories. The 20 Newsgroups dataset is a collection of approximately 20,000

[5]http://www.nist.gov/speech/tests/tdt/tdt98/index.htm
[6]http://www.daviddlewis.com/resources/testcollections/reuters21578/
[7]http://people.csail.mit.edu/jrennie/20Newsgroups/

newsgroup documents, partitioned (nearly) evenly across 20 different newsgroups. Reuters-21578 corpus contains 21578 documents in 135 categories. We use here the ModApte version of the Reuters dataset. Documents with multiple category labels are discarded leaving 8293 documents in 65 categories. In our experiments we removed documents with low number of words. Table II shows the summary of each dataset that we use in our analysis. We compare CNMA with HDP, LDA,

TABLE II
TEXT DATASET SUMMARY

	TDT2	20Newsgroup	Reuters
No. of documents	3807	4342	3228
Vocabulary size	4350	4612	3071
No. of category	30	20	10
Minimum no. of words per document (n_d)	180	150	50

and low-rank SVD approximation of matrix $\hat{\Psi}$. We use multi-class liblinear SVM[8], which is well suited for document classification. We use 5-fold cross validation to optimize the parameter C of the SVM algorithm. Figure 10 shows the results of classification for different datasets. We omitted the legend of Fig. 10(a) and Fig. 10(b) which are identical to the legend of Fig. 10(c). By comparing the results in Fig. 10, we observe that the performance of CNMA is competitive with HDP, LDA, and SVD. Moreover, the number of topics found by both CNMA and HDP algorithms is quite similar. This suggests that the dimension of the latent space discovered by HDP can be recovered by CNMA as well.

VIII. CONCLUSION

In this paper, we provided the framework of confidence-constrained rank minimization to recover the true number of topics (rank of the term-by-document matrix) in topic models and defined the problem as a parameter free convex optimization. We proposed the conditions under which the exact rank of the probability matrix Ψ can be recovered . Moreover, we showed that the reconstruction error is $\mathcal{O}(\sqrt{M/n})$, where M/n is the ratio of the number of document to the number of words per document. We devised a fast and accurate algorithms to solve CNM which enhances the applicability of CNM for a large real datasets.

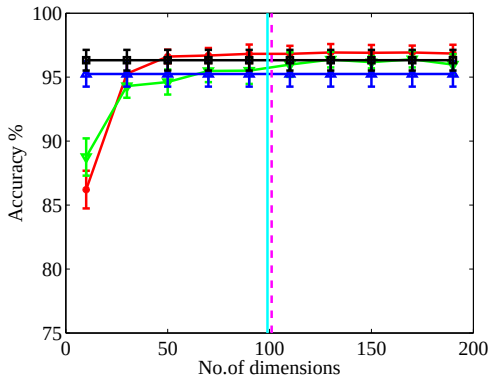
As future research direction, one can consider the following. The rank minimization problem was replaced heuristically with the nuclear norm minimization. Obtaining the conditions for which both rank and nuclear norm provide the same results can be considered. Our approach is an unsupervised technique in dimension reduction. Developing a new model which accounts for the useful discriminative information in the dataset is another future research direction.

APPENDIX A
DERIVATIVE OF $\frac{\lambda_1}{2}\|D_{\frac{1}{\lambda_1}}(\Psi')\|_F^2$ WITH RESPECT TO λ_1

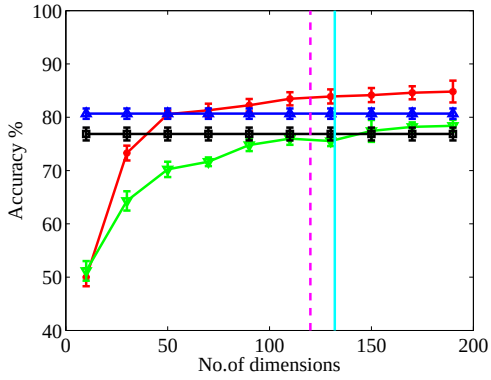
The derivative of $\frac{\lambda_1}{2}\|D_{\frac{1}{\lambda_1}}(\Psi')\|_F^2$ with respect to λ_1 is

$$\frac{d\frac{\lambda_1}{2}\|D_{\frac{1}{\lambda_1}}(\Psi')\|_F^2}{d\lambda_1} = \frac{1}{2}\|D_{\frac{1}{\lambda_1}}(\Psi')\|_F^2 + \frac{1}{\lambda_1}\|D_{\frac{1}{\lambda_1}}(\Psi')\|_*$$

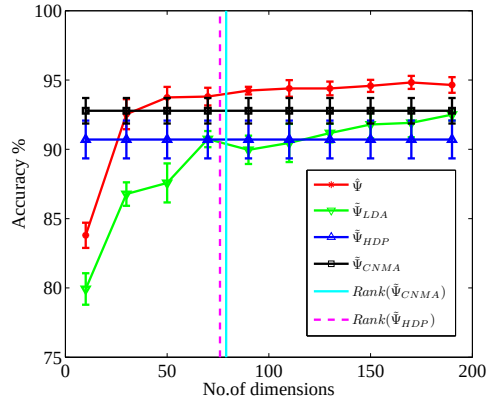
[8]http://www.csie.ntu.edu.tw/cjlin/liblinear/



(a)



(b)



(c)

Fig. 10. Classification accuracy for (a) TDT2, b) 20Newsgroup, and (c) Reuters

$$-\frac{1}{\lambda_1}\text{tr}((1\lambda_2^2 + \Lambda_3)^T D_{\frac{1}{\lambda_1}}(\Psi'))). \quad (18)$$

Proof:

Using the product rule, the derivative of $\frac{\lambda_1}{2}\|D_{\frac{1}{\lambda_1}}(\Psi')\|_F^2$ with respect to λ_1 can be expressed as:

$$\frac{d\frac{\lambda_1}{2}\|D_{\frac{1}{\lambda_1}}(\Psi')\|_F^2}{d\lambda_1} = \frac{1}{2}\|D_{\frac{1}{\lambda_1}}(\Psi')\|_F^2 + \frac{\lambda_1}{2}\frac{d\|D_{\frac{1}{\lambda_1}}(\Psi')\|_F^2}{d\lambda_1}. \quad (19)$$

Since $D_{\frac{1}{\lambda_1}}(\Psi') = U(S - \frac{1}{\lambda_1}I)_+V^T$, we have $\|D_{\frac{1}{\lambda_1}}(\Psi')\|_F^2 = \text{tr}\left(D_{\frac{1}{\lambda_1}}(\Psi')^T D_{\frac{1}{\lambda_1}}(\Psi')\right) = \text{tr}\left((S - \frac{1}{\lambda_1}I)_+^2\right)$. Therefore, the

second term on the RHS of (19) is

$$\frac{\lambda_1}{2}\frac{d}{d\lambda_1}(\|D_{\frac{1}{\lambda_1}(\Psi')}\|_F^2) = \frac{\lambda_1}{2}\frac{d}{d\lambda_1}\text{tr}\left((S-\frac{1}{\lambda_1}I)_+^2\right)$$
$$= \lambda_1\text{tr}\left(\frac{d(S-\frac{1}{\lambda_1}I)}{d\lambda_1}(S-\frac{1}{\lambda_1}I)_+\right)$$
$$= \lambda_1\text{tr}\left(\frac{dS}{d\lambda_1}(S-\frac{1}{\lambda_1}I)_+\right) + \frac{1}{\lambda_1}\text{tr}\left((S-\frac{1}{\lambda_1}I)_+\right)$$
$$= \lambda_1\text{tr}\left(\frac{dS}{d\lambda_1}(S-\frac{1}{\lambda_1}I)_+\right) + \frac{1}{\lambda_1}\|D_{\frac{1}{\lambda_1}}(\Psi')\|_*. \quad (20)$$

Since $\text{tr}\left(\frac{dS}{d\lambda_1}(S-\frac{1}{\lambda_1}I)_+\right) = \text{tr}\left((\frac{d\Psi'}{d\lambda_1})^T D_{\frac{1}{\lambda_1}}(\Psi')\right)$ [57], we have $\lambda_1\text{tr}\left(\frac{dS}{d\lambda_1}(S-\frac{1}{\lambda_1}I)_+\right) = -\frac{1}{\lambda_1}\text{tr}((1\underline{\lambda}_2^T + \Lambda_3)^T D_{\frac{1}{\lambda_1}}(\Psi')))$ and consequently

$$\frac{\lambda_1}{2}\frac{d\|D_{\frac{1}{\lambda_1}}(\Psi')\|_F^2}{d\lambda_1} = -\frac{1}{\lambda_1}\text{tr}((1\underline{\lambda}_2^T + \Lambda_3)^T D_{\frac{1}{\lambda_1}}(\Psi')))$$
$$+\frac{1}{\lambda_1}\|D_{\frac{1}{\lambda_1}}(\Psi')\|_*. \quad (21)$$

Substituting (21) into (19), we obtain (18).

APPENDIX B
PROOF OF PROBABILITY BOUND FOR ESTIMATION ERROR

To prove the probability bound for the estimation error of rank recovery in CRM, we defined two random quantities $Q = \sum_{d=1}^{M} n_d Q_d$ and $Q' = \sum_{d=1}^{M} Q_d$, where $Q_d = \sum_{l=1}^{L}(\Psi_{ld} - \hat{\Psi}_{ld})^2$. We use the one-tailed Chebyshev's inequality for random variable X as following:

$$P\left(X \geq E(X) + k\sqrt{Var(X)}\right) \leq \frac{1}{1+k^2}. \quad (22)$$

To compute the Chebyshev bound, we need to evaluate mean and variance of random quantity Q_d. First we start with calculation of the expected value of random variable Q_d.

$$E(Q_d) = \sum_{l=1}^{L} E(\hat{\Psi}_{ld} - \Psi_{ld})^2 = \text{Var}(\hat{\Psi}_{ld}) =$$
$$\sum_{l=1}^{L}\frac{\Psi_{ld}(1-\Psi_{ld})}{n_d} = \frac{1}{n_d}(1 - \sum_{l=1}^{L}\Psi_{ld}^2) \quad (23)$$

Note that $\text{Var}(\hat{\Psi}_d) = \frac{\Psi_{ld}(1-\Psi_{ld})}{n_d}$.

1) $Var(Q_d)$: The variance of Q_d can be calculated as follows (for notational ease we define $I_{ij} = I(X_i = j)$):

$$\text{Var}(Q_d) = \sum_{l=1}^{L}\sum_{m=1}^{L}\left(E\left[\left(\frac{1}{n_d}\sum_{i=1}^{n_d} I_{il} - \Psi_{ld}\right)^2\left(\frac{1}{n_d}\right.\right.\right.$$
$$\left.\left.\left.\sum_{j=1}^{n_d} I_{jm} - \Psi_{md}\right)^2\right] - E\left[\left(\frac{1}{n_d}\sum_{i=1}^{n_d} I_{il} - \Psi_{ld}\right)^2\right]\times$$
$$\left.E\left[\left(\frac{1}{n_d}\sum_{j=1}^{n_d} I_{jm} - \Psi_{md}\right)^2\right]\right) \quad (24)$$

We compute the second term on the RHS of (24) as follows:

$$E\left[\left(\frac{1}{n_d}\sum_{i=1}^{n_d} I_{il} - \Psi_{ld}\right)^2\right]E\left[\left(\frac{1}{n_d}\sum_{i=1}^{n_d} I_{jm} - \Psi_{md}\right)^2\right]$$
$$= \frac{\Psi_{ld}(1-\Psi_{ld})}{n_d} \times \frac{\Psi_{md}(1-\Psi_{md})}{n_d}$$

For the first term on the RHS of (24), we have:

$$E\left[\left(\frac{1}{n_d}\sum_{i=1}^{n_d} I_{il} - \Psi_{ld}\right)^2\left(\frac{1}{n_d}\sum_{j=1}^{n_d} I_{jm} - \Psi_{md}\right)^2\right] =$$
$$\frac{1}{n_d^4}\left(\sum_i\sum_j\sum_k\sum_t E\left[\left(I_{il} - \Psi_{ld}\right)\left(I_{jl} - \Psi_{ld}\right)\right.\right.$$
$$\left.\left.\left(I_{km} - \Psi_{md}\right)\left(I_{tm} - \Psi_{md}\right)\right]\right)$$

To evaluate $E[(I_{il} - \Psi_{ld})(I_{jl} - \Psi_{ld})(I_{km} - \Psi_{md})(I_{tm} - \Psi_{md})]$, we consider all the alternatives of i, j, k, l as follows (the enumeration of each alternative is specified in the bracket):

1) $[n_d]$ $i = j = k = t$

$$(I_{il} - \Psi_{ld})^2 = I_{il}(1 - 2\Psi_{ld}) + \Psi_{ld}^2$$
$$E[\left(I_{il}(1-2\Psi_{ld}) + \Psi_{ld}^2\right)(I_{im}(1-2\Psi_{md}) +$$
$$\Psi_{md}^2)] = \delta_{lm}\Psi_{ld}(1-2\Psi_{ld})^2 + \Psi_{ld}(1-2\Psi_{ld})$$
$$\Psi_{md}^2 + \Psi_{ld}^2\Psi_{md}(1-2\Psi_{md}) + \Psi_{ld}^2\Psi_{md}^2$$

2) $[4n_d(n_d-1)]$ $(i = j = k \neq t, i = j = t \neq k, i = k = t \neq j, j = k = t \neq i)$

$$E\left[(I_{il} - \Psi_{ld})^2(I_{im} - \Psi_{md})(I_{tm} - \Psi_{md})\right] = 0$$

3) $[n_d(n_d-1)]$ $i = j \neq k = t$

$$E\left[(I_{il} - \Psi_{ld})^2\right]E\left[(I_{jm} - \Psi_{md})^2\right] =$$
$$\frac{\Psi_{ld}(1-\Psi_{ld})}{n_d} \times \frac{\Psi_{md}(1-\Psi_{md})}{n_d}$$

4) $[2n_d(n_d-1)]$ $(i = k \neq j = t, i = t \neq j = k)$

$$2E\left[(I_{il} - \Psi_{ld})(I_{jm} - \Psi_{md})\right]^2 = 2[\delta_{lm}\Psi_{ld} -$$
$$\Psi_{ld}\Psi_{md} - \Psi_{ld}\Psi_{md} + \Psi_{ld}\Psi_{md}]^2$$
$$= 2(\delta_{lm}\Psi_{ld} - \Psi_{ld}\Psi_{md})^2 = 2(\delta_{lm}\Psi_{ld}^2(1-2\Psi_{ld})$$
$$+\Psi_{ld}^2\Psi_{md}^2)$$

5) $[6n_d(n_d-1)(n_d-2)]$ $(i = j \neq k \neq t,$ and all the combinations of 3 out of 4)

$$E[(I_{il} - \Psi_{ld})^2(I_{km} - \Psi_{md})(I_{tm} -$$
$$\Psi_{md})] = 0$$

6) $[n_d(n_d-1)(n_d-2)(n_d-3)]$ $i \neq j \neq k \neq t$

$$E[(I_{il} - \Psi_{ld})(I_{jl} - \Psi_{ld})(I_{km} - \Psi_{md})(I_{tm} -$$
$$\Psi_{md})] = 0$$

By adding all the alternatives from one to six and organizing them, we get the following expression for $Var(Q_d)$:

$$\begin{aligned}
Var(Q_d) &= \frac{2}{n_d^2}\sum_{l=1}^{L}\sum_{m=1}^{L}\left(\delta_{lm}\Psi_{ld}^2\left(1-2\Psi_{ld}\right)+\Psi_{ld}^2\Psi_{md}^2\right)\\
&+\frac{1}{n_d^3}\sum_{l=1}^{L}\sum_{m=1}^{L}\left(\delta_{lm}\Psi_{ld}\left(1-2\Psi_{ld}\right)^2+\Psi_{ld}\left(1-2\Psi_{ld}\right)\Psi_{md}^2\right.\\
&+\Psi_{ld}^2\Psi_{md}\left(1-2\Psi_{md}\right)+\Psi_{ld}^2\Psi_{md}^2\\
&-\Psi_{ld}(1-\Psi_{ld})\Psi_{md}(1-\Psi_{md})-2(\delta_{lm}\Psi_{ld}^2(1-2\Psi_{md})+\\
&\Psi_{ld}^2\Psi_{md}^2))\\
&=\frac{2}{n_d^2}\left(\sum_{l=1}^{L}\Psi_{ld}^2-2\sum_{l=1}^{L}\Psi_{ld}^3+\left(\sum_{l=1}^{L}\Psi_{ld}^2\right)^2\right)+\frac{1}{n_d^3}\left(8\sum_{l=1}^{L}\Psi_{ld}^3\right.\\
&\left.-6\left(\sum_{l=1}^{L}\Psi_{ld}^2\right)^2-2\sum_{l=1}^{L}\Psi_{ld}^2\right) \qquad (25)
\end{aligned}$$

The first component on RHS of (25) can be bounded using Cauchy-Schwartz as $\left(\sum\Psi_{ld}^{1.5}\Psi_{ld}^{0.5}\right)^2 \le \sum_l\Psi_{ld}^3\sum_l\Psi_{ld}$. Hence, $\left(\sum_l\Psi_{ld}^2\right)^2\le\sum_l\Psi_{ld}^3$. Thus,

$$\begin{aligned}
&\frac{2}{n_d^2}\left(\sum_{l=1}^{L}\Psi_{ld}^2-2\sum_{l=1}^{L}\Psi_{ld}^3+\left(\sum_{l=1}^{L}\Psi_{ld}^2\right)^2\right)\\
&\le\frac{2}{n_d^2}\left(\sum_{l=1}^{L}\Psi_{ld}^2-2(\sum_{l=1}^{L}\Psi_{ld}^2)^2+\left(\sum_{l=1}^{L}\Psi_{ld}^2\right)^2\right)\\
&=\frac{2}{n_d^2}\left(\sum_{l=1}^{L}\Psi_{ld}^2-\left(\sum_{l=1}^{L}\Psi_{ld}^2\right)^2\right)\\
&=\frac{2}{n_d^2}(t-t^2)=\frac{2}{n_d^2}(1/4-(t-1/2)^2)\le\frac{1}{2n_d^2},
\end{aligned}$$

where $t=\sum_{l=1}^{L}\Psi_{ld}^2$. For the second component term on RHS of (25) since $\sum_l\Psi_{ld}^3\le\sum_l\Psi_{ld}^2$, we have

$$\begin{aligned}
&\frac{1}{n_d^3}\left(8\sum_{l=1}^{L}\Psi_{ld}^3-6(\sum_{l=1}^{L}\Psi_{ld}^2)^2-2\sum_{l=1}^{L}\Psi_{ld}^2\right)\le\\
&\frac{6}{n_d^3}\left(\sum_{l=1}^{L}\Psi_{ld}^2-(\sum_{l=1}^{L}\Psi_{ld}^2)^2\right)\\
&=\frac{6}{n_d^3}(1/4-(t-1/2)^2)\le\frac{3}{2n_d^3}.
\end{aligned}$$

The mean of Q and Q' can be bounded as follows:

$$\begin{aligned}
E(Q)&=\sum_{d=1}^{M}n_dE(Q_d)=M-\sum_{d=1}^{M}\sum_{l=1}^{L}\Psi_{ld}^2\le M,\\
E(Q')&=\sum_{d=1}^{M}E(Q_d)=\sum_{d=1}^{M}\frac{1}{n_d}-\sum_{d=1}^{M}\sum_{l=1}^{L}\Psi_{ld}^2\le\sum_{d=1}^{M}\frac{1}{n_d},
\end{aligned}$$

since $-\sum_{d=1}^{M}\sum_{l=1}^{L}\Psi_{ld}^2\le 0$. Note that Q_d, $d=1,\ldots,M$ are $i.i.d.$ random variables, thus the variance of Q and Q' can be computed as the sum of variance of Q_d.

$$\begin{aligned}
Var(Q)&=\sum_{d=1}^{M}n_d^2Var(Q_d)\le\frac{M}{2}+\frac{3}{2}\sum_{d=1}^{M}\frac{1}{n_d}\\
Var(Q')&=\sum_{d=1}^{M}Var(Q_d)\le\sum_{d=1}^{M}\frac{1}{2n_d^2}+\sum_{d=1}^{M}\frac{3}{2n_d^3}.
\end{aligned}$$

Using the one-tailed Chebyshev inequality, we have the following probability bound for Q and Q':

$$\begin{aligned}
&P\left(Q\ge M+k\sqrt{\frac{M}{2}\left(1+3/M\sum_{d=1}^{M}\frac{1}{n_d}\right)}\right)\le\frac{1}{1+k^2},\\
&P\left(Q'\ge\sum_{d=1}^{M}\frac{1}{n_d}+k\sqrt{\left(\sum_{d=1}^{M}\frac{1}{2n_d^2}+\sum_{d=1}^{M}\frac{3}{2n_d^3}\right)}\right)\le\frac{1}{1+k^2}.
\end{aligned}$$

Alternatively, we say w.p. $1-\delta_k$, ,$\delta_k=\frac{1}{1+k^2}$, we have $Q=\sum_{d=1}^{M}\sum_{l=1}^{L}n_d\left(\hat{\Psi}_{ld}-\Psi_{ld}\right)^2\le\epsilon^2(\delta_k)$, where

$$\epsilon^2(\delta_k)=\epsilon^{*2}(\delta_k)=M+k\sqrt{\frac{M}{2}\left(1+3/M\sum_{d=1}^{M}\frac{1}{n_d}\right)},$$

and $Q'=\sum_{d=1}^{M}\sum_{l=1}^{L}\left(\hat{\Psi}_{ld}-\Psi_{ld}\right)^2\le\epsilon'^2(\delta_k)$, where

$$\epsilon'^2(\delta_k)=\epsilon'^{*2}(\delta_k)=\sum_{d=1}^{M}\frac{1}{n_d}+k\sqrt{\left(\sum_{d=1}^{M}\frac{1}{2n_d^2}+\sum_{d=1}^{M}\frac{3}{2n_d^3}\right)}.$$

REFERENCES

[1] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," in *Proceedings of Conference on Advances in Neural Information Processing Systems*, 2003, pp. 561–568.

[2] Z.J. Zha, X.S. Hua, T. Mei, J. Wang, G.J. Qi, and Z. Wang, "Joint multi-label multi-instance learning for image classification," in *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

[3] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American society for information science*, vol. 41, no. 6, pp. 391–407, 1990.

[4] T. Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of the 22nd International Conference on Research and Development in Information Retrieval*. ACM, 1999, pp. 50–57.

[5] D.M. Blei, A.Y. Ng, and M.I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[6] D.M. Blei, "Introduction to probabilistic topic models," *Available from http://www.cs.princeton.edu/ blei/papers/*, 2011.

[7] M. Welling, C. Chemudugunta, and N. Sutter, "Deterministic latent variable models and their pitfalls," in *Proceedings of International Conference on Data Mining*, 2008.

[8] L. AlSumait, D. Barbará, J. Gentle, and C. Domeniconi, "Topic significance ranking of LDA generative models," *Journal of Machine Learning and Knowledge Discovery in Databases*, pp. 67–82, 2009.

[9] Y.W. Teh, M.I. Jordan, M.J. Beal, and D.M. Blei, "Hierarchical Dirichlet processes," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1566–1581, 2006.

[10] D.M. Blei, T.L. Griffiths, and M.I. Jordan, "The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies," *Journal of the ACM*, vol. 57, no. 2, pp. 7, 2010.

[11] Z. Ghahramani, P. Sollich, and T. L. Griffiths, "Bayesian nonparametric latent feature models," *Bayesian Statistics*, 2007.

[12] A. Asuncion, M. Welling, P. Smyth, and Y.W. Teh, "On smoothing and inference for topic models," in *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 2009, pp. 27–34.

[13] B. Kanagal and V. Sindhwani, "Rank selection in low-rank matrix approximations: A study of cross-validation for NMFs," in *Proceedings of Conference on Advances in Neural Information Processing Systems*, 2010, vol. 1, pp. 10–15.
[14] P. Chen and D. Suter, "Recovering the missing components in a large noisy low-rank matrix: Application to SFM," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 8, pp. 1051–1063, 2004.
[15] T.T. Do, Y. Chen, N. Nguyen, L. Gan, and T.D. Tran, "A fast and efficient heuristic nuclear-norm algorithm for affine rank minimization," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*. 2009, pp. 3393–3396, IEEE.
[16] J.H. Manton, R. Mahony, and Y. Hua, "The geometry of weighted low-rank approximations," *IEEE Transactions on Signal Processing*, vol. 51, no. 2, pp. 500–514, 2003.
[17] G. Tang and A. Nehorai, "Lower bounds on the mean-squared error of low-rank matrix reconstruction," *IEEE Transactions on Signal Processing*, , no. 99, pp. 1–1, 2011.
[18] E.J. Candes, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis," *Journal of ACM*, vol. 58, no. 1, pp. 1–37, 2009.
[19] M. Fazel, *Matrix rank minimization with applications*, Ph.D. thesis, Stanford University, 2002.
[20] B. Recht, M. Fazel, and P.A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization, 2007," *SIAM Review*, vol. 52, pp. 471–501, 2010.
[21] E.J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Foundations of Computational Mathematics*, vol. 9, no. 6, pp. 717–772, 2009.
[22] B. Recht, W. Xu, and B. Hassibi, "Necessary and sufficient conditions for success of the nuclear norm heuristic for rank minimization," in *Proceedings of 47th IEEE Conference on Decision and Control*. IEEE, 2008, pp. 3065–3070.
[23] K. Konishi and T. Furukawa, "A nuclear norm heuristic approach to fractionally spaced blind channel equalization," *Signal Processing Letters*, vol. 18, no. 1, pp. 59–62, 2011.
[24] J.F. Cai, E.J. Candes, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *Journal on Optimization*, vol. 20, pp. 615–640, 2008.
[25] K.C. Toh and S. Yun, "An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems," *Pacific Journal of Optimization*, vol. 6, pp. 615–640, 2010.
[26] Y.J. Liu, D. Sun, and K.C. Toh, "An implementable proximal point algorithmic framework for nuclear norm minimization," *Mathematical Programming*, pp. 1–38, 2009.
[27] C. Lemaréchal and C. Sagastizábal, "Practical aspects of the Moreau-Yosida regularization I: theoretical properties," *Rapport de Recherche-Institut National de Recherche en Informatique et en Automatique*, 1994.
[28] Z. Lin, M. Chen, L. Wu, and Y. Ma, "The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices," *Mathematical Programming*, 2009.
[29] E.J. Candes and Y. Plan, "Matrix completion with noise," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 925–936, 2010.
[30] R.H. Keshavan, A. Montanari, and S. Oh, "Matrix completion from noisy entries," *Journal of Machine Learning Research*, vol. 99, pp. 2057–2078, 2010.
[31] H. Attias, "Inferring parameters and structure of latent variable models by variational Bayes," in *Proceedings of 15th Conference on Uncertainty in Artificial Intelligence*, 1999, vol. 2.
[32] H. Wallach, D. Mimno, and A. McCallum, "Rethinking LDA: Why priors matter," in *Proceedings of Conference on Advances in Neural Information Processing Systems*, 2009, vol. 22, pp. 1973–1981.
[33] M. Steyvers and T. Griffiths, "Probabilistic topic model," *Handbook of Latent Semantic Analysis*, pp. 1–15, 2007.
[34] N. Srebro, J.D.M. Rennie, and T. Jaakkola, "Maximum-margin matrix factorization," in *Proceedings of Conference on Advances in Neural Information Processing Systems*, 2005, vol. 17, pp. 1329–1336.
[35] T.K. Pong, P. Tseng, S. Ji, and J. Ye, "Trace norm regularization: Reformulations, algorithms, and multi-task learning," *Submitted to SIAM Journal on Optimization*, 2009.
[36] P.E. Gill, W. Murray, and M.H. Wright, *Practical optimization*, vol. 1, Academic press, 1981.
[37] M. Wax and T. Kailath, "Detection of signals by information theoretic criteria," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 33, no. 2, pp. 387–392, 1985.
[38] P.O. Perry and P.J. Wolfe, "Minimax rank estimation for subspace tracking," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 3, pp. 504–513, 2010.
[39] Z. Liu and L. Vandenberghe, "Semidefinite programming methods for system realization and identification," in *Proceedings of the 48th IEEE Conference on Decision and Control*. IEEE, 2009, pp. 4676–4681.
[40] K. Mohan and M. Fazel, "Reweighted nuclear norm minimization with application to system identification," in *Proceedings of Conference on American Control*. IEEE, 2010, pp. 2953–2959.
[41] F.R. Bach, "Consistency of trace norm minimization," *Journal of Machine Learning Research*, vol. 9, pp. 1019–1048, 2008.
[42] S. Negahban and M.J. Wainwright, "Estimation of (near) low-rank matrices with noise and high-dimensional scaling," *Submitted to the Annals of Statistics*, 2009.
[43] R. Meka, P. Jain, and I.S. Dhillon, "Guaranteed rank minimization via singular value projection," in *Proceedings of Conference on Advances in Neural Information Processing Systems*, 2010.
[44] B. Behmardi and R. Raich, "On provable exact low-rank recovery in topic models," in *Proceedings of IEEE International Workshop on Statistical Signal Processing*, 2011, pp. 265 –268.
[45] G.W. Stewart, "On the early history of the singular value decomposition," *SIAM review*, pp. 551–566, 1993.
[46] R. Meka, P. Jain, C. Caramanis, and I.S. Dhillon, "Rank minimization via online learning," in *Proceedings of the 25th International Conference on Machine learning*. ACM, 2008, pp. 656–663.
[47] J.P. Haldar and D. Hernando, "Rank-constrained solutions to linear matrix equations using powerfactorization," *Signal Processing Letters*, vol. 16, no. 7, pp. 584–587, 2009.
[48] D. Bertsekas, "On the Goldstein-Levitin-Polyak gradient projection method," *IEEE Transactions on Automatic Control*, vol. 21, no. 2, pp. 174–184, 1976.
[49] Y. Nesterov, "A method of solving a convex programming problem with convergence rate O (1/k2)," *Soviet Mathematics Doklady*, vol. 27, pp. 372–376, 1983.
[50] S.P. Boyd and L. Vandenberghe, *Convex optimization*, Cambridge University Press, 2004.
[51] B. Behmardi and R. Raich, "Convex optimization for exact rank recovery in topic models," in *Proceedings of IEEE International Workshop on Machine Learning for Signal Processing*, 2011, pp. 1–6.
[52] A.A. Goldstein, "Convex programming in Hilbert space," *American Mathematics Society*, vol. 70, no. 5, pp. 709–710, 1964.
[53] D.L. Donoho, I. Drori, Y. Tsaig, and J.L. Starck, "Sparse solution of underdetermined linear equations by stagewise orthogonal matching pursuit," *Citeseer*, 2006.
[54] Chih-Chung Chang and Chih-Jen Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 1–27, 2011, Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.
[55] B. Lakshminarayanan and R. Raich, "Inference in supervised latent Dirichlet allocation," in *Proceedings of IEEE International Workshop on Machine Learning for Signal Processing*. IEEE, 2011.
[56] D.G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the 7th IEEE International Conference on Computer Vision*. IEEE, 1999, vol. 2, pp. 1150–1157.
[57] T. Papadopoulo and M. Lourakis, "Estimating the jacobian of the singular value decomposition: Theory and applications," *Computer Vision-ECCV 2000*, pp. 554–570, 2000.

Behrouz Behmardi (S'10-11) received his B.S. degree in Industrial Engineering from Amir Kabir University of Technology (Tehran Polytechnic), Tehran, Iran, in 2002. He received his M.S. degree in Industrial Engineering from Khaje Nasir Toosi University of Technology, Tehran, Iran, in 2005. He received his M.S. in Industrial Engineering from Oregon State University in 2008 and is currently pursuing his Ph.D. in Electrical Engineering at the School of Electrical Engineering and Computer Science, Oregon State University, Corvallis. Behrouz Behmardi's research interests are in machine learning and signal processing with specific focus on statistical manifold learning and rank recovery. His main interest is in probabilistic latent space discovery with applications to text processing and computer vision.

Raviv Raich (S'98–M'04) received the B.Sc. and M.Sc. degrees from Tel Aviv University, Tel-Aviv, Israel, in 1994 and 1998, respectively, and the Ph.D. degree from Georgia Institute of Technology, Atlanta, in 2004, all in electrical engineering. Between 1999 and 2000, he was a Researcher with the Communications Team, Industrial Research, Ltd., Wellington, New Zealand. From 2004 to 2007, he was a Postdoctoral Fellow with the University of Michigan, Ann Arbor. Since fall 2007, he has been an Assistant Professor in the School of Electrical Engineering and Computer Science, Oregon State University, Corvallis. Raviv Raichs research interests are in statistical signal processing and machine learning. He has particular interest in applications concerning structure discovery in high dimensions. Raviv Raich serves as an Associate Editor for the IEEE Transactions on Signal Processing. He is a member of the Machine Learning for Signal Processing (MLSP) Technical Committee of the IEEE Signal Processing Society.