A Algorithm Analysis

A.1 ADMM update derivation

For completeness, we derive the ADMM steps of the problem in (12). Given current iterates V_1^t, γ^t, and Λ^t,

$$\begin{aligned}
V_1^{t+1} &= \operatorname*{arg\,min}_{V_1 \in \mathbb{R}^{n\times k}} \Big\{ \mu \|V_1\|_h + \frac{\rho}{2}\|\widetilde{W}\gamma^t - V_1\|_F^2 \\
&\qquad + \langle \Lambda_1, \widetilde{W}\gamma^t - V_1 \rangle \Big\} \\
&= \operatorname*{arg\,min}_{V \in \mathbb{R}^{n\times K}} \left\{ \frac{\mu}{\rho}\|V_1\|_h + \frac{1}{2}\|V_1 - \frac{\rho \widetilde{W}\gamma^t + \Lambda_1^t}{\rho}\|_F^2 \right\} \\
&= \text{Soft-Threshold}_{\mu/\rho}\left(\frac{\rho\widetilde{W}\gamma^t + \Lambda_1^t}{\rho} \right)
\end{aligned} \tag{14}$$

where we soft-threshold the matrix with the regularization parameter $\frac{\lambda}{\rho}$.

$$\begin{aligned}
V_2^{t+1} &= \operatorname*{arg\,min}_{V_2 \in \mathbb{R}^{k\times k}} \Big\{ \frac{\rho}{2}\|\gamma^t - V_2\|_F^2 + \langle \Lambda_2, \gamma^t - V_2 \rangle \\
&\qquad + \mathbb{1}(\lambda_{min}(V_2 V_2^T) \geq \frac{1}{R^2}) \Big\} \\
&= \operatorname*{arg\,min}_{V_2 \in \mathbb{R}^{k\times k}} \Big\{ \frac{1}{2}\|V_2 - \frac{\rho\gamma^t + \Lambda_2^t}{\rho}\|_F^2 \\
&\qquad + \mathbb{1}(\sigma_{min}(V_2) \geq \frac{1}{R}) \Big\} \\
&= \text{Proj}_{G_R}\left(\frac{\rho\gamma^t + \Lambda_2^t}{\rho} \right)
\end{aligned} \tag{15}$$

where $G_R = \{X \in \mathbb{R}^{n\times K} | \sigma_{min}(X) \geq \frac{1}{R}\}$ and Proj_{G_R} is the projection onto the set G_R.

$$\begin{aligned}
\gamma^{t+1} &= \operatorname*{arg\,min}_{\gamma \in \mathbb{R}^{k\times k}} \Big\{ -\log|\det \gamma\gamma^T| + \frac{\rho}{2}\|\widetilde{W}\gamma - V_1\|_F^2 \\
&\qquad + \langle \Lambda, \widetilde{W}\gamma - V_1 \rangle + \frac{\rho}{2}\|\gamma - V_2\|_F^2 \\
&\qquad + \langle \Lambda_2, \gamma - V_2 \rangle \Big\} \quad \text{s.t.} \quad \gamma \mathbf{1}_K = \mathbf{a} \\
&= \operatorname*{arg\,min}_{\gamma \in \mathbb{R}^{k\times k}} \Big\{ -\log|\det \gamma\gamma^T| + \frac{\rho}{2}\|C^{1/2}(\gamma - A)\|_F^2 \Big\} \\
&\qquad \text{s.t.} \quad \gamma \mathbf{1}_K = \mathbf{a}
\end{aligned} \tag{16}$$

where we have that

$$\begin{aligned}
A &= C^{-1}B^T = U D_A V^T \\
B &= (V_1^{t+1})^T \widetilde{W} + (V_2^{t+1})^T - \frac{(\Lambda_2)^t)^T}{\rho} - \frac{(\Lambda_1)^t)^T \widetilde{W}}{\rho} \\
C &= I + \widetilde{W}^T \widetilde{W}
\end{aligned}$$

We can derive the update for γ^{t+1}, as it is a convex problem with a linear constraint. First, consider the (16) without the linear constraint $\gamma \mathbf{1} = \mathbf{a}$. Then, we can rewrite the unconstrained γ-subproblem as

$$\begin{aligned}
\gamma_+ &= \operatorname*{arg\,min}_{\gamma \in \mathbb{R}^{k\times k}} \Big\{ -\log(\det \gamma\gamma^T) + \frac{\rho}{2}\|C^{1/2}(\gamma - A)\|_F^2 \Big\} \\
&= \operatorname*{arg\,min}_{\gamma \in \mathbb{R}^{k\times k}} \Big\{ -\log(\det \gamma\gamma^T) + \frac{\rho}{2}\operatorname{tr}(\gamma^T C\gamma) \\
&\qquad - \rho \operatorname{tr}(\gamma^T C A) \Big\} \\
&= \operatorname*{arg\,min}_{\gamma = UDV^T} \Big\{ -\log(\det \gamma\gamma^T) + \frac{\rho}{2}\operatorname{tr}(\gamma^T C\gamma) \\
&\qquad - \rho \operatorname{tr}(\gamma^T C A) \Big\} \\
&= \operatorname*{arg\,min}_{\gamma = UDV^T} \Big\{ -\log(\det D^2) + \frac{\rho}{2}\operatorname{tr}(UD^2U^T C) \\
&\qquad - \rho \operatorname{tr}(U D_A D U^T C) \Big\} \\
&= \operatorname*{arg\,min}_{\gamma = UDV^T} \Big\{ -\sum_{i=1}^{K} 2\log|D_{ii}| + \frac{\rho}{2}\operatorname{tr}(ED^2) - \rho \operatorname{tr}(FD) \Big\} \\
&= \operatorname*{arg\,min}_{\gamma = UDV^T} \Big\{ -\sum_{i=1}^{K} 2\log|D_{ii}| + \frac{\rho}{2}E_{ii}D_{ii}^2 - \rho F_{ii} D_{ii} \Big\}
\end{aligned}$$

where $E = U^T C U$ and $F = U^T C U D_A$. Then we can solve the above problem element by element. Looking at the i-th entry, we can take the derivative and set it to zero. That is

$$\frac{\partial}{\partial D_{ii}}\left(\log|D_{ii}| + \frac{\rho}{2}E_{ii}D_{ii}^2 - \rho F_{ii} D_{ii} \right) = 0$$

leading to the following quadratic formula

$$D_{ii}^2 - \frac{F_{ii}}{E_{ii}} D_{ii} - \frac{2}{\rho E_{ii}} = 0$$

which has the solution

$$\widehat{D}_{ii} = \frac{\frac{F_{ii}}{E_{ii}} + \sqrt{\frac{F_{ii}^2}{E_{ii}^2} + \frac{8}{\rho E_{ii}}}}{2}$$

Then, using these diagonal elements $\hat{D}_{ii}$, it follows that

$$\begin{aligned}
\gamma_+ &= \operatorname*{arg\,min}_{\gamma \in \mathbb{R}^{k\times k}} \Big\{ -\log(\det \gamma\gamma^T) + \frac{\rho}{2}\|C^{1/2}(\gamma - A)\|_F^2 \Big\} \\
&= U\widehat{D}V^T
\end{aligned}$$

We make the final adjustment to satisfy the linear constraint. Thus, the γ update is

$$\gamma^{(t+1)} = \gamma_+ - (\gamma_+ \mathbf{1} - \mathbf{a})(\mathbf{1}^T C^{-1}\mathbf{1})^{-1}\mathbf{1}^T C^{-1}$$

A.2 Proof of Proposition 3.2

Proof. The first order conditions of the updates in Algorithm 1 give us

$$
\begin{aligned}
&0 \in \partial\|\cdot\|_{h,\mu}(V_1^{t+1}) - \rho(\widetilde{W}\gamma^t - V_1^{t+1}) - \Lambda_1^t \\
&0 \in \mathbb{1}_{G_R}(V_2^{t+1}) - \rho(\gamma^t - V_2^{t+1}) - \Lambda_2^t \\
&0 \in -2(\gamma^{t+1})^{-T} + \rho\widetilde{W}^T(\widetilde{W}\gamma^{t+1} - V_1^{t+1}) + \widetilde{W}^T\Lambda_1^t + \\
&\quad \rho(\gamma^{t+1} - V_2^{t+1}) + \Lambda_2^t + \mathbf{1}_K(\nu^{t+1})^T \text{ s.t. } \gamma^{t+1}\mathbf{1} = \mathbf{a}
\end{aligned} \tag{17}
$$

Note that the first order condition for γ^{t+1} is different as it is a equality constrained convex problem. Also, by the definitions of Λ_1^{t+1} and Λ_2^{t+1}

$$
\begin{aligned}
\Lambda_1^{t+1} &= \Lambda_1^t + \rho(\widetilde{W}\gamma^{t+1} - V_1^{t+1}) \\
\Lambda_2^{t+1} &= \Lambda_2^t + \rho(\gamma^{t+1} - V_2^{t+1})
\end{aligned} \tag{18}
$$

Then, combining these two sets of equations, we have that

$$
\begin{aligned}
&\Lambda_1^{t+1} + \rho\widetilde{W}(\gamma^t - \gamma^{t+1}) \in \partial\|\cdot\|_{h,\mu}(V_1^{t+1}) \\
&\Lambda_2^{t+1} + \rho(\gamma^t - \gamma^{t+1}) \in \partial\mathbb{1}_{G_R}(V_2^{t+1}) \\
&2(\gamma^{t+1})^{-T} - \mathbf{1}_K(\nu^{t+1})^T = \widetilde{W}^T\Lambda_1^{t+1} + \Lambda_2^{t+1} \\
&\frac{1}{\rho}(\Lambda_1^{t+1} - \Lambda_1^t) = \widetilde{W}\gamma^{t+1} - V_1^{t+1} \\
&\frac{1}{\rho}(\Lambda_2^{t+1} - \Lambda_2^t) = \gamma^{t+1} - V_2^{t+1}
\end{aligned} \tag{19}
$$

Then, let us define $(\gamma^t, V_1^t, V_2^t, \Lambda_1^t, \Lambda_2^t)_{t=1}^\infty$ be a sequence of iterates with a limit point $(\gamma^*, V_1^*, V_2^*, \Lambda_1^*, \Lambda_2^*)$. Then, by the last two equations of (19), we have that $\widetilde{W}\gamma^* = \widetilde{W}V_2^* = V_1^*$. Therefore, the first two equations give us that

$$
\begin{aligned}
\Lambda_1^* &\in \partial\|\cdot\|_{h,\mu}(V_1^*) = \partial\|\cdot\|_{h,\mu}(\widetilde{W}\gamma^*) \\
\Lambda_2^* &\in \partial\mathbb{1}_{G_R}(V_2^*) = \partial\mathbb{1}_{G_R}(\gamma^*)
\end{aligned}
$$

Lastly, using the third equation in (19), it follows that

$$
\begin{aligned}
&2(\gamma^*)^{-T} - \mathbf{1}_K(\nu^*)^T = \\
&\quad \widetilde{W}^T\Lambda_1^* + \Lambda_2^* \in \partial\|\cdot\|_{h,\mu}(\widetilde{W}\gamma^*) + \partial\mathbb{1}_{G_R}(\gamma^*)
\end{aligned}
$$

Noting that the optimality condition for $\arg\min_\gamma -\log|\det(\gamma\gamma^T)|$ s.t. $\gamma\mathbf{1} = \mathbf{a}$ is

$$
-2(\gamma^*)^{-T} + \mathbf{1}_K(\nu^*)^T = 0 \quad \text{and } \gamma^*\mathbf{1} = \mathbf{a}
$$

We have that

$$
\begin{aligned}
\mathbf{0} &= -2(\gamma^*)^{-T} + \mathbf{1}_K(\nu^*)^T + 2(\gamma^*)^{-T} - \mathbf{1}_K(\nu^*)^T \\
&= -2(\gamma^*)^{-T} + \mathbf{1}_K(\nu^*)^T + \widetilde{W}^T\Lambda_1^* + \Lambda_2^* \in \partial f(\gamma^*)
\end{aligned}
$$

and we have that $\gamma^*\mathbf{1} = \mathbf{a}$ by the formulation of our update for γ^t. This shows that γ^* satisfies the optimality condition of (10) and thus a stationary point for f. □

B Simulations

We demonstrate the computational benefit as well as the accuracy of our model in terms of perplexity. The experiments are based on the simulated data from the LDA model, and we focus on the comparison to the variational EM (VEM) and Gibbs sampling to illustrate the advantages of our method. As part of the future work, we plan to compare the stochastic implementation of MVTM with GDM (Yurochkin and Nguyen, 2016) and the imporved implementations of the Gibbs sampling presented in Li et al. (2014) and Yuan et al. (2015) at a much larger scale.

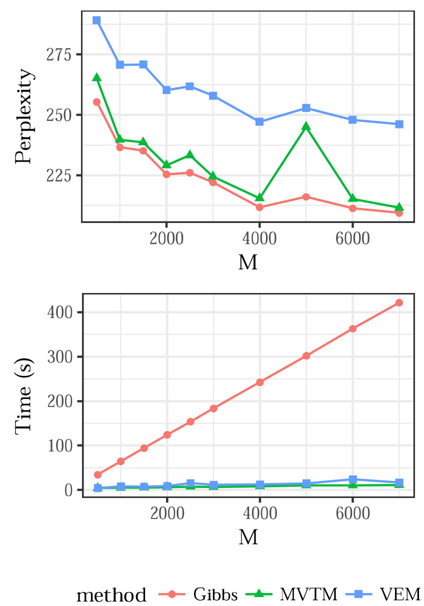

Figure 7: Perplexity of the held-out data and the corresponding time complexity of each method at varying values of the number of documents M with $N_m = 1000$, $K = 5$, $V = 1200$, $\eta = 0.1$ and $\alpha = 0.1$

We first look at the behavior of the algorithms as M increases when $N_m = 1000$ (Figure 8). At $N_m = 1000$, we are working with the setting that is close to the asymptotic regime, and MVTM has the computational speed comparable to VEM and the statistical performance similar to the Gibbs sampling.

In a more challenging case with the shorter documents at $N_m = 100$, MVTM continues to perform as well as the Gibbs sampling with a little additional computational cost. This performance comparison would be of interest for the researchers who are working with shorter documents present in the modern application. As discussed in Tang et al. (2014) and Nguyen (2015), the limitation of LDA comes from the document lengths. Our results show that MVTM

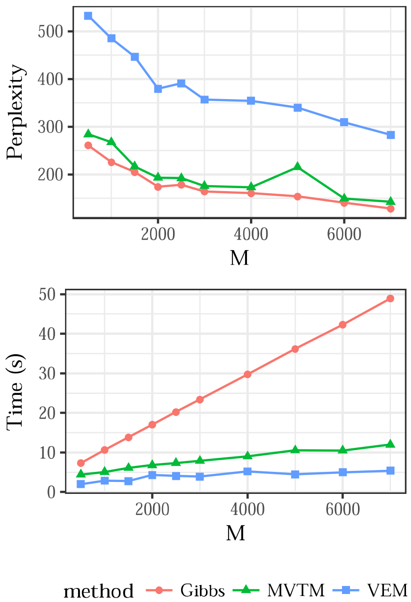

Figure 8: Perplexity of the held-out data and the corresponding time complexity of each method at varying values of the number of documents M with $N_m = 100$, $K = 5$, $V = 1200$, $\eta = 0.1$ and $\alpha = 0.1$

do not suffer from the short documents in terms of statistical performance, when the regularization parameter μ for the hinge loss is appropriately chosen. The current batch implementation, however, suffers from the number of documents present in the dataset, as it has to soft-threshold every document. This computational limitation, however, can be alleviated by the stochastic implementation as demonstrated in the stochastic implementation of the variational method in Hoffman et al. (2013).

C NIPS dataset Topics

C.1 Computational Time

Figure 9 shows the time complexities of different algorithms on the NIPS dataset as we increase the number of topics. Compared to GDM, the proposed MVTM improvement on performance comes at a little computational cost. RecoverKL could achieve similar computational speed if the anchor words are provided. However, when we include the computational cost of finding the anchor words, GDM and MVTM show computational advantages over RecoverKL.

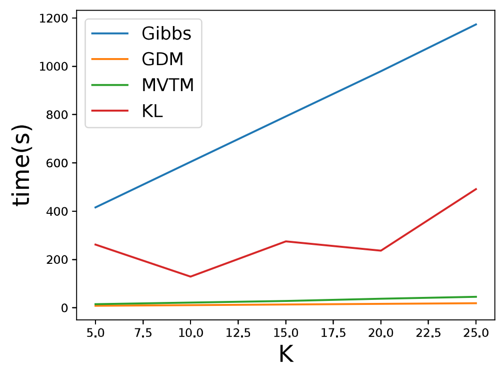

Figure 9: The computational performance of different algorithms as a function of the number of topics. NIPS dataset includes 1491 documents and 4492 unique words.

C.2 Top 10 topics

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
neuron	input	training	training	algorithm	unit	model	network	function	learning
network	output	set	error	learning	network	data	neural	set	system
input	system	network	set	data	input	parameter	system	approximation	control
model	circuit	recognition	data	problem	hidden	distribution	problem	result	function
pattern	signal	data	cell	weight	weight	system	training	linear	action
neural	neural	algorithm	input	method	output	object	control	bound	algorithm
synaptic	network	vector	network	function	layer	gaussian	dynamic	number	task
learning	chip	learning	classifier	distribution	learning	likelihood	unit	point	reinforcement
cell	weight	classifier	weight	vector	pattern	cell	result	network	error
spike	analog	word	test	parameter	training	mixture	point	threshold	model

Table 2: Top 10 MVTM topic for NIPS dataset

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
neuron	circuit	recognition	set	model	network	function	image	model	learning
cell	signal	speech	training	memory	input	algorithm	object	data	control
model	system	word	data	representation	unit	learning	images	distribution	system
input	neural	system	algorithm	node	weight	point	field	gaussian	action
activity	analog	training	error	rules	neural	vector	map	parameter	model
synaptic	chip	hmm	performance	tree	output	result	visual	mean	dynamic
pattern	output	character	classifier	structure	learning	case	motion	algorithm	policy
response	current	model	classification	level	training	problem	feature	probability	algorithm
firing	input	network	number	graph	layer	parameter	direction	method	reinforcement
cortex	neuron	context	learning	rule	hidden	equation	features	component	problem

Table 3: Top 10 Gibbs topic for NIPS dataset

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
neuron	input	word	data	image	network	model	cell	learning	learning
network	output	speech	set	images	unit	data	visual	algorithm	control
spike	weight	recognition	training	object	neural	parameter	motion	function	model
synaptic	neural	system	error	point	weight	likelihood	direction	problem	system
input	network	training	function	features	hidden	mixture	response	action	task
pattern	net	character	vector	graph	training	distribution	orientation	policy	movement
firing	chip	hmm	method	representation	output	algorithm	neuron	optimal	controller
model	layer	speaker	classifier	feature	input	set	model	gradient	motor
activity	analog	context	kernel	information	error	gaussian	frequency	convergence	dynamic
neural	bit	network	gaussian	recognition	function	variables	field	step	reinforcement

Table 4: Top 10 GDM topic for NIPS dataset