A Geometric structure of normalized points from a cone

Lemma A.1. Let \( y_i = \frac{z_i}{\|z_i\|} \), then \( y_i^T = r_i \phi_i^T Y_P \) for \( r_i = \frac{m_i^T 1}{\|m_i^T Y_P\|} \geq 1 \), and \( \phi_i = (\phi_{i1}, \phi_{i2}, \cdots, \phi_{iK})^T \), \( \phi_{ij} = \frac{m_{ij}}{\sum_j m_{ij}} \).

Proof. \( y_i^T = \frac{z_i}{\|z_i\|} = \frac{m_i^T Y_P}{\|m_i^T Y_P\|} = \frac{m_i^T 1}{\|m_i^T Y_P\|} m_i^T Y_P = r_i \phi_i^T Y_P \). Clearly \( m_i^T Y_P \| = \| \sum_j m_{ij} y_{1(j)} \| \leq \sum_j m_{ij} \| y_{1(j)} \| = \sum_j m_{ij} = m_i^T 1 \), so \( r_i \geq 1 \).

Proof of Lemma 2.1. Since \( \text{rank}(P) = K \), we have \( VEV^T = P = \rho \Theta B \Theta^T \Gamma \). W.L.O.G, let \( \Theta(1,:) = I \), then \( V_P EV^T = \rho \Gamma P \Theta^T \Gamma \). Now \( VE = PV = \rho \Theta B \Theta^T \Gamma V = \Theta \Gamma_2^{-1} V_P \Theta^T EV = \Theta \Gamma_2^{-1} V_P E \), right multiplying \( E^{-1} \) gives \( V = \Theta \Gamma_2^{-1} V_P \). Also consider that \( V_P EV^T = \rho \Gamma P \Theta^T \), \( V_P \) is full rank.

B Identifiability of DCMMSB-type Models

Lemma B.1. For DCMMSB-type models such that \( f(\theta_i) = 1 \), \( \forall i \in [n] \) for some degree 1 homogeneous function \( f \) (e.g., \( f(\theta) = \|\theta\|_p \)), the sufficient conditions for \( (\Theta, B, \Gamma) \) to be identifiable up to a permutation of the communities are (a) there is at least one pure node in each community, (b) \( \sum_i \gamma_i = n \), (c) \( B \) has unit diagonal.

Proof. From Lemma 2.1, we have \( V = \Theta \Gamma_2^{-1} V_P \) and \( V_P \) is full rank. Suppose two set of parameters \( \{\Gamma^{(1)}, \Theta^{(1)}, B^{(1)}\} \) and \( \{\Gamma^{(2)}, \Theta^{(2)}, B^{(2)}\} \) yield the same \( P \) (W.L.O.G, we abort \( \rho \) in \( B \)) and each has a pure node set \( P_1 \) and \( P_2 \) and W.L.O.G, assume the permutation of the communities is fixed, i.e., \( \Theta^{(1)}_{P_1} = \Theta^{(2)}_{P_2} = I \). Then,

\[
\Gamma^{(1)} \Theta^{(1)} (\Gamma^{(1)}_{P_1})^{-1} V_{P_1} = \Gamma^{(2)} \Theta^{(2)} (\Gamma^{(2)}_{P_2})^{-1} V_{P_2}.
\]

(3)

Taking indices \( P_1 \) and \( P_2 \) respectively on \( V \), we have,

\[
V_{P_1} = \Gamma^{(2)}_{P_1} \Theta^{(2)}_{P_1} (\Gamma^{(2)}_{P_2})^{-1} V_{P_2} \quad \text{and} \quad V_{P_2} = \Gamma^{(1)}_{P_2} \Theta^{(1)}_{P_2} (\Gamma^{(1)}_{P_1})^{-1} V_{P_1}.
\]

(4)

Then,

\[
V_{P_1} = \Gamma^{(2)}_{P_1} \Theta^{(2)}_{P_1} (\Gamma^{(2)}_{P_2})^{-1} \Gamma^{(1)}_{P_2} \Theta^{(1)}_{P_2} (\Gamma^{(1)}_{P_1})^{-1} V_{P_1} \Rightarrow I = \Gamma^{(2)}_{P_1} \Theta^{(2)}_{P_1} (\Gamma^{(2)}_{P_2})^{-1} \Gamma^{(1)}_{P_2} \Theta^{(1)}_{P_2} (\Gamma^{(1)}_{P_1})^{-1}, \quad \text{as } V_{P_1} \text{ is full rank.}
\]

(5)

As \( \Gamma^{(2)}_{P_1} \Theta^{(2)}_{P_1} (\Gamma^{(2)}_{P_2})^{-1} \) and \( \Gamma^{(1)}_{P_2} \Theta^{(1)}_{P_2} (\Gamma^{(1)}_{P_1})^{-1} \) are all nonnegative, using Lemma 1.1 of [5], they are both generalized permutation matrices. Also since \( \Gamma^{(2)}_{P_1}, \Gamma^{(1)}_{P_2} \) are diagonal matrix, \( \Theta^{(2)}_{P_1} \) must be
Algorithm B

SVM-cone-topic

Algorithm A

SVM-cone-DCMMSB

In this section we provide the detailed algorithms for parameter estimations of DCMMSB, OCCAM (Algorithm A) and Topic Models (Algorithm B). These algorithms both reply on the one class SVM (Algorithm A) for finding the corner rays and then use those for parameter estimation, the details of which vary from model to model. Note for Algorithm A step 7 is to normalize rows of $\theta$ by $\ell_1$ norm, if we normalize by $\ell_2$ norm, then it can be used for estimation of OCCAM.

Algorithm A SVM-cone-DCMMSB

Input: Adjacency matrix $A \in \mathbb{R}^{n \times n}$, number of communities $K$

Output: Estimated degree parameters $\Gamma$, community memberships $\hat{\Theta}$, and community interaction matrix $B$

1: Get top-$K$ eigen-decomposition of $A$ as $\tilde{V}\tilde{E}\tilde{V}^T$
2: Normalize rows of $\tilde{V}$ by $\ell_2$ norm
3: Use SVM-cone to get pure node set $C$ and estimated $\tilde{M}$
4: $\tilde{V}_C = \tilde{V}(C,:)$, get $N_C$ from row norms of $\tilde{V}_C$
5: $\tilde{D}_C = \sqrt{\text{diag}(N_C\tilde{V}_C\tilde{E}\tilde{V}_C^T\tilde{N}_C)}$
6: $\hat{F} = \text{diag}(\tilde{M}\tilde{D}_C\mathbf{1}_K)$
7: $\hat{\Theta} = \hat{F}^{-1}\tilde{M}D$
8: $\hat{\Gamma} = n\hat{F}/(1\hat{F}1)$, $\hat{\Gamma}_C = \hat{\Gamma}(C,C)$
9: $B = \hat{\Gamma}_C^{-1}\tilde{V}_C\tilde{E}\tilde{V}_C\hat{\Gamma}_C^{-1}$
10: $B = B/\max_{i,j}B_{ij}$

Algorithm B SVM-cone-topic

Input: Word-document count matrix $A \in \mathbb{R}^{V \times D}$, number of topics $K$

Output: Estimated word-topic matrix $\hat{T}$

1: Randomly splitting the words in each document to two halves to get $A_1$ and $A_2$
2: Normalize columns of $A_1$ and $A_2$ by $\ell_1$ norm to get $\hat{A}_1$ and $\hat{A}_2$
3: Get top-$K$ SVD of $U = \hat{A}_1\hat{A}_2^T$ as $\bar{V}\bar{E}\bar{V}^T$
4: Normalize rows of $\bar{V}$ by $\ell_2$ row norm
5: Use SVM-cone to get pure node set $C$ and estimated $\hat{M}$
6: Normalizing columns of $\hat{D}_C$ by $\ell_1$ norm to get $\hat{T}$
D Corner finding with One-class SVM with population inputs

Lemma D.1. If Proj$_{\text{Conv}(y_P^i)}(0)$ is an interior point in Conv$(y_P^T)$, then One-class SVM can find all the K corners with $m_{ij} = 1$ as support vectors given $y_i, i \in [n]$ as inputs. And a sufficient condition for this to hold is $(y_P y_P^T)^{-1}1 > 0$.

Proof. The primal problem of One-class SVM in (7) is

$$\min \frac{1}{2} \|w\|^2 - b \quad \text{s.t.} \quad w^T y_i \geq b, \; i \in [n].$$

First of all note that $b \geq 0$ because if $b < 0$, we can always make $b = 0$ to satisfy the condition and decrease the value of the objective function. From Lemma 2.1 we have $y_i^T = r_i \phi_i^T y_P$. As $r_i \geq 1$, if there exists $(w,b)$ that $w^T y_i \geq b, \; i \in I$, we have

$$w^T y_i = r_i \phi_i^T y_P w = r_i \sum_{j} \phi_{ij} w^T y_{(j)} \geq r_i b \geq b, \; i \in [n].$$

So we can reduce the problem to using points $i \in I$ as inputs. Furthermore, we consider an equivalent primal problem and its dual:

$$\text{Primal : } \max \quad b \quad \text{s.t.} \quad \|w\| \leq 1, \quad w^T y_i \geq b, \; i \in I$$

$$\text{Dual : } \min \frac{1}{2} \sum_{i,j} \beta_{ij} y_i^T y_j \quad \text{s.t.} \quad \sum_{i} \beta_i = 1, \quad \beta_i \geq 0, \; i \in I$$

The dual problem is basically to find a point in Conv$(y_P^T)$ that has the minimum norm (closest to origin). Now denote the optimal function value for the dual problem as $L_{y_P}$ and for any subset $S \subset I$, let $L_{y_P(S;i)}$ be the optimal value when we want to find a point in Conv$(y_P^T(S;i))$ that has the minimum norm.

Let $N \in \mathbb{R}^{n \times n}$ be a diagonal matrix such that $N_{ii} = 1/\|z_i\|$, then $y_P = N_P z_P$ is also full rank. If for $\beta^* = \arg \min_{\beta} L_{y_P}(\beta)$, each coordinate is strictly larger than 0, it is easy to see that $L_{y_P} > L_{y_P(S;i)}$ since $y_P$ is full rank. So a sufficient condition for One-class SVM to find all K corners of $L_{y_P}$ is $\beta^* > 0$, which means the closest point to origin in Conv$(y_P^T)$ is an interior point (also the projection of origin to Conv$(y_P^T)$).

Now we will show a sufficient condition for this.

Suppose the $\beta^* > 0$. First let us find a hyperplane $(w,d)$ that is through columns of $y_P^T$ with $d < 0$ (since $y_P$ is full rank, we must have $d \neq 0$). We have $y_P^T w = d1$. Since the distance from origin to hyperplane $(w,d)$ is $\frac{d}{\|w\|}$, Proj$_{\text{Conv}(y_P^i)}(0)$ is an interior point in Conv$(y_P^T)$, we have

$$y_P^T \beta^* = \text{Proj}_{\text{Conv}(y_P^i)}(0) = \frac{d}{\|w\|} \frac{w}{\|w\|}$$

Then,

$$w^T y_P^T \beta^* = \frac{dw^T w}{\|w\|^2} = d.$$

As $w^T y_P^T = d1^T$, we have $d1^T \beta^* = d$, so $1^T \beta^* = 1$. So the only condition left to be satisfied is that $\beta^* > 0$, using Eq. (7).

$$y_P y_P^T \beta^* = \frac{d y_P w}{\|w\|^2} = \frac{d (d1)}{\|w\|^2} = \frac{d^2}{\|w\|^2}.$$

so $\beta^* = \frac{d^2}{\|w\|^2} (y_P y_P^T)^{-1}1 > 0$ and all we require is:

$$(y_P y_P^T)^{-1}1 > 0.$$

Proof of Theorem 2.3 Using Lemma 2.1 we have:

$$I = V^T V = V_P^T \Gamma^{-1} \Theta^T \Gamma^2 \Theta \Gamma^{-1} v_P \quad \Rightarrow \quad (V_P V_P^T)^{-1} = \Gamma^{-1} \Theta^T \Gamma^2 \Theta \Gamma^{-1}.$$  (8)
Since $Y_P = N_p V_P$, we have:

$$(Y_P Y_P^T)^{-1} = N_p^{-1} \Gamma_p^{-1} \Theta^T \Gamma^2 \Theta^{-1} N_p^{-1}. \quad (9)$$

On the RHS of Eq. (9), $N_p^{-1}, \Gamma_p^{-1}$, and $\Gamma$ are all diagonal matrix with strictly positive diagonal elements, then diagonal of $(Y_P Y_P^T)^{-1}$ must be strictly positive, as the $i$-th element on its diagonal is proportional to $\|\Theta(\cdot, i)\|^2$, and since $\Theta$ is nonnegative, we can easily get that $(Y_P Y_P^T)^{-1} 1 > 0$. So for DCMMSE-type models, it is always true that the closest point in Conv$(Y_P^T)$ to origin is an interior point of Conv$(Y_P^T)$.

\[\square\]

E Corner finding with One-class SVM with empirical inputs

**Lemma E.1.** Let $\epsilon = \max_i \|y_i - \hat{y}_i\|$. Denote $(w, b)$ and $(\hat{w}, \hat{b})$ be the optimal solution for the primal problem of One-class SVM in (6) with population $(y_1, y_2, \ldots, y_n)$ and empirical inputs $(\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n)$ respectively, then $|\hat{b} - b| \leq \epsilon$.

**Proof.** First we have $w^T y_i \geq b, \forall i \in [n]$, and $\|w^T (\hat{y}_i - y_i)\| \leq \epsilon$. Then $w^T \hat{y}_i = w^T y_i + w^T (\hat{y}_i - y_i) \geq b - \epsilon$. As $(w, b - \epsilon)$ is a feasible solution of the primal problem with empirical inputs, by optimality of $b$, we have $\hat{b} \geq b - \epsilon$. Similarly we can get $b \geq \hat{b} - \epsilon$, so $|\hat{b} - b| \leq \epsilon$. \[\square\]

**Lemma E.2.** Let $(w, b)$, $(\hat{w}, \hat{b})$ be the hyperplane of the optimal solution of One-class SVM with population and empirical inputs respectively, then $\|\hat{w} - w\| \leq \zeta \epsilon$, for $\zeta = \frac{4K}{\eta b^2 \sqrt{\lambda_K(YPY_P^T)}} < \pi$.

**Proof.** Let $\beta_i, i \in I$ be the solution of the dual problem in Eq. (6) with population inputs, from the construction of this dual problem, we know $w = \frac{\sum_{i \in I} \beta_i y_i}{\sum_{i \in I} \beta_i y_i^2} = b Y_P Y_P^T \mathbf{1}$. From the condition of the primal problem, $Y_P \hat{w} \geq \hat{b} \mathbf{1}$, then we have $Y_P \hat{w} = \hat{Y}_P \hat{w} - (Y_P - Y_P) \hat{w} \geq (\hat{b} - \epsilon) \mathbf{1} \geq (b - 2\epsilon) \mathbf{1}$. Then there exists a vector $c \geq 0$ such that $Y_P \hat{w} = (b - 2\epsilon) \mathbf{1} + c$. Now let $\tilde{w} = Y_P^T \varphi + \hat{\omega}_\perp$, where $Y_P \hat{\omega}_\perp = \mathbf{0}$. So $Y_P \hat{w} = Y_P Y_P^T \varphi = (b - 2\epsilon) \mathbf{1} + c$, which gives $\hat{w} = Y_P^T (Y_P Y_P^T)^{-1} ((b - 2\epsilon) \mathbf{1} + c) + \hat{\omega}_\perp$. Since $\tilde{w} = 1$, we have

$$1 = \|\tilde{w}\|^2 = ((b - 2\epsilon) \mathbf{1} + c + \hat{\omega}_\perp)^T (Y_P Y_P^T)^{-1} ((b - 2\epsilon) \mathbf{1} + c + \hat{\omega}_\perp)^2 + 2 \|\hat{\omega}_\perp\|^2$$

$$= b^2 1^T (Y_P Y_P^T)^{-1} 1 + 2b 1^T (Y_P Y_P^T)^{-1} (c - 2\epsilon \mathbf{1}) + (c - 2\epsilon \mathbf{1})^T (Y_P Y_P^T)^{-1} (c - 2\epsilon \mathbf{1}) + 2 \|\hat{\omega}_\perp\|^2.$$

Since $1 = \|w\|^2 = b^2 1^T (Y_P Y_P^T)^{-1} 1$, we have

$$0 \leq (c - 2\epsilon \mathbf{1})^T (Y_P Y_P^T)^{-1} (c - 2\epsilon \mathbf{1}) + \|\hat{\omega}_\perp\|^2$$

$$= -2b 1^T (Y_P Y_P^T)^{-1} (c - 2\epsilon \mathbf{1}) - 4bc 1^T (Y_P Y_P^T)^{-1} \mathbf{1} \leq 2\epsilon / b^2.$$

which uses that $(Y_P Y_P^T)^{-1}$ is positive definite. This gives

$$2b 1^T (Y_P Y_P^T)^{-1} \mathbf{1} \leq 4bc 1^T (Y_P Y_P^T)^{-1} \mathbf{1} = 4bc / b^2$$

which by Condition 2, we know $(\min_i 1^T (Y_P Y_P^T)^{-1} e_i) \geq \eta$, so $\|c\| \leq \epsilon / (2\epsilon / b^2)$.

Let $\hat{P}$ be the set of support vectors returned by empirical One-class SVM, and $\hat{\beta}$ as the optimal solution for the dual problem, then $\tilde{w} = \hat{Y}_P \hat{\beta} / b$ and $\sum_{j \in \hat{P}} \hat{\beta}_j = 1$. Now we will give an upper bound on $\|\hat{w}_\perp\|$. For any $v \in \text{span}(Y_P)$, we have $\|\hat{w}_\perp\| \leq \|\hat{w} - v\|$. Now take $v = Y_P^T \hat{\beta} / b$, since all rows of $Y$ lie in the span of $Y_P$, this choice of $v$ also lies in the span of $Y_P$. Thus,

$$\|\hat{w}_\perp\| \leq \|\hat{w} - v\| = \|\hat{Y}_P \hat{\beta} - Y_P^T \hat{\beta} / b\| = \|\sum_{j \in \hat{P}} \hat{\beta}_j (y_j - \hat{y}_j) / \hat{b} \| \leq \epsilon / (b - \epsilon).$$
Now, we have
\[
\hat{w} - w = Y^T_P (Y_P Y_P^T)^{-1}((b - 2\epsilon)1 + c) + \hat{w}_\perp - bY^T_P (Y_P Y_P^T)^{-1}1 = Y^T_P (Y_P Y_P^T)^{-1}(c - 2\epsilon 1) + \hat{w}_\perp,
\]
\[
\|\hat{w} - w\|^2 = (c - 2\epsilon 1)^T (Y_P Y_P^T)^{-1}(c - 2\epsilon 1) + \|\hat{w}_\perp\|^2 \leq c^T (Y_P Y_P^T)^{-1}c + 4\epsilon^2/b^2 + \epsilon^2/(b - \epsilon)^2
\]
\[
\leq \|c\|^2 \lambda_1 ((Y_P Y_P^T)^{-1}) + 4\epsilon^2/b^2 + \epsilon^2/(b - \epsilon)^2 \leq \left(\frac{4}{\eta^2 b^2 \lambda_1 (Y_P Y_P^T)^{-1}} + \frac{4}{b^2} + \frac{1}{(b - \epsilon)^2}\right) \epsilon^2,
\]
where we use Eq. (10) to get that the cross terms are non-negative for the first inequality. First
\[
\eta \leq \frac{4}{\eta^2 b^2 \lambda_1 (Y_P Y_P^T)^{-1}} + \frac{4}{b^2} + \frac{1}{(b - \epsilon)^2}
\]

Using Lemma E.2,
\[
1/b^2 = 1^T (Y_P Y_P^T)^{-1}1 \leq K \lambda_1 ((Y_P Y_P^T)^{-1}) = K/\lambda_1 (Y_P Y_P^T).
\]

Lemma E.3. Let $(\hat{w}, \hat{b})$ be the hyperplane of the optimal solution of One-class SVM with empirical inputs, then $\hat{b}1 \leq Y_P \hat{w} \leq \hat{b}1 + (\zeta + 2)\epsilon 1$.

Proof. Using Lemma E.2
\[
\hat{Y}_P \hat{w} = Y_P \hat{w} + (\hat{Y}_P - Y_P) \hat{w} \leq Y_P w + Y_P (\hat{w} - w) + \epsilon 1 \leq b1 + (\zeta + \epsilon)1 \leq \hat{b}1 + (\zeta + 2)\epsilon 1.
\]

Lemma E.4. Let $(w, b)$, $(\hat{w}, \hat{b})$ be the hyperplane of the optimal solution of One-class SVM with population and empirical inputs respectively, and $S$ be the set of nodes selected as support vectors in the optimal solution of the dual problem with empirical inputs. Then for $r_i$ defined in Lemma A.1
\[
r_i - 1 \leq \frac{1}{b(2\epsilon)^{-1}}, \forall i \in S. Furthermore, \forall i \in [n], if $\hat{w}^T \hat{y}_i \leq \hat{b} + (\zeta + 2)\epsilon$, then $r_i - 1 \leq \frac{1}{\epsilon b^{-2\epsilon}}$.
\]

Proof. First $\forall i \in S$, we have,
\[
\hat{b} = \hat{w}^T \hat{y}_i = \hat{w}^T y_i + \hat{w}^T (\hat{y}_i - y_i) = r_i \sum_j \phi_{ij} \hat{w}^T y_{I(j)} + \hat{w}^T (\hat{y}_i - y_i)
\]
\[
= r_i \sum_j \phi_{ij} \hat{w}^T \hat{y}_{I(j)} + r_i \sum_j \phi_{ij} \hat{w}^T (y_{I(j)} - \hat{y}_{I(j)}) + \hat{w}^T (y_i - y_i)
\]
\[
\geq r_i \hat{b} - r_i \epsilon - \epsilon.
\]
This gives
\[
r_i \leq \frac{\hat{b} + \epsilon}{b - \epsilon} \implies r_i - 1 \leq \frac{2\epsilon}{b - \epsilon} \leq \frac{2\epsilon}{b - \epsilon} = \frac{1}{b(2\epsilon)^{-1}}
\]
where the last step uses $b \geq \hat{b} - \epsilon$ from Lemma E.1 Similarly, for $i \in [n]$ such that $\hat{w}^T \hat{y}_i \leq \hat{b} + (\zeta + 2)\epsilon$, we have $\hat{b} + (\zeta + 2)\epsilon \geq r_i \hat{b} - r_i \epsilon - \epsilon$ and this gives $r_i - 1 \leq \frac{1}{\epsilon b^{-2\epsilon}}$.

Lemma E.5. For S defined in Lemma E.4 $\forall i \in S$, $\exists j \in [K]$ such that for $\phi_{ij}$ defined in Lemma A.1 $\phi_{ij} \geq 1 - \epsilon_1$, for $\epsilon_1 = \frac{2\epsilon}{b \lambda_1 (Y_P Y_P^T)^{-1}}$. Furthermore, $\forall i \in [n]$, if $\hat{w}^T \hat{y}_i \leq \hat{b} + (\zeta + 2)\epsilon$, then $\exists j \in [K]$, $\phi_{ij} \geq 1 - \epsilon_2$, for $\epsilon_2 = \frac{2\epsilon}{b(1 + (\zeta + 2)\epsilon) \lambda_1 (Y_P Y_P^T)^{-1}}$.

Proof. By Lemma E.4, we have $r_i \leq 1 + \frac{1}{b(2\epsilon)^{-1}} = 1 + 1/b$. As $y_i = r_i \phi_i^T Y_P$, we have $1 = \|y_i\| = r_i \|\phi_i^T Y_P\|$, so $\|\phi_i^T Y_P\| \geq 1 - 2\epsilon/b$. Let $y_{-k} = \sum_{j \neq k} \phi_{ij} y_{I(j)}$, $\forall k \in [K]$. then $\phi_i^T Y_P = \phi_{ik} y_{I(k)} + (1 - \phi_{ik}) y_{-k}$. It is easy to see that $\|y_{-k}\| \leq 1$, then
\[
\|\phi_i^T Y_P\|^2 = \phi_i^2 + (1 - \phi_{ik})^2 + 2 \phi_{ik}(1 - \phi_{ik}) y_{I(k)} y_{-k},
\]
\[
y_{I(k)} y_{-k} = \sum_{j \neq k} \phi_{ij} y_{I(j)} y_{I(k)} y_{-k} \leq \max_{j \neq k} y_{I(k)} y_{I(j)} \leq \max_{i \neq l} y_{I(i)} y_{I(l)}.
\]
Using $2\mathbf{x}_1^T \mathbf{x}_2 = \|\mathbf{x}_1\|^2 + \|\mathbf{x}_2\|^2 - \|\mathbf{x}_1 - \mathbf{x}_2\|^2$ for any same length vectors $\mathbf{x}_1$ and $\mathbf{x}_2$, and

\[
\|\mathbf{y}_{I_{(i)}} - \mathbf{y}_{I_{(j)}}\|^2 = \| (\mathbf{e}_i - \mathbf{e}_j)^T \mathbf{y}_p \|^2 = (\mathbf{e}_i - \mathbf{e}_j)^T \mathbf{y}_p \mathbf{y}_p^T (\mathbf{e}_i - \mathbf{e}_j) \\
\geq \min_{|x|=1} x^T \mathbf{y}_p \mathbf{y}_p^T x = 2 \lambda_K (\mathbf{y}_p \mathbf{y}_p^T),
\]

we have $\max_{i \neq j} \mathbf{y}_p^T \mathbf{y}_{I_{(i)}} \leq 1 - \lambda_K (\mathbf{y}_p \mathbf{y}_p^T)$. Then,

\[
(1 - 2\epsilon/b)^2 \leq \| \phi_i^T \mathbf{y}_p \|^2 \leq \phi_i^2 (1 - \phi_{ik})^2 + 2 \phi_{ik} (1 - \phi_{ik}) (1 - \lambda_K (\mathbf{y}_p \mathbf{y}_p^T))
\]

\[
= 1 - 2 \phi_{ik} (1 - \phi_{ik}) \lambda_K (\mathbf{y}_p \mathbf{y}_p^T),
\]

which gives $\phi_{ik} (1 - \phi_{ik}) \leq \frac{2(\lambda_K (\mathbf{y}_p \mathbf{y}_p^T))}{b(1 - \lambda_K (\mathbf{y}_p \mathbf{y}_p^T))} =: \epsilon_1, \forall k \in [K]$. Since $\sum_k \phi_{ik} = 1$, we must have $\exists j \in [K], \phi_{ij} \geq 1 - \epsilon_1$. Similarly, for $i \in [n]$ such that $\mathbf{w}^T \mathbf{y}_i \leq \hat{b} + (\zeta + 2)\epsilon$, we have $r_i - 1 \leq \frac{(\zeta + 4)\epsilon}{b - 2\epsilon}$ from Lemma E.4 then $\phi_i^T \mathbf{y}_p = \frac{1}{r_i} \geq 1 - \frac{(\zeta + 4)\epsilon}{b + (\zeta + 2)\epsilon}$, and this gives that $\phi_{ik} (1 - \phi_{ik}) \leq \frac{2(\lambda_K (\mathbf{y}_p \mathbf{y}_p^T))}{b(1 - \lambda_K (\mathbf{y}_p \mathbf{y}_p^T))} =: \epsilon_2 < \frac{2\epsilon}{b(1 - \lambda_K (\mathbf{y}_p \mathbf{y}_p^T))}$, using $\zeta \geq 4$ and $(\zeta + 2)\epsilon \geq 0$. Also since $\sum_k \phi_{ik} = 1$, we must have $\exists j \in [K], \phi_{ij} \geq 1 - \epsilon_2$. \hfill \Box

**Remark** E.1. Lemma E.5 shows that for One-class SVM with empirical inputs, the support vectors selected are all nearly corner points. Lemma E.6 shows that each corner point is closed to the hyperplane ($\hat{w}, \hat{b}$) selected by One-class SVM by $(\zeta + 2)\epsilon$, and then Lemma E.7 shows that each point close to ($\hat{w}, \hat{b}$) by $(\zeta + 2)\epsilon$ are all nearly corner points. So choosing points that are $(\zeta + 2)\epsilon$ close to ($\hat{w}, \hat{b}$) will guarantee us all the K corner points and some nearly corner points.

**Lemma E.6.** Let $S_c = \{i : \mathbf{w}^T \mathbf{y}_i \leq \hat{b} + (\zeta + 2)\epsilon\}$, then $\forall i, j \in S_c$, for $\epsilon_3 = \epsilon + \frac{(\zeta + 4)\epsilon}{b - 2\epsilon}$, we have

\[
\| \phi_i - \phi_j \| \sqrt{\lambda_K (\mathbf{y}_p \mathbf{y}_p^T)} - 2\epsilon_3 \leq \| \mathbf{y}_i - \mathbf{y}_j \| \leq \| \phi_i - \phi_j \| \sqrt{\lambda_1 (\mathbf{y}_p \mathbf{y}_p^T)} + 2\epsilon_3.
\]

**Proof.** First we have, $\| \mathbf{y}_i - \phi_i^T \mathbf{y}_p \| = \| \mathbf{y}_i - r_i \phi_i^T \mathbf{y}_p + (r_i - 1) \phi_i^T \mathbf{y}_p \| \leq \epsilon + \frac{\epsilon}{b - 2\epsilon} =: \epsilon_3$, where last step is by Lemma E.4. This gives $\| \phi_i^T \mathbf{y}_p - \phi_j^T \mathbf{y}_p \| \leq 2\epsilon_3$. Then we have $\| \phi_i^T \mathbf{y}_p - \phi_j^T \mathbf{y}_p \| \leq 2\epsilon_3 \leq \| \mathbf{y}_i - \mathbf{y}_j \| = \| \phi_i - \phi_j \| \sqrt{\lambda_K (\mathbf{y}_p \mathbf{y}_p^T)} + 2\epsilon_3$. Combining

\[
\| \phi_i - \phi_j \| \sqrt{\lambda_K (\mathbf{y}_p \mathbf{y}_p^T)} \leq \| \phi_i^T \mathbf{y}_p - \phi_j^T \mathbf{y}_p \| \leq \| \phi_i - \phi_j \| \sqrt{\lambda_1 (\mathbf{y}_p \mathbf{y}_p^T)},
\]

we have the result. \hfill \Box

**Lemma E.7.** Let $S_c = \{i : \mathbf{w}^T \mathbf{y}_i \leq \hat{b} + (\zeta + 2)\epsilon\}$, then there exists exact K clusters in $S_c$, given $\epsilon \leq c \frac{n(\zeta + 4)\epsilon}{K^{1.5}}$, for some constant $c$. \hfill \Box

**Proof.** First because $I \in S_c$, from Lemma E.5 there exists at least K clusters in $S_c$. By Lemma E.5 $\forall i \in S_c, \exists k_i \in [K], \phi_{ik_i} \geq 1 - \epsilon_2$. If $k_i = k_j$, by Lemma E.6

\[
\| \mathbf{y}_i - \mathbf{y}_j \| \leq \| \phi_i - \phi_j \| \sqrt{\lambda_1 (\mathbf{y}_p \mathbf{y}_p^T)} + 2\epsilon_3 \leq \sqrt{3} \epsilon_2 \sqrt{\lambda_1 (\mathbf{y}_p \mathbf{y}_p^T)} + 2\epsilon_3.
\]

This means if $j$ is a corner point, $i$ will be close to it, and will be in the same cluster as long as there is enough separation between different clusters. Now we will prove this is true. Similarly, if $k_i \neq k_j$,

\[
\| \mathbf{y}_i - \mathbf{y}_j \| \geq \| \phi_i - \phi_j \| \sqrt{\lambda_K (\mathbf{y}_p \mathbf{y}_p^T)} - 2\epsilon_3 \geq \sqrt{2(1 - 2\epsilon_2)} \sqrt{\lambda_K (\mathbf{y}_p \mathbf{y}_p^T)} - 2\epsilon_3.
\]

In order to have enough separation between $p$ clusters, we need

\[
\sqrt{2(1 - 2\epsilon_2)} \sqrt{\lambda_K (\mathbf{y}_p \mathbf{y}_p^T)} - 2\epsilon_3 = \sqrt{2} \sqrt{\lambda_K (\mathbf{y}_p \mathbf{y}_p^T)} - 2\sqrt{2} \epsilon_2 \sqrt{\lambda_K (\mathbf{y}_p \mathbf{y}_p^T)} - 2\epsilon_3
\]

\[
> c' (\sqrt{3} \epsilon_2 \sqrt{\lambda_1 (\mathbf{y}_p \mathbf{y}_p^T)} + 2\epsilon_3),
\]

where $c'$ is a constant.
for some constant $c' > 2$. This is equivalent to show

$$\sqrt{2} > (2\sqrt{2} + \sqrt{3c'} \sqrt{\kappa(Y_p Y_p^T)}) \epsilon_2 + \frac{2 + 2c'}{\sqrt{\lambda_K(Y_p Y_p^T)}} \epsilon_3.$$ 

As

$$\leq (2\sqrt{2} + \sqrt{3c'} \sqrt{\kappa(Y_p Y_p^T)}) \epsilon_2 + \frac{2 + 2c'}{b \lambda_K(Y_p Y_p^T)} \epsilon_3 \leq \sqrt{2},$$

where $c_i, i \in [4]$ are some constants we do not specify and we use $1/b^2 \leq K/\lambda_K(Y_p Y_p^T)$ in the second last inequality. So a sufficient condition for separated clusters is $c_i = \frac{K^{1.5}}{\eta(\lambda_K(Y_p Y_p^T))^3} \epsilon$, which is

$$\epsilon \leq c_i \frac{\eta(\lambda_K(Y_p Y_p^T))^3}{K^{1.5}} \sqrt{\kappa(Y_p Y_p^T)},$$

for some constant $c_i$.

### F Consistency of inferred parameters

#### Lemma F.1. For set $C$ returned by Algorithm 7, there exits a permutation matrix $\Pi \in \mathbb{R}^{K \times K}$ that $\|\hat{Y}_C - \Pi Y_p\|_F \leq \epsilon_4$, for $\epsilon_4 = \frac{c_Y \kappa \eta}{(\lambda_K(Y_p Y_p^T))^{1.5}} \epsilon$ and $c_Y$ is some constant.

**Proof.** By Lemma E.5 we know that $\forall i \in S_C, \exists j \in [K]$ such that $\phi_{ij} \geq 1 - \epsilon_2$. Then we have:

$$\|\hat{y}_i - y_{I(j)}\| \leq \|\hat{y}_i - y_i\| + \|y_i - y_{I(j)}\| \leq \epsilon + \sum_{i \neq j} \phi_{ij}Y_{I(j)} + \|y_{I(j)}\| + \|Y_{I(1)} - 1\|Y_{I(j)}\|

\leq \epsilon + \left(1 + \frac{(\zeta + 4)\epsilon}{b - 2\epsilon}\right) (2\epsilon_2) + \frac{(\zeta + 4)\epsilon}{b - 2\epsilon}

\leq \epsilon + \frac{4\epsilon}{b} + 4\epsilon_2 \leq \frac{c_Y \sqrt{K}}{b \lambda_K(Y_p Y_p^T)} \epsilon \leq \frac{c_Y \sqrt{K}}{b \lambda_K(Y_p Y_p^T)} \epsilon \leq \frac{c_Y \sqrt{K}}{b \lambda_K(Y_p Y_p^T)} \epsilon,$$

where we use $\epsilon \leq b/(4\zeta)$ and $\zeta \geq 4$. And $c_Y$ is a constant. Then $\|\hat{Y}_C - \Pi Y_p\|_F \leq \frac{c_Y \kappa \eta}{(\lambda_K(Y_p Y_p^T))^{1.5}} \epsilon$. 

#### Lemma F.2. Let $\max_i \|e_i^T(Z - \hat{Z})\| = \epsilon_0$, then $\|y_i - \hat{y}_i\| \leq 2\epsilon_0 / |z_i|$. 

**Proof.** First note that by definition $\|z_i\| - \|\hat{z}_i\| \leq \epsilon_0$, then,

$$\|y_i - \hat{y}_i\| = \left\|\frac{z_i}{|z_i|} - \frac{\hat{z}_i}{|\hat{z}_i|}\right\| \leq \left\|\frac{z_i}{|z_i|} - \frac{\hat{z}_i}{|\hat{z}_i|}\right\| + \left\|\frac{\hat{z}_i}{|\hat{z}_i|} - \frac{z_i}{|z_i|}\right\| \leq \left\|\frac{z_i}{|z_i|} - \frac{\hat{z}_i}{|\hat{z}_i|}\right\| + \left\|\frac{\hat{z}_i}{|\hat{z}_i|} - \frac{z_i}{|z_i|}\right\| \leq 2\epsilon_0 / |z_i|.$$
Proof of Theorem 2.8: First let us get some important intermediate bounds. Using Weyl’s inequality,
\[ |\sigma_i(\hat{Y}_C) - \sigma_i(Y_P)| \leq \|\hat{Y}_C - \Pi Y_P\| \leq \epsilon_4 \]
\[ |\lambda_i(\hat{Y}_C\hat{Y}_C^T) - \lambda_i(Y_P Y_P^T)| = |\sigma_i^2(\hat{Y}_C) - \sigma_i^2(Y_P)| \leq (\sigma_i(\hat{Y}_C) + \sigma_i(Y_P)) \epsilon_4 \]
\[ \leq (2\sigma_i(Y_P) + \epsilon_4) \epsilon_4. \]
Secondly,
\[ \|\left(\hat{Y}_C\hat{Y}_C^T\right)^{-1}\| = \frac{1}{\lambda_K(\hat{Y}_C\hat{Y}_C^T)} \leq \frac{1}{\lambda_K(Y_P Y_P^T) - (2\sigma_K(Y_P) + \epsilon_4) \epsilon_4} \leq \frac{2}{\lambda_K(Y_P Y_P^T)}, \]
where we use \((2\sigma_K(Y_P) + \epsilon_4) \epsilon_4 < \lambda_K(Y_P Y_P^T)/2\). Then,
\[ \|\Pi(Y_P Y_P^T)^{-1} - (\hat{Y}_C\hat{Y}_C^T)^{-1}\Pi\| = \|\Pi(Y_P Y_P^T)^{-1} - (\hat{Y}_C\hat{Y}_C^T)^{-1}\Pi\| = \|\Pi(Y_P Y_P^T)^{-1} - (\hat{Y}_C\hat{Y}_C^T)^{-1}\Pi\| \leq \|\Pi(Y_P Y_P^T)^{-1} - (\hat{Y}_C\hat{Y}_C^T)^{-1}\Pi\| \leq \frac{1}{2} \|\Pi(Y_P Y_P^T)^{-1}\| \|\Pi(Y_P Y_P^T)^{-1}\| \leq \frac{1}{2} \|\Pi(Y_P Y_P^T)^{-1}\|. \]

Note that \(M = Z(Y_P Y_P^T)^{-1}. \) Let \(\max_i \|e_i^T(Z - \hat{Z})\| = \epsilon_0\), then,
\[ \|e_i^T(M - Z\hat{Y}_C\hat{Y}_C^T)^{-1}\Pi\| = \|e_i^T(Z(Y_P Y_P^T)^{-1} - \hat{Y}_C\hat{Y}_C^T)^{-1}\Pi\| \leq \|e_i^T(Z - \hat{Z})\| \|\Pi(Y_P Y_P^T)^{-1}\| \|\Pi(Y_P Y_P^T)^{-1}\| \leq \|e_i^T(Z - \hat{Z})\| \|\Pi(Y_P Y_P^T)^{-1}\| \|\Pi(Y_P Y_P^T)^{-1}\| \leq \frac{1}{2} \|\Pi(Y_P Y_P^T)^{-1}\|. \]

where we uses \(\epsilon_4 \leq \|Y_P\|/2, \epsilon_0 < \|e_i^T Z\|/2\) for relaxations. \(\square\)

G  Equivalence of using \(\hat{V}\) and \(\hat{V}V^T\) as input of Algorithm [1]

Lemma G.1. Let \(u_i = e_i^T U = v_i/\|v_i\|, y_i = e_i^T Y = V v_i/\|V v_i\|, \hat{u}_i = e_i^T \hat{U} = \hat{v}_i/\|\hat{v}_i\|, \hat{y}_i = e_i^T \hat{Y} = \hat{V} \hat{v}_i/\|\hat{V} v_i\|. One-class SVM using rows of \(U\) (or \(\hat{U}\)) and rows of \(Y\) (or \(\hat{Y}\)) will return the same solution \(\beta\).

Proof. Since \(y_i = V v_i/\|V v_i\| = V v_i/\|v_i\| = v_i, \) and \(\hat{y}_i = \hat{V} \hat{v}_i/\|\hat{V} v_i\| = \hat{V} \hat{v}_i/\|v_i\| = \hat{V} \hat{u}_i, \) we have \(y_i^T y_j = u_i^T V^T v_j = u_i^T v_j \) and \(\hat{y}_i^T \hat{y}_j = \hat{u}_i^T \hat{V}^T \hat{v}_j = \hat{u}_i^T \hat{u}_j. \) It is easy to see that One-class SVM using rows of \(U\) (or \(\hat{U}\)) and rows of \(Y\) (or \(\hat{Y}\)) have the same objective function (Eq. 6) and thus will have the same solution of \(\beta, i \in [n]. \)

Remark G.1. By Lemmas [G.1, D.1] and Theorem 2.3 One-class SVM with \(y_i = V v_i/\|V v_i\|, \) \(i \in [n]\) as inputs can find all the \(K\) corners corresponding to the pure nodes as support vectors for DCMMSS-type models. Furthermore, as \(Y_C = \hat{U}C\hat{V}^T, \)
\[ \hat{M} = \hat{Y}\hat{Y}_C^{-1} = \hat{V}\hat{V}_C^{-1} = \hat{V}\hat{U}_C\hat{U}^T \]
which shows that outputs of Algorithm [1] using \(\hat{V}\) and \(\hat{V}V^T\) as input are same.
H DCMMSB-type models properties

Lemma H.1. For DCMMSB-type models, $\gamma_i / \sqrt{\lambda_1(\Theta^T \Gamma^2 \Theta)} \leq \|v_i\| \leq \psi \gamma_i / \sqrt{\lambda_1(\Theta^T \Gamma^2 \Theta)}$, $\forall i \in [n]$, and $\gamma_i / \sqrt{\lambda_1(\Theta^T \Gamma^2 \Theta)} \leq \|v_i\| \leq \gamma_i / \sqrt{\lambda_1(\Theta^T \Gamma^2 \Theta)}$, $\forall i \in I$. Note that $\psi = 1$ for DCMMSB and $\sqrt{K}$ for OCCAM.

Proof. Eq. (8) gives $((\Gamma_p^{-1} V_p) (\Gamma_p^{-1} V_p)^T)^{-1} = \Theta^T \Gamma^2 \Theta$, then,

$$\max_i \|e_i(\Gamma_p^{-1} V_p)\|^2 = \max_i e_i^T (\Gamma_p^{-1} V_p) (\Gamma_p^{-1} V_p)^T e_i \leq \max_i x^T (\Gamma_p^{-1} V_p) (\Gamma_p^{-1} V_p)^T x$$

$$= \lambda_1((\Gamma_p^{-1} V_p) (\Gamma_p^{-1} V_p)^T) = 1/\lambda_K(\Theta^T \Gamma^2 \Theta)$$

$$\min_i \|e_i(\Gamma_p^{-1} V_p)\|^2 = \min_i e_i^T (\Gamma_p^{-1} V_p) (\Gamma_p^{-1} V_p)^T e_i \geq \min_i \|x\|^2 (\Gamma_p^{-1} V_p) (\Gamma_p^{-1} V_p)^T x$$

$$= \lambda_1((\Gamma_p^{-1} V_p) (\Gamma_p^{-1} V_p)^T) = 1/\lambda_1(\Theta^T \Gamma^2 \Theta).$$

By Lemma 2.1 $\forall i \in [n]$, $v_i^T = \gamma_i \theta_i^T \Gamma_p^{-1} V_p$, then $\|v_i\| \leq \psi \gamma_i \max_i \|e_i(\Gamma_p^{-1} V_p)\| \leq \psi \gamma_i / \sqrt{\lambda_K(\Theta^T \Gamma^2 \Theta)}$, where $\psi = \max_i \|\theta_i\|_1$. Similarly, $\|v_i\| \geq \gamma_i \min_i \|\theta_i\| \min_i \|e_i(\Gamma_p^{-1} V_p)\| \geq \gamma_i / \sqrt{\lambda_1(\Theta^T \Gamma^2 \Theta)}$. Note that $\psi = 1$ for DCMMSB and $\sqrt{K}$ for OCCAM. In general $\psi \geq 1$ and $\min_i \|\theta_i\| = 1$ for any DCMMSB-type model that satisfies $\|\theta_i\|_p \geq 1$.

\[\square\]

Lemma H.2. For DCMMSB-type models whose eigenvectors has the form in Lemma 2.1 if using $Z = VV^T$, $M = \Gamma \Theta \Gamma_p^{-1} N_p^{-1}$, then:

$$\lambda_1(\gamma_p Y_p Y_p^T) \leq \kappa(\Theta^T \Gamma^2 \Theta), \quad \lambda_K(\gamma_p Y_p Y_p^T) \geq 1/\kappa(\Theta^T \Gamma^2 \Theta), \quad \text{and} \quad \kappa(\gamma_p Y_p Y_p^T) \leq (\kappa(\Theta^T \Gamma^2 \Theta))^2.$$  

Proof. For DCMMSB-type models, we have $V = \Gamma \Theta \Gamma_p^{-1} V_p$, and $(V_p V_p^T)^{-1} = \Gamma_p^{-1} \Theta^T \Gamma^2 \Theta \Gamma_p^{-1}$ by Lemma 2.1 and Theorem 2.3. Note that $Y_p = N_p Z_p$, then we have

$$\lambda_1(\gamma_p Y_p Y_p^T) = \lambda_1(N_p Z_p Z_p^T N_p) = \lambda_1(N_p V_p \Gamma_p \Theta^T \Gamma^2 \Theta \Gamma_p^{-1} N_p)$$

$$\leq (\lambda_1(N_p \Gamma_p))^2 \lambda_1((\Theta^T \Gamma^2 \Theta)^{-1}) \leq (\max_{i \in I} \gamma_i / \|v_i\|)^2 / \lambda_K(\Theta^T \Gamma^2 \Theta)$$

$$\leq \lambda_1(\Theta^T \Gamma^2 \Theta) / \lambda_K(\Theta^T \Gamma^2 \Theta) = \kappa(\Theta^T \Gamma^2 \Theta) \quad \text{(by proof of Lemma H.1)}$$

Similarly, we have:

$$\lambda_K(\gamma_p Y_p Y_p^T) = \lambda_K(N_p \Gamma_p (\Theta^T \Gamma^2 \Theta)^{-1} \Gamma_p N_p) \geq (\lambda_K(N_p \Gamma_p))^2 \lambda_K((\Theta^T \Gamma^2 \Theta)^{-1})$$

$$\geq (\min_{i \in I} \gamma_i / \|v_i\|)^2 / \lambda_1(\Theta^T \Gamma^2 \Theta) \geq \lambda_K(\Theta^T \Gamma^2 \Theta) / (\psi \lambda_1(\Theta^T \Gamma^2 \Theta))$$

$$= 1 / (\kappa(\Theta^T \Gamma^2 \Theta)) \quad \text{(by proof of Lemma H.1)}$$

And finally we have,

$$\kappa(\gamma_p Y_p Y_p^T) \leq (\kappa(\Theta^T \Gamma^2 \Theta))^2.$$

\[\square\]

Lemma H.3. For DCMMSB-type models, let $v_i = V^T e_i$, $\hat{v}_i = \hat{V}^T e_i$, $y_i = V v_i / \|V v_i\|$, and $\hat{y}_i = \hat{V} \hat{v}_i / \|\hat{V} \hat{v}_i\|$, $i \in [n]$. Also let $\epsilon_0 = \max_i \|v_i - \hat{v}_i\|$, then,

$$\|y_i - \hat{y}_i\| \leq \frac{2 \epsilon_0}{\|V v_i\|} \leq \frac{2 \epsilon_0 \sqrt{\lambda_1(\Theta^T \Gamma^2 \Theta)}}{\gamma_i}.$$

Proof. From Lemma 2.2 we have

$$\|y_i - \hat{y}_i\| \leq \frac{2 \epsilon_0}{\|V v_i\|} = \frac{2 \epsilon_0 \sqrt{\lambda_1(\Theta^T \Gamma^2 \Theta)}}{\gamma_i},$$

where the last step uses Lemma H.1.

\[\square\]
Lemma H.4. For DCMMSE-type models, \( \lambda^*(P) \geq \rho \lambda^*(B) \lambda_K (\Theta^T \Gamma^2 \Theta) \).

Proof. Let \( X = B \Theta^T \Gamma^2 \Theta B \), it is easy to see that \( X \) is full rank and positive definite, then

\[
\begin{align*}
\lambda^*(P) &= \rho \lambda^*(\Gamma \Theta \Theta^T \Gamma) = \rho \sqrt{\lambda_K (\Gamma \Theta \Theta^T \Gamma)} \\
&= \rho \sqrt{\lambda_K (X^{1/2} \Theta \Theta^T \Theta X^{1/2})} = \rho \sqrt{\lambda_K (X \Theta^T \Theta)} \\
&\geq \rho \sqrt{\lambda_K (B)^2 (\lambda_K (\Theta^T \Gamma^2 \Theta))^2} = \rho \lambda^*(B) \lambda_K (\Theta^T \Gamma^2 \Theta),
\end{align*}
\]

where we use that \( LL^T \) and \( L^T L \) have the same leading \( K \) eigenvalues for a matrix \( L \in \mathbb{R}^{n \times K} \) with rank \( K < n \). \( \square \)

I DCMMSB error bounds

Lemma I.1. For DCMMSB-type models, if \( \theta_i \sim \text{Dirichlet}(\alpha) \), let \( \alpha_0 = 1_K \alpha, \alpha_{\text{max}} = \max_i \alpha_i, \alpha_{\text{min}} = \min_i \alpha_i \), \( \nu = \alpha_0/\alpha_{\text{min}} \), then

\[
P \left( \lambda_1 (\Theta^T \Gamma^2 \Theta) \leq \frac{3 \gamma_{\text{max}}^2 n (\alpha_{\text{max}} + \| \alpha \|^2)}{2 \alpha_0 (1 + \alpha_0)} \right) \geq 1 - K \exp \left( -\frac{n}{36 \nu^2 (1 + \alpha_0)^2} \right)
\]

\[
P \left( \lambda_K (\Theta^T \Gamma^2 \Theta) \geq \frac{\gamma_{\text{min}}^2 n}{2 \nu (1 + \alpha_0)} \right) \geq 1 - K \exp \left( -\frac{n}{36 \nu^2 (1 + \alpha_0)^2} \right)
\]

\[
P \left( \kappa(\Theta^T \Gamma^2 \Theta) \leq 3 \frac{\alpha_{\text{max}} + \| \alpha \|^2}{\alpha_{\text{min}}} \right) \geq 1 - 2K \exp \left( -\frac{n}{36 \nu^2 (1 + \alpha_0)^2} \right)
\]

\[
P \left( \lambda^*(P) \geq \frac{\gamma_{\text{min}}^2 \lambda^*(B) \nu n}{2 \nu n (1 + \alpha_0)} \right) \geq 1 - K \exp \left( -\frac{n}{36 \nu^2 (1 + \alpha_0)^2} \right)
\]

where \( \lambda^*(P) \) is the \( K \)-th singular value of \( P \).

Proof. First note that

\[
\lambda_1 (\Theta^T \Gamma^2 \Theta) = \lambda_1 (\Gamma \Theta \Theta^T \Gamma) \leq (\lambda_1 (\Gamma))^2 \lambda_1 (\Theta \Theta^T) = (\lambda_1 (\Gamma))^2 \lambda_1 (\Theta^T \Theta).
\]

Here we use that \( XX^T \) and \( X^T X \) have the same leading \( K \) eigenvalues for \( X \in \mathbb{R}^{n \times K} \) with rank \( K < n \). Also, as \( \Theta^T (\Gamma^2 - \gamma_{\text{min}}^2 I) \Theta \) is positive semidefinite, we have

\[
\lambda_K (\Theta^T \Gamma^2 \Theta) = \lambda_K (\Theta^T (\Gamma^2 - \gamma_{\text{min}}^2 I) \Theta + \gamma_{\text{min}}^2 \Theta^T \Theta) \geq \lambda_K (\Theta^T (\Gamma^2 - \gamma_{\text{min}}^2 I) \Theta) + \lambda_K (\gamma_{\text{min}}^2 \Theta^T \Theta) \geq \gamma_{\text{min}}^2 \lambda_K (\Theta^T \Theta)
\]

By Lemma A.2 of [5],

\[
P \left( \lambda_1 (\Theta^T \Theta) \leq \frac{3 n (\alpha_{\text{max}} + \| \alpha \|^2)}{2 \alpha_0 (1 + \alpha_0)} \right) \geq 1 - K \exp \left( -\frac{n}{36 \nu^2 (1 + \alpha_0)^2} \right)
\]

\[
P \left( \lambda_K (\Theta^T \Theta) \geq \frac{n}{2 \nu (1 + \alpha_0)} \right) \geq 1 - K \exp \left( -\frac{n}{36 \nu^2 (1 + \alpha_0)^2} \right)
\]

\[
P \left( \kappa(\Theta^T \Theta) \leq 3 \frac{\alpha_{\text{max}} + \| \alpha \|^2}{\alpha_{\text{min}}} \right) \geq 1 - 2K \exp \left( -\frac{n}{36 \nu^2 (1 + \alpha_0)^2} \right)
\]

So \( \kappa(\Theta^T \Gamma^2 \Theta) = \frac{\lambda_1 (\Theta^T \Gamma^2 \Theta)}{\lambda_K (\Theta^T \Gamma^2 \Theta)} \leq \frac{\gamma_{\text{max}}^2 \kappa(\Theta^T \Theta)}{\gamma_{\text{min}}^2} \leq 3 \frac{\gamma_{\text{max}}^2}{\gamma_{\text{min}}} \frac{\alpha_{\text{max}} + \| \alpha \|^2}{\alpha_{\text{min}}} \) with high probability. Using Lemma H.4, we have,

\[
\lambda^*(P) \geq \rho \lambda^*(B) \lambda_K (\Theta^T \Gamma^2 \Theta) \geq \frac{\gamma_{\text{min}}^2 \lambda^*(B) \nu n}{2 \nu n (1 + \alpha_0)} \rho n,
\]

with probability at least \( 1 - K \exp \left( -\frac{n}{36 \nu^2 (1 + \alpha_0)^2} \right) \). \( \square \)
Lemma I.2. For DCMMSB-type models, we have \((Y_pY_p^T)^{-1}1 \geq (\min_i\gamma_i)^2(1-P)^{-1}\). Furthermore, if \(\theta_i \sim \text{Dirichlet}(_\alpha)\), with probability larger than \(1 - 1/n^3 - K\ exp\left(-\frac{n}{6\nu^2(1+\alpha)}\right)\),

\[(Y_pY_p^T)^{-1}1 \geq (\min_i\gamma_i)^2(1-P)^{-1}\nu\frac{1}{\nu}\]

where \(\nu = \sum_{i=1}^n\alpha_i\).

Proof. First note that, for diagonal matrices \(D \in \mathbb{R}^{n \times n}\) and \(\Gamma \in \mathbb{R}^{n \times n}\) that have strictly positive elements on the diagonal, and some matrices \(G \in \mathbb{R}^{m \times m}\) and \(H_1 \in \mathbb{R}^{n \times m}, H_2 \in \mathbb{R}^{n \times m}\) have

\[GD1 \geq (\min_iD_{ii})^2G1,\]

\[(H_i^T\Gamma H_1 \geq \min_i\Gamma_iH_i^T H_1,\]

where last step follows that \(D, G\) and \((D - \min_iD_{ii}I)\) are all non-negative. Eq. (12) can be proved in a similar way. Now use these on Eq. (9), we have

\[(Y_pY_p^T)^{-1}1 = N_p^{-1}P^{-1}\theta^T\Gamma^2\Theta_{\min}^{-1}N_p^{-1}1 \geq \left(\min_i\|\theta_{\min}(i)\|\right)^2\theta^T\Gamma^2\Theta_{\min}1\]

where the last step follows Lemma [H.1.] By Lemma C.1. of [4], we know if rows of \(\Theta\) are from Dirichlet distribution with parameter \(\alpha = (\alpha_0, \alpha_2, \cdots, \alpha_K)\), \(\alpha_0 = \sum_i\alpha_i\), \(\nu = \alpha_0/\min_i\alpha_i\),

\[\Theta1 \geq \frac{n}{\nu} \left(1 - O_P(\sqrt{\nu \log n})\right)1\]

with probability larger than \(1 - 1/n^3\). Now by Lemma [I.1] we have, with probability larger than \(1 - 1/n^3 - K\ exp\left(-\frac{n}{6\nu^2(1+\alpha)}\right)\),

\[(Y_pY_p^T)^{-1}1 \geq \frac{\min_i\gamma_i^2}{\lambda_1(\Theta^T\Gamma^2\Theta)}\theta^T1 \geq \frac{\min_i\gamma_i^2}{\lambda_1(\Theta^T\Gamma^2\Theta)}\nu\frac{1}{\nu}\frac{\nu}{\nu}(1 - 1 - 1/n^3)\]

where \(\nu = \sum_{i=1}^n\alpha_i\\gamma_i\). further, \(\mu_2 \geq \frac{2\gamma_i^2(1 + \alpha)\nu}{3\gamma_i^2(\alpha_{\max} + \|\alpha\|^2)}\frac{1}{\nu}\frac{\nu}{\nu}(1 - 1 - 1/n^3)\]

We use a crucial result from [5] that shows row-wise eigenspace concentration for general low rank matrix.

Theorem I.3 (Row-wise eigenspace concentration [5]). Suppose \(P\) has rank \(K, \max_i\|P_{ij}\| \leq \rho\). Let \(A_{ij} = A_{ji} \sim \text{Ber}(P_{ij})\), \(V\) and \(\bar{V}\) are \(P\) and \(A\)’s top-\(K\) eigenvectors respectively. If \(P_{\max}\|V_{\max}\| > \sqrt{\rho}\) \(\leq \delta_1\), and for some constant \(\xi > 1\), \(\rho n = \Omega((\log n)^{2\xi})\) and \(P(\lambda^*(P) < 4\sqrt{\rho}(\log n)^{\xi}) < \delta_2\), then for a fixed \(i \in [n]\), with probability at least \(1 - \delta_1 - \delta_2 - O(Kn^{-3})\),

\[\|e_i^T(\bar{V}V^T - VV^T)\| = O\left(\frac{\min\{K, \kappa(P)\}\sqrt{Kn\rho}}{\lambda^*(P)}\right)\left(\frac{\min\{K, \kappa(P)\} + (\log n)^{\xi}}{\max_i\|V_i\|_\infty + (K + 1)n^{-2\xi}}\right).\]
We will use $\psi$ to denote a model specific number which is 1 for DCMSMSB and $\sqrt{K}$ for OCCAM, defined in Lemma H.1.

**Corollary I.4.** For DCMSMSB-type models with $\theta_i \sim$ Dirichlet($\alpha$), let $v_i = V^T e_i$, $\hat{v}_i = \hat{V}^T e_i$, $y_i = V v_i / \|V v_i\|$, and $\hat{y}_i = \hat{V} \hat{v}_i / \|\hat{V} \hat{v}_i\|$, $i \in [n]$. Also let $\epsilon_0 = \max_i \|v_i - \hat{v}_i\|$, then, if $\nu := \alpha_0 / \min\{n^{\frac{3}{12}}, n \rho\} \geq 2(1 + \alpha_0) / \nu$, then, with probability at least $\nu$, $\epsilon_0 \leq \min\{\sqrt{\frac{\nu}{2}} + \epsilon, \sqrt{\frac{\nu}{2}}\}$ for some constant $\xi > 1$.

**Proof.** First by Lemma H.3,

$$\epsilon = \max_i \|y_i - \hat{y}_i\| = \tilde{O}\left(\frac{\psi_{\max} \min\{K^2, (\kappa(\mathbf{P}))^2\} K^{0.5} \nu (1 + \alpha_0)}{\gamma_3 \min \lambda^* (\mathbf{B}) \sqrt{n \rho}}\right)$$

with probability at least $1 - O(Kn^{-3})$.

**Proof of Theorem 3.2.** Note that $\mathbf{P} = \rho \Gamma \Theta \Theta^T \Gamma = \mathbf{V} \mathbf{E} \mathbf{V}^T$, we have $\rho \Gamma_p \beta \Gamma_p = \mathbf{V} \mathbf{E} \mathbf{V}^T$, then $\rho \mathbf{N} \mathbf{P} \Gamma_p \beta \Gamma_p \mathbf{N} \mathbf{P} = \mathbf{N} \mathbf{P} \mathbf{V} \mathbf{E} \mathbf{V}^T \mathbf{Y} \mathbf{P} \mathbf{E} \mathbf{V}^T \mathbf{Y}^T \mathbf{P}$. As $\mathbf{B}$ has unit diagonal, let $B(i,i) = c^2$, then $c^2 \rho_\gamma^2_{(i,i)} / \|v_{(i,i)}\|^2 = c^2 e_i^T \mathbf{N} \mathbf{P} \Gamma_p \beta \Gamma_p \mathbf{N} \mathbf{P} e_i = c^2 e_i^T \mathbf{Y} \mathbf{P} \mathbf{E} \mathbf{V}^T \mathbf{Y} \mathbf{P} e_i = d_i^2$. Since our estimation for $c^2 \rho_\gamma^2_{(i,i)} / \|v_{(i,i)}\|^2$ is $c^2 \Gamma_p^T \hat{Y} \hat{C} \hat{V} \hat{E} \hat{V}^T \hat{Y} \hat{C} \hat{P} e_i$, and note that $\|e_i\| \leq \max_i \|e_i^T \mathbf{P}\|_1 = O(\rho n)$,
we have,
\[ \| \hat{E} \| \leq \| E \| + \| A - P \| = O(p) \] using Weyl’s inequality and Theorem 5.2 of [3], and \( \| VEV^T - V\hat{E}V^T \| \leq \lambda_{K+1}(A) + \| P - A \| \leq 2 \| P - A \| = O(\sqrt{p}) \). Let \( \hat{d}_i^2 = e_i^T \hat{Y}_C VEV^T \hat{Y}_C e_i \), then we have,
\[
\begin{align*}
| \hat{d}_i - \hat{d}_2(i) | &= | e_i^T (Y_P \Pi^T \hat{Y}_C VEV^T \hat{Y}_C e_i - e_i^T \Pi^T \hat{Y}_C \hat{V}EV^T \hat{Y}_C e_i) | \\
&\leq | e_i^T (Y_P \Pi^T \hat{Y}_C) VEV^T \hat{Y}_C e_i | + | e_i^T \Pi^T \hat{Y}_C (VEV^T - \hat{V}EV^T) \hat{Y}_C e_i | \\
&+ | e_i^T \Pi^T \hat{Y}_C VEV^T (Y_P - \hat{Y}_C \Pi) e_i | \\
&\leq | e_i^T (Y_P \Pi^T \hat{Y}_C) \| E \| + \| VEV^T - \hat{V}EV^T \| + \| \hat{E} \| \| e_i^T (Y_P - \Pi^T \hat{Y}_C) \| \\
&\leq O(\sqrt{p}) \epsilon_4 \sqrt{K} + O(\sqrt{p}).
\end{align*}
\]

Using Lemma [H.1] \( c \sqrt{\rho \lambda_K(\Theta^T \Theta)} \leq \hat{d}_i \leq c \sqrt{\rho \lambda_1(\Theta^T \Theta)} \), and by Lemma [I.1]
\[ \lambda_1(\Theta^T \Theta) \leq \frac{3\gamma_{max} n(\sigma_{max} + \| \alpha \|^2)}{2\sigma(1 + \alpha)}, \]
and \( \hat{d}_i \leq c \sqrt{\frac{3\gamma_{max} n(\sigma_{max} + \| \alpha \|^2)}{2\sigma(1 + \alpha)}} \) with probability at least \( 1 - 2K \exp \left( -\frac{n}{36\sigma^2(1 + \alpha)^2} \right) \). Then, using Lemma [H.2]
\[
\begin{align*}
| \hat{d}_i - \hat{d}_2(i) | &\leq \frac{O(\sqrt{p}) \epsilon_4 \sqrt{K} + O(\sqrt{p})}{\min(\hat{d}_j + \hat{d}_2(i))} \\
&\leq \frac{O(\sqrt{p}) \epsilon_4 \sqrt{K} (\sqrt{\frac{\gamma_i}{\lambda_K(\Theta^T \Theta)}})^{1.5} + O(\sqrt{p})}{\sqrt{\rho \lambda_K(\Theta^T \Theta)}} \\
&= O \left( \frac{K^{1.5}(\lambda(\Theta^T \Theta))^{1.5} \sqrt{p}}{\eta \sqrt{\lambda_K(\Theta^T \Theta)}} \right).
\end{align*}
\]

Let \( D = \text{diag}(d_1, d_2, \ldots, d_K) \) and \( \hat{D} = \text{diag}(\hat{d}_1, \hat{d}_2, \ldots, \hat{d}_K) \), then \( D = c \sqrt{p} \hat{D} \). Now as we estimate \( c \sqrt{p} \hat{D} \) by \( \hat{D} \), we have
\[
\begin{align*}
\| e_i^T (c \sqrt{p} \Theta - \hat{c} \sqrt{p} \hat{D} \hat{\Pi}) \| &= \| e_i^T (MD - \hat{M} \hat{D} \hat{\Pi}) \| \\
&\leq \| e_i^T (M - \hat{M}) \| \| D \| + \| e_i^T \hat{M} \| \| D - \Pi^T \hat{D} \Pi \| \\
&\leq \epsilon_M \max_j | d_j | + (| \| e_i^T \hat{M} \| + \epsilon_M) \max_j | d_j - \hat{d}_2(i) | \\
&\leq \frac{\gamma_i}{\lambda_K(\Theta^T \Theta)} + \frac{\epsilon}{\lambda_K(\Theta^T \Theta)} \left( \frac{K^{1.5}(\lambda(\Theta^T \Theta))^{1.5} \sqrt{p}}{\sqrt{\lambda}(\Theta^T \Theta)} \right).
\end{align*}
\]
where we use \( | \| e_i^T \hat{M} \| | D - \Pi^T \hat{D} \Pi \| \leq \gamma_i \| \hat{\Pi} \| \| v_j \| / \| \hat{\Pi} \| \) and \( \| \hat{\Pi} \| \leq 1 \) for DCMMSB and OCCAM for the second last inequality. As
\[
\begin{align*}
\epsilon_M &= \frac{c_M \sqrt{K(\gamma_i \| Y_P Y_P^T \|)^2} \epsilon}{(\lambda_K(\Theta^T \Theta))^{2.5}} \\
&\leq \frac{\gamma_i \| e_i^T Z \| (\lambda_K(\Theta^T \Theta))^{6.5} K^2}{\lambda(\Theta^T \Theta)} \epsilon \\
&\leq \frac{c_i \| e_i^T Z \| (\lambda_K(\Theta^T \Theta))^{6.5} K^2}{\eta} \epsilon.
\end{align*}
\]
Then
\[
\begin{align*}
\epsilon_5 &= \| e_i^T (c \sqrt{p} \Theta - \hat{c} \sqrt{p} \hat{D} \hat{\Pi}) \| \\
&= c \sqrt{\rho \lambda_1(\Theta^T \Theta)} \epsilon_M \\
&= c \sqrt{\rho \lambda_1(\Theta^T \Theta)} \frac{\gamma_i}{\lambda_K(\Theta^T \Theta)} + \frac{\epsilon}{\lambda_K(\Theta^T \Theta)} \left( \frac{K^{1.5}(\lambda(\Theta^T \Theta))^{1.5} \sqrt{p}}{\sqrt{\lambda}(\Theta^T \Theta)} \right).
\end{align*}
\]
As $|c\sqrt{p_{\gamma_i}} - \hat{c}\sqrt{\hat{p}_{\gamma_i}}| = \|e_i^T(c\sqrt{p}\Theta - \hat{c}\sqrt{\hat{p}}\hat{\Theta}\Pi)\| \leq \psi\sqrt{K}\epsilon_5$, let $X_i = e_i^Tc\sqrt{p}\Theta$ and $\hat{X}_i = e_i^T\hat{c}\sqrt{\hat{p}}\hat{\Theta}\Pi$, then as $\|\hat{X}_i\| \leq \hat{c}\sqrt{\hat{p}_{\gamma_i}}$,

$$\|e_i^T(\Theta - \hat{\Theta}\Pi)\| = \left\|\frac{X_i}{c\sqrt{p}_{\gamma_i}} - \frac{\hat{X}_i}{\hat{c}\sqrt{\hat{p}_{\gamma_i}}}\right\| \leq \frac{\|\hat{X}_i\|}{c\sqrt{p}_{\gamma_i}} + \frac{\|\hat{X}_i\|}{\hat{c}\sqrt{\hat{p}_{\gamma_i}}} = \frac{\psi\sqrt{K}\epsilon_5}{c\sqrt{p}_{\gamma_i}} + \frac{\psi\sqrt{K}\epsilon_5}{\hat{c}\sqrt{\hat{p}_{\gamma_i}}} = O \left(\frac{\psi\sqrt{K}\epsilon_5}{\gamma_{\min}\sqrt{\hat{p}}}\right)$$

$$= O \left(\max \left\{\psi K^{0.5}(\kappa(\Theta^T\Gamma^2\Theta))^{3.5}, \frac{n}{\lambda_K(\Theta^T\Gamma^2\Theta)}\right\} \frac{\psi\gamma_{\max}K^{1.5}(\kappa(\Theta^T\Gamma^2\Theta))^{3.5}}{\gamma_{\min}\eta}\right)$$

$$= O \left(\max \left\{\psi K^{0.5}(\kappa(\Theta^T\Gamma^2\Theta))^{3.5}, \frac{n}{\lambda_K(\Theta^T\Gamma^2\Theta)}\right\} \frac{\psi^{2\gamma_{\max}K^{2}\min\{K^2,(\kappa(P))^2\}n^{3/2}}{\gamma_{\min}\eta\lambda^*(B)\lambda_K(\Theta^T\Gamma^2\Theta)\sqrt{\hat{p}}}ight)$$

$$= O \left(\frac{\psi^{2\gamma_{\max}K^{2}\min\{K^2,(\kappa(P))^2\}n^{3/2}}{\gamma_{\min}\eta\lambda^*(B)\lambda_K(\Theta^T\Gamma^2\Theta)\sqrt{\hat{p}}}ight).$$

Note that this bound works for both DCMMSB and OCCAM, and $\lambda_K(\Theta^T\Gamma^2\Theta) = \Omega(n)$, so the bound is about $O(1/\sqrt{\hat{p}n})$. Specifically, for DCMMSB,

$$\|e_i^T(\Theta - \Phi\Pi)\| = O \left(\frac{\psi^{2\gamma_{\max}K^{2}\min\{K^2,(\kappa(P))^2\}n^{3/2}}{\gamma_{\min}\eta\lambda^*(B)\lambda_K(\Theta^T\Gamma^2\Theta)\sqrt{\hat{p}}}ight)$$

$$= O \left(\frac{\psi^{2\gamma_{\max}K^{2}\min\{K^2,(\kappa(P))^2\}n^{3/2}}{\gamma_{\min}\eta\lambda^*(B)\lambda_K(\Theta^T\Gamma^2\Theta)\sqrt{\hat{p}}}ight).$$

\[\Box\]

## J Topic model error bounds

### J.1 Eigenspace concentration for topic models

Consider the following setup similar to [11].

$$A_{ij} \sim \text{Binomial}(N, \lambda_{ij}) \quad \text{For } i \in [V], j \in [D] \tag{13}$$

Here $A$ is the probability matrix for words appearing in documents. Furthermore, we have $A = TH$, where $T$ is the word to topic probabilities with columns summing to 1 and $H$ is the topic to document matrix with columns summing to 1. Also note that, $\sum_i \|\hat{e}_i^TA\hat{A}^T\|_1 = D$, since the columns of $A$ sum to one. We will construct a matrix $A_1A_2^T$, where $A_1$ and $A_2$ are obtained by dividing the words in each document uniformly randomly in two equal parts. For simplicity denote $N_1 = N/2$.

Consider the matrix $U = \frac{A_1A_2^T}{N_1}$. We have $E[U] = A\hat{A}^T$.

**Lemma J.1.** For topic models, we have $(Y_P^TY_P^{-1}1 \geq \frac{\min_i \|e_i^T\|}{\lambda_1(T^TT)} \geq \frac{\min_i \|e_i^T\|}{K}1$, where $T$ is the word-topic probability matrix.

**Proof.** Noting that $T = \Phi\Pi$ for topic models, and following the steps of Lemma I.2 we find

$$(Y_P^TY_P^{-1}1 \geq \frac{\min_i \gamma_i}{\lambda_1(\Theta^T\Gamma^2\Theta)}1 = \frac{\min_i \gamma_i}{\gamma_1^*(\Gamma^2\Theta)^T}1 = \frac{\min_i \gamma_i}{\gamma_1^*(T^TT)}1 \geq \frac{\min_i \|e_i^T\|}{K}1,$$

where the last step is true because $\lambda_1(T^TT) \leq \text{trace}(T^TT) = \sum_i \|Te_i\|^2 \leq K$. So we can choose $\eta = \frac{\min_i \|e_i^T\|}{K}$.

\[\Box\]

**Lemma J.2.** Using Eq I.3, we see that under Assumption I.1

$$P \left( \|R\| \leq \frac{18D \log \max(V, D)}{N_1} \right) \geq 1 - \frac{1}{\max(V, D)}.$$
Proof. Let \( R := U - \mathcal{A} \mathcal{A}^T \).

\[
R_{ik} = \frac{\sum_{j=1}^{D} A_1(ij) A_2(kj)}{N_1^2} - g_{ik}
\]

Note that \( E[R_{ij}] = 0 \), and \( A_1(ij) A_2(kj)/N_1^2 \) is bounded by 1. Also \( A_1(i, j) \) and \( A_2(i, j) \) are independent. For independent \( X := A_1(ij)/N_1, Y := A_1(kj)/N_1 \),

\[
\text{var}(XY) = \text{var}(X)\text{var}(Y) + \text{var}(X)E[X]^2 + \text{var}(Y)E[Y]^2 \leq \frac{3A_1(ij)A_2(kj)}{N_1}
\]

\[
\text{var}(R_{ik}) \leq 3g_{ik}/N_1
\]

When \( g_{ik} = 0, U_{ik} = 0 \). When \( g_{ik} > 0 \), using Bernstein’s inequality, we have:

\[
P \left( |R_{ik}| \geq t_{ik} \right) \leq 2 \exp \left( -\frac{t_{ik}^2}{2(3g_{ik}/N_1 + t_{ik}/3)} \right),
\]

Setting, \( t_{ik} = \sqrt{24 \log \max(V, D)g_{ik}/N_1} \), we see that,

\[
\sum_{i,k} t_{ik}^2 = 24 \log \max(V, D) \sum_{i,k} g_{ik}/N_1 = 24 \log \max(V, D)D/N_1
\]

Then,

\[
P \left( ||R||_F \geq \sum_{i,k} t_{ik}^2 \right) \leq V^2 \max_{i,k} P \left( |R_{ik}| \geq t_{ik} \right) \leq V^2 / \max(V, D)^3 \leq 1 / \max(V, D),
\]

This yields the result. \( \square \)

**Lemma J.3.** Using Eq (13), we see that, under Assumption J.1, there exists constants \( C, r \) such that,

\[
P \left( \|U - \mathcal{A} \mathcal{A}^T\| \geq C_r \sqrt{D \log D/N} \right) \leq 1/D^r.
\]

**Proof.** We use the Matrix Bernstein bound in [8]. Let \( S_k := \frac{A_{1k}A_{2k}^T}{N_1^2} - \mathcal{A}_k \mathcal{A}_k^T \), where \( M_k \) is the \( k \)th column of matrix \( M \). Note that \( E[S_k] \) is the \( V \times V \) zeros matrix. We also see that by symmetry of the random splitting, \( E[S_kS_k^T] = E[S_k^T S_k] \).

We will now note some theoretical properties of the \( S_k \) matrices. Let \( X \) be a vector of size \( V \), such that, \( X_i \sim \text{Binomial}(N_1, a_i) \).

\[
\frac{E[X^T X]}{N_1^2} = \sum_{i=1}^{V} \frac{E[X_i^2]}{N_1^2} = \sum_{i=1}^{V} \frac{E[X_i]^2 + \text{var}(X_i)}{N_1^2} = \sum_{i=1}^{V} \frac{N_1^2 a_i^2 + N_1 a_i (1 - a_i)}{N_1^2} = \left( 1 - \frac{1}{N_1} \right) ||a||_2^2 + \frac{1}{N_1}
\]

(14)

Furthermore, let

\[
\text{Cov}(X) = \Sigma, \quad \Sigma_{ij} = N_1 a_i (1 - a_i) 1(i = j)
\]

(15)

Then,

\[
E[S_kS_k^T] = E \left[ \frac{A_{1k}A_{2k}^T}{N_1^2} - \mathcal{A}_k \mathcal{A}_k^T \right]
\]

(By independence)

\[
= E \left[ \frac{A_{2k}^T A_{1k}}{N_1^2} \right] - \mathcal{A}_k \mathcal{A}_k^T
\]

(By Eq (14) and (15))

\[
= \left( \frac{1}{N_1} + \mathcal{A}_k^2 (1 - \frac{1}{N_1}) \right) \left( \frac{\Sigma_k}{N_1^2} + \mathcal{A}_k \mathcal{A}_k^T \right) - ||\mathcal{A}_k||_2^2 \mathcal{A}_k \mathcal{A}_k^T
\]

\[
= ||\mathcal{A}_k||_2^2 \left( \frac{\Sigma_k}{N_1^2} + \frac{1 - ||\mathcal{A}_k||_2^2}{N_1} \right) \left( \frac{\Sigma_k}{N_1^2} + \mathcal{A}_k \mathcal{A}_k^T \right)
\]

\[
= \mathcal{A}_k \mathcal{A}_k^T
\]

(15)
Since \( \| \Sigma_k \| \leq N_1 \| A_k \|_1 = N_1, \| A \|_F^2 \leq D \),
\[
v(S) = \left\| \sum_k E[S_k S_k^T] \right\| \leq 2 \frac{\| A \|_F^2}{N_1} + \frac{D}{N_1^2} \leq \frac{D}{N_1} \left( 2 + \frac{1}{N_1} \right).
\]
Furthermore,
\[
\| S_k \| \leq \| A_k \|^2 + \frac{\| A_{1k} \| \| A_{2k} \|}{N_1^2} \leq 2 =: L
\]
So the Matrix Bernstein bound gives us:
\[
P \left( \left\| \sum_k S_k \right\| \geq t \right) \leq 2V \exp \left( -\frac{t^2/2}{v(S) + Lt/3} \right) = 2V \exp \left( -\frac{t^2/2}{3D/N_1 + 2t/3} \right)
\]
Using \( t = C_r \sqrt{D \log D/N} \), and using the condition in Assumption 3.1 we get the bound.

**Proof of Lemma 3.3.** First note the proof is under Assumption 3.1. Let \( R = U - AA^T \). Using the Davis Kahan lemma [9], we see that:
\[
\| \hat{V}O - V \|_F \leq \frac{c(2\lambda_1(\hat{A}A^T) + \| R \|_2) \min(\sqrt{K} \| R \|_2, \| R \|_F)}{\lambda_K^2(\hat{A}A^T)}
\]
where \( \lambda_1 \) and \( \lambda_K \) are the largest and \( K^{th} \) largest singular values (and also eigenvalue) of \( \hat{A}A^T \) respectively. Thus,
\[
\| \hat{V}O - V \|_F \leq \frac{\sqrt{8}(2\lambda_1(\hat{A}A^T) + \| R \|_2) \min(\sqrt{K} \| R \|_2, \| R \|_F)}{\lambda_K^2(\hat{A}A^T)}
\]
\[
\leq \frac{\sqrt{8}2\lambda_1(\hat{A}A^T) + \sqrt{D \log D/N}}{\lambda_K^2(\hat{A}A^T)^2} \max \left( C_r \sqrt{K} \frac{D \log D}{N}, \sqrt{KD \log \max(V, D)} \right)
\]
\[
\leq \frac{\lambda_1(\hat{HH}^T)\lambda_1(T^T T)}{\lambda_K^2(\hat{HH}^T)^2} O_P \left( \sqrt{K \log \max(V, D)} \right)
\]
\[
= \frac{\kappa(\hat{HH}^T)\lambda_1(T^T T)}{\lambda_K^2} O_P \left( \sqrt{K \log \max(V, D)} \right)
\]
where the third inequality follows Lemma H.4 with \( P = \hat{A}A^T, \Gamma = T, B = HH^T \) and \( \rho = 1 \). Now we bound \( \| e_i^T (\hat{VV}^T - VV^T) \| \) as:
\[
\| e_i^T (\hat{VV}^T - VV^T) \| \leq \| \hat{VV}^T - VV^T \|_2 \leq \| (\hat{V}O - V)V^T + V(\hat{V}O - V)^T \|
\]
\[
= \frac{\kappa(\hat{HH}^T)\lambda_1(T^T T)}{\lambda_K^2} O_P \left( \sqrt{K \log \max(V, D)} \right)
\]

**J.2 Parameter estimation for topic models**

**Proof of Theorem 3.4.** For topic models, \( M = TD \), where \( T = \Gamma \Theta, D = (N_p \Gamma_p)^{-1} \), \( \gamma_i = \Gamma \Theta \Gamma_i = \| e_i^T T \|_1 \). For empirical estimation we have \( \hat{M} = \hat{T}D \), where \( \hat{D}(i, i) = \| \hat{M}(i, i) \|_1 \). First we have \( \forall i \in K, \| T(:, i) \|_1 = 1 \), then \( \| \hat{M}(i, :) \|_1 = \| D(i, i) = \| V_{\hat{I}(i)} \| / \gamma_{I(i)} \), let \( \pi \) be the permutation
function for permutation matrix $\Pi$ in Theorem 2.3 then,

$$|D(i, i) - \tilde{D}(\pi(i), \pi(i))| = \|M(:, i)\|_1 - \|\tilde{M}(\cdot, \pi(i))\|_1 \leq \|M(:, i) - \tilde{M}(\cdot, \pi(i))\|_1$$

$$= \sum_{j=1}^{V} |M(j, i) - \tilde{M}(j, \pi(i))| \leq \sum_{j=1}^{V} \|M(j, :) - \tilde{M}(j, :)\Pi\|_1$$

$$= \sum_{j=1}^{V} \|e_{j}^T T\|_1 \frac{\|M(j, :) - \tilde{M}(j, :)\Pi\|_1}{\|e_{j}^T T\|_1}$$

$$\leq K_{\max} \frac{\|M(j, :) - \tilde{M}(j, :)\Pi\|_1}{\|e_{j}^T T\|_1} \leq K_{\max} \frac{\|M(j, :) - \tilde{M}(j, :)\Pi\|_1}{\|e_{j}^T T\|_1}$$

$$\leq \frac{K_{1.5} \epsilon_M}{\min_j \|e_{j}^T T\|_1} = \epsilon_D$$

Note that $T^T T = \Theta^T T^2 \Theta$, and from Lemma H.1 we know

$$1/\sqrt{\lambda_1(T^T T)} \leq \|V_i\|/\gamma_i \leq 1/\sqrt{\lambda_K(T^T T)}, \forall i \in [n]$$

Using Lemma H.2 we have $\lambda_1(Y_p Y_p^T) \leq \kappa(\Theta^2 \Theta) = \kappa(T^T T)T$, $\lambda_K(Y_p Y_p^T) \geq 1/\kappa(\Theta \Theta) = 1/\kappa(T^T T)$, and $\kappa(Y_p Y_p^T) \leq (\kappa(\Theta \Theta))^2 = (\kappa(T^T T))^2$.

Then the error for each row of $T$ is

$$\|e_i^T (T - \Pi \Pi^T)\| = \|e_i^T (MD^{-1} - MD^{-1} \Pi^T)\|$$

$$\leq \|e_i^T (\tilde{M} - M \Pi \Pi^T)\| \|D(\cdot, j)\| + \|e_i^T M \Pi \Pi^T\| \|D(\cdot, j)\| \frac{\|D(\cdot, j) - \tilde{D}(\pi(\cdot, \pi(\cdot))\|}{\|D(\cdot, j)\|}$$

$$\leq \frac{2 \epsilon_M}{\min_j \|D(j, j)\|} + \frac{2 \epsilon_D}{\min_j \|D(j, j)\|} \|e_i^T M\|$$

$$\leq \frac{2 \epsilon_M}{\min_j \|V_i(j)\|/\gamma_i(j)} + \frac{2 \epsilon_D}{\min_j \|V_i(j)\|/\gamma_i(j)} \min_j \|D(j, j)\| \|e_i^T T\|$$

$$\leq 2 \sqrt{\frac{\lambda_1(T^T T) e_M}{\kappa(T^T T)}} + 2 \sqrt{\frac{\lambda_1(T^T T)}{\lambda_K(T^T T)}}\frac{\sqrt{\lambda_1(T^T T)}}{\sqrt{\lambda_K(T^T T)}} \|e_i^T T\| \frac{K_{1.5} \epsilon_M}{\min_j \|e_i^T T\|_1}$$

$$\leq 4 \sqrt{\lambda_1(T^T T)} \sqrt{\kappa(T^T T)} \|e_i^T T\| \frac{K_{1.5} \epsilon_M}{\min_j \|e_i^T T\|_1} \frac{\kappa(\gamma_p \gamma_p^T) e_i^T Z_i}{(\kappa(\gamma_p \gamma_p^T))^2}$$

$$\leq c_1 \sqrt{\lambda_1(T^T T)} (\kappa(T^T T))^{5.5} \|e_i^T T\| \frac{K_{3.5} \max_j \|e_i^T T\|_1 \|e_i^T T\|_1}{\eta \min_j \|e_i^T T\|_1}$$

$$\leq c_2 \sqrt{\lambda_1(T^T T)} (\kappa(T^T T))^{7} K_{\gamma} \max_j \|e_i^T T\|_1 \|e_i^T T\|_1$$

where we use $\epsilon_D \leq \|D(i, j)\|/2$ for relaxation in the 3rd inequality and $c_1$ and $c_2$ are some constants.

Under Assumption 3.1 by Lemma 3.3 we have

$$\epsilon_0 = \|e_i^T (V V^T - V V^T)\|_F = \frac{\kappa(H H^T) \lambda_1(T^T T)}{\lambda_K(T^T T)} \|D(\cdot, j)\| \frac{\sqrt{K \log \max(V, D)}}{D N}$$

By Lemma H.3 $\|y_i - \tilde{y}_i\| \leq 2 \epsilon_0 \|V_i\| \leq 2 \epsilon_0 \|V_i\|_{11}$. So,

$$\epsilon = \max_i \|y_i - \tilde{y}_i\| = \frac{\kappa(H H^T) (\lambda_1(T^T T))^{1.5}}{\min_j \|e_j^T T\|_1 \lambda_K(T^T T)} \|D(\cdot, j)\| \frac{\sqrt{K \log \max(V, D)}}{D N}$$
Then,
\[
\frac{\|e^T_i (T - TH^T)\|}{\|e^T_i T\|} \leq \frac{\sqrt{\lambda_1(T^T T)} (\kappa(T^T T))^7 K^{3.5}}{\eta} \max_j \frac{\|e^T_j T\|_1}{\min_j \|e^T_j T\|_1} \\
= \frac{\sqrt{\lambda_1(T^T T)} (\kappa(T^T T))^7 K^{3.5}}{\eta} \max_j \frac{\|e^T_j T\|_1}{\min_j \|e^T_j T\|_1} \min_j \|e^T_j T\|_1 \lambda_{K}^2 (T^T T) \Omega \left( \sqrt{\frac{K \log \max(V, D)}{DN}} \right)
\]
\[
= \max_j \frac{\|e^T_j T\|_1}{(\min_j \|e^T_j T\|_1)^2} \frac{\kappa(HH^T)(\kappa(T^T T))^9}{\eta} O_P \left( K^4 \sqrt{\frac{\log \max(V, D)}{DN}} \right)
\]
\[
= O_P \left( \frac{K^4 \max_j \|e^T_j T\|_1}{\eta(\min_j \|e^T_j T\|_1)^2} \sqrt{\frac{\log \max(V, D)}{DN}} \right)
\]
\[
(\text{if } \kappa(T^T T) = \Theta(1) \text{ and } \kappa(HH^T) = \Theta(1))
\]
M Network statistics for DBLP datasets

Table 1: Network statistics

(a) Author-author DBLP Graphs.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>DBLP1</th>
<th>DBLP2</th>
<th>DBLP3</th>
<th>DBLP4</th>
<th>DBLP5</th>
</tr>
</thead>
<tbody>
<tr>
<td># nodes</td>
<td>30,566</td>
<td>16,817</td>
<td>13,315</td>
<td>25,481</td>
<td>42,351</td>
</tr>
<tr>
<td># communities</td>
<td>6</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Average Degree</td>
<td>8.9</td>
<td>7.6</td>
<td>8.5</td>
<td>5.2</td>
<td>6.8</td>
</tr>
<tr>
<td>Overlap %</td>
<td>18.2</td>
<td>14.9</td>
<td>21.1</td>
<td>14.4</td>
<td>18.5</td>
</tr>
</tbody>
</table>

(b) Bipartite Graphs.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>DBLP1</th>
<th>DBLP2</th>
<th>DBLP3</th>
<th>DBLP4</th>
<th>DBLP5</th>
</tr>
</thead>
<tbody>
<tr>
<td># nodes</td>
<td>103,660</td>
<td>50,699</td>
<td>42,288</td>
<td>53,369</td>
<td>81,245</td>
</tr>
<tr>
<td># communities</td>
<td>12</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>Average Degree</td>
<td>3.4</td>
<td>3.4</td>
<td>3.6</td>
<td>2.6</td>
<td>3.0</td>
</tr>
<tr>
<td>Overlap %</td>
<td>6.3</td>
<td>5.6</td>
<td>5.7</td>
<td>6.9</td>
<td>9.7</td>
</tr>
</tbody>
</table>

N DBLP bipartite author-paper networks

Figure 1: DBLP coauthorship wall-clock time

Figure 2: The wall-clock time of the competing methods respectively on the bipartite author-paper DBLP network. BSNMF was out of memory for DBLP1 and DBLP5.
Statistics of topic modeling datasets

Table 2: Statistics of topic modeling datasets

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Vocabulary size $V$</th>
<th>Number of documents $D$</th>
<th>Total number of words</th>
</tr>
</thead>
<tbody>
<tr>
<td>NIPS$^1$</td>
<td>5002</td>
<td>1,491</td>
<td>1,589,280</td>
</tr>
<tr>
<td>NYTimes$^1$</td>
<td>5004</td>
<td>296,784</td>
<td>68,876,786</td>
</tr>
<tr>
<td>PubMed$^1$</td>
<td>5001</td>
<td>7,829,043</td>
<td>485,719,597</td>
</tr>
<tr>
<td>20NG$^2$</td>
<td>5000</td>
<td>9,540</td>
<td>886,043</td>
</tr>
<tr>
<td>Enron$^1$</td>
<td>5003</td>
<td>29,823</td>
<td>4,963,162</td>
</tr>
<tr>
<td>KOS$^1$</td>
<td>5001</td>
<td>3,412</td>
<td>405,190</td>
</tr>
</tbody>
</table>

Topics in Real Data

Table 3: Top-10 word of 5 topics for different topic modeling datasets

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Top-10 words</th>
</tr>
</thead>
<tbody>
<tr>
<td>NIPS</td>
<td>algorithm data problem method parameter point vector distribution error space</td>
</tr>
<tr>
<td></td>
<td>neuron output pattern signal circuit visual synaptic unit layer current</td>
</tr>
<tr>
<td></td>
<td>data unit training output image information object recognition pattern point</td>
</tr>
<tr>
<td></td>
<td>unit hidden output layer weight object pattern visual representation connection</td>
</tr>
<tr>
<td></td>
<td>error algorithm training weight data parameter method problem classifier</td>
</tr>
<tr>
<td>NYT</td>
<td>con son solo era mayor zzz_mexico director sin fax sector</td>
</tr>
<tr>
<td></td>
<td>zzz_bush government school campaign show american member country zzz_united_states law</td>
</tr>
<tr>
<td></td>
<td>company companies market stock business billion plan money analyst government</td>
</tr>
<tr>
<td></td>
<td>team game season play player games run coach win won</td>
</tr>
<tr>
<td></td>
<td>file sport zzz_los_angeles notebook internet zzz_calif read output web computer</td>
</tr>
<tr>
<td>PubMed</td>
<td>receptor expression gene binding system function region genes dna mechanism</td>
</tr>
<tr>
<td></td>
<td>concentration strain gene dna system expression region genes test function</td>
</tr>
<tr>
<td></td>
<td>tumor gene expression disease genes lesion mutation region dna clinical</td>
</tr>
<tr>
<td></td>
<td>rat concentration plasma day serum animal liver drug response administration</td>
</tr>
<tr>
<td></td>
<td>children disease clinical year test therapy women system diagnosis drug</td>
</tr>
<tr>
<td>20NG</td>
<td>key government car chip state including information cs number long</td>
</tr>
<tr>
<td></td>
<td>god jesus bible question things tite christian world chirst true</td>
</tr>
<tr>
<td></td>
<td>year michael game team cs games win play including car</td>
</tr>
<tr>
<td></td>
<td>drive mb scsi windows card hard disk dos computer drives</td>
</tr>
<tr>
<td></td>
<td>windows window dos file files program card fax run win</td>
</tr>
<tr>
<td>Enron</td>
<td>report status changed payment approved approval amount paid due expense</td>
</tr>
<tr>
<td></td>
<td>database error operation perform hourahead data file process start message</td>
</tr>
<tr>
<td></td>
<td>power california customer gas order deal list office forward comment</td>
</tr>
<tr>
<td></td>
<td>message contract corp receive offer free send list received click</td>
</tr>
<tr>
<td></td>
<td>hourahead final file hour data price process error detected variances</td>
</tr>
<tr>
<td>KOS</td>
<td>iraq administration military iraqi president american troops busses officials soldiers</td>
</tr>
<tr>
<td></td>
<td>voting vote senate polls governor electoral voter media voters primary</td>
</tr>
<tr>
<td></td>
<td>percent senate race elections republican party state voters campaign polls</td>
</tr>
<tr>
<td></td>
<td>senate polls governor electoral primary vote ground races voter contact</td>
</tr>
<tr>
<td></td>
<td>dean edwards primary clark gephardt lieberman iowa results polls kucinich</td>
</tr>
</tbody>
</table>

https://archive.ics.uci.edu/ml/datasets/Bag+of+Words
http://qwone.com/~jason/20Newsgroups/
References


