MULTIVARIATE GAUSSIANS AND MAXIMUM LIKELIHOOD DEGREE

FRANK WANG

1. Foundations and Background

Let k be a positive integer. For a fixed $\mu \in \mathbb{R}^k$ and $\Sigma \in \mathbb{S}^k_+$, the multivariate normal distribution $\mathcal{N}(\mu, \Sigma)$ of a k-dimensional vector $\mathbf{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_k \end{bmatrix} \in \mathbb{R}^k$ has expectation

$$\mu := \mathbb{E}[\mathbf{X}] = \begin{bmatrix} \mathbb{E}[X_1] \\ \vdots \\ \mathbb{E}[X_k] \end{bmatrix}$$

and covariance matrix $\Sigma \in \mathbb{S}^k_+$ with

$$\Sigma_{ij} := \mathbb{E}[\ (X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])\ ]$$

We will assume that the covariance matrix has full rank so the *concentration matrix* or *precision matrix* Σ^{-1} is the inverse of the covariance matrix Σ.

Let m be positive integer. Every positive definite matrix $\Sigma \in \mathbb{S}^m_{++}$ is the covariance matrix of multivariate normal distribution on $\mathbb{R}^m$. This paper studies statistical models for which the concentration matrix $K := \Sigma^{-1}$ is written as a linear combination

$$K = \sum_{i=1}^{d} \lambda_i K_i \tag{1}$$

for some linearly independent matrices $K_i \in \mathbb{S}^m$ and unknown values $\lambda_i \in \mathbb{R}$. Such a model is called a *linear concentration model*, which was introduced by Anderson in 1970 [1].

Fix the linearly independent matrices K_i, and let $\mathcal{L} := \operatorname{span}\{K_1, \ldots, K_d\}$ be a linear subspace of $\mathbb{S}^m$. The basic statistical problem is to estimate the unknown parameters λ_i in (1) from n observations $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(n)}$ drawn from a multivariate normal distribution $\mathcal{N}(\mu, \Sigma)$ where $K := \Sigma^{-1} \in \mathcal{L}$.

Let $\bar{\mathbf{X}}$ be the sample mean of $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(n)}$. From $\bar{\mathbf{X}}$ and the n observations $\mathbf{X}^{(i)}$, we obtain the sample covariance matrix

$$S = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{X}^{(i)} - \bar{\mathbf{X}})(\mathbf{X}^{(i)} - \bar{\mathbf{X}})^\top \in \mathbb{S}^m_+$$

The linear concentration model also refers to any nonempty set of the form

$$\mathcal{L}^{-1}_{++} := \{\Sigma \in \mathbb{S}^m_{++} : \Sigma^{-1} \in \mathcal{L}\}$$

where $\mathcal{L}$ is a linear subspace of $\mathbb{S}^m$.

The *log-likelihood function* for the linear concentration model (1) is

$$\ell(K; \mathbf{X}^{(1)}, \dots \mathbf{X}^{(n)}) = \frac{n}{2}\left[\log\det(K) - \langle S, K\rangle\right]$$

This is a strictly concave real-valued function on the relatively open cone $\mathbb{S}^m_{++} \cap \mathcal{L}$.

If a maximum (known as the *maximum likelihood estimate* or *MLE*) for the log-likelihood function exists, then it is attained by a unique matrix $\hat{K} \in \mathbb{S}^m_{++} \cap \mathcal{L}$. Its inverse $\hat{\Sigma} = \hat{K}^{-1}$ is uniquely determined by the linear equations

$$\langle \hat{\Sigma}, K_i\rangle = \langle S, K_i\rangle \qquad \text{for } i = 1, \dots, d$$

This characterization of the inverse $\hat{\Sigma}$ follows from a theory of *exponential families* by Brown in 1986 [2]. In that theory, the scalars λ_i are the *canonical parameters* and the values $\langle S, K_i\rangle$ are the *sufficient statistics* of the exponential family in (1).

Let

$$\text{fiber}_{\mathcal{L}}(S) := \{\Sigma \in \mathbb{S}^m_{++} : \langle \Sigma, K\rangle = \langle S, K\rangle \text{ for all } K \in \mathcal{L}\}$$

This is the set of all full-rank covariance matrices whose sufficient statistics are given by S.

For a fixed linear subspace $\mathcal{L}$ of $\mathbb{S}^m$, and for a fixed basis $\{K_1, \dots, K_d\}$ of $\mathcal{L}$, the *cone of concentration matrices* is the relatively open cone

$$\mathcal{K}_{\mathcal{L}} := \mathcal{L} \cap \mathbb{S}^m_{++}$$

By the Universal Property of Quotients, consider the canonical map from $\mathbb{S}^m \to \mathbb{S}^m/\mathcal{L}^\perp$. Since $\mathbb{S}^m/\mathcal{L}^\perp \cong \mathbb{R}^d$, and with the fixed basis $\{K_1, \dots, K_d\}$, this allows us to consider the map $\pi_{\mathcal{L}} : \mathbb{S}^m \to \mathbb{R}^d$ given by

$$\pi_{\mathcal{L}}(S) = \begin{bmatrix} \langle S, K_1\rangle \\ \vdots \\ \langle S, K_d\rangle \end{bmatrix}$$

Let $C := \pi_{\mathcal{L}}(\mathbb{S}^m_{++})$ be the image of $\pi_{\mathcal{L}}$ on the positive definite cone. Again, considering the canonical isomorphism $\tau : \mathbb{R}^d \to \mathbb{S}^m/\mathcal{L}^\perp$, $\mathcal{C}_{\mathcal{L}} := \tau(C)$ is the *cone of sufficient statistics.*

Proposition 1.1. *$\mathcal{C}_{\mathcal{L}}$ is dual to $\mathcal{K}_{\mathcal{L}}$ in the sense that*

$$\mathcal{C}_{\mathcal{L}} = \{S \in \mathbb{S}^m/\mathcal{L}^\perp : \langle S, K\rangle > 0, \qquad \forall K \in \mathcal{K}_{\mathcal{L}}\}$$

Note that the MLE exists for a sample covariance matrix S if and only if $\text{fiber}_{\mathcal{L}}(S)$ is nonempty. By formulating the MLE as a convex problem constraint to a convex cone and taking the Lagrangian dual, the set of dual solutions will end up begin $\text{fiber}_{\mathcal{L}}(S)$.

Also, the map $K \mapsto T = \pi_{\mathcal{L}}(K^{-1})$ is a homeomorphism between $\mathcal{K}_{\mathcal{L}}$ and $\mathcal{C}_{\mathcal{L}}$. The inverse map $T \mapsto K$ takes the sufficient statistics to the MLE of the concentration matrix.

For a statistical model, define the *ML degree* as the degree of the algebraic function which maps the sufficient statistics to the MLE. [3]

An *undirected Gaussian graphical model* is when the subspace $\mathcal{L}$ is defined by vanishing some off-diagonal entries of the concentration matrix K. Let $G = ([m], E)$ be a graph, where E contains all self-loops. Then, a basis for E is the set $\{e_ie_j + e_je_i : ij \in E\}$. In this undirected Gaussian graphical model, write $\mathcal{K}_G, \mathcal{C}_G, \text{fiber}_G(S)$ for $\mathcal{K}_{\mathcal{L}}, \mathcal{C}_{\mathcal{L}}, \text{fiber}_{\mathcal{L}}(S)$, respectively.

Also, for a sample covariance matrix S,

$$\text{fiber}_G(S) = \{\Sigma \in \mathbb{S}^m_{++} : \Sigma_{ij} = S_{ij}, \quad \forall\ ij \in E\}$$

2. Main Results and Techniques

One of the leading questions asked by the authors is the following.

Question. The map taking a sample covariance matrix S to its maximum likelihood estimate $\hat{\Sigma}$ is an algebraic function. Its degree is the ML degree of the model $\mathcal{L}$. Can we find a formula for this ML degree? Which models $\mathcal{L}$ have their ML degree equal to 1?

The authors investigated this question when applied to the undirected Gaussian graphical model for m-holes, chordal graphs, suspensions, wheel graphs, and graphs with 5 or fewer vertices.

2.1. **The m-hole.** Let C_m be an m-hole (an induced cycle on m vertices). The following conjecture was given by Drton et al. in 2009:

Conjecture 2.1.

$$\text{ML-degree}(C_m) = (m-3)\cdot 2^{m-2} + 1, \qquad \textit{for } m \geq 3$$

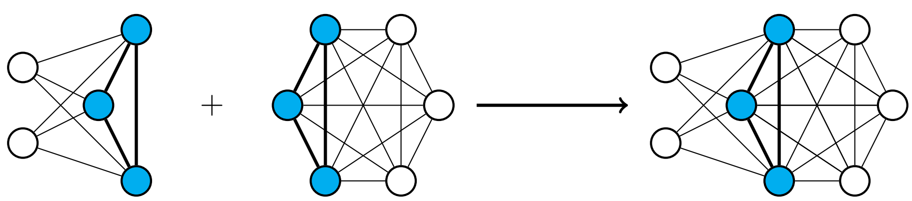
The following is a lemma claimed by the authors.

Conjecture 2.2. *For $m \geq 4$,*

$$\text{ML-degree}(C_m) > 1$$

The authors were unable to provide a proof of this claim.

2.2. **Chordal Graphs.** A graph G is *chordal* or *decomposable* if every induced cycle on at least 4 vertices contains a chord. A characterization of chordal graphs is that they are the clique-sum of complete graphs. The *clique-sum* of two graphs is the gluing of two graphs together at a clique:



Proposition 2.3. *Let G be a clique sum of n graphs $G_1, \dots, G_n$. Then,*

$$\text{ML-degree}(G) = \prod_{i=1}^{n} \text{ML-degree}(G_i)$$

Proof. By induction on n, it suffices to show the result when $n = 2$.

Let $G = (V, E)$ be a graph, where (A, B, C) is a partition of V such that C is a clique, and there are no edges between A and B.

Let $S \in \mathbb{S}^m$, and fix $\Sigma \in \mathbb{S}^m$ with $\Sigma_{ij} = S_{ij}$ for $ij \in E$, and variables $z_{ij} = \Sigma_{ij}$ for $ij \notin E$. Then, the ML degree of G is the number of complex solutions to the equations

$$(\spadesuit) \qquad (\Sigma)^{-1}_{ij} = 0, \qquad \forall\ ij \notin E$$

such that $\Sigma \succ 0$. Let

$$K = \Sigma^{-1}, \qquad K^1 = (\Sigma[A \cup C])^{-1}, \qquad K^2 = (\Sigma[B \cup C])^{-1}$$

Then,

$$K = \begin{bmatrix} K^1_{AA} & K^1_{AC} & 0 \\ (K^1_{AC})^\top & K_{CC} & K^2_{CB} \\ 0 & (K^2_{CB})^\top & K^2_{BB} \end{bmatrix}, K^1 = \begin{bmatrix} K^1_{AA} & K^1_{AC} \\ (K^1_{AC})^\top & K_{CC} \end{bmatrix}, K^2 = \begin{bmatrix} K_{CC} & K^2_{BC} \\ (K^2_{BC})^\top & K^2_{BB} \end{bmatrix}$$

Then, by Schur complement, a positive-definite completion of K^1 and K^2 gives a positive-definite completion of K, and every positive-definite completion of K gives a positive-definition completion of K^1 and K^2. Thus,

$$\text{ML-degree}(G) = \text{ML-degree}(G[A \cup C]) \cdot \text{ML-degree}(G[B \cup C])$$

□

Theorem 2.4. *If a graph G is chordal, then* ML-degree$(G) = 1$.

Proof. The undirected Gaussian graphical model for a complete graph K_m does not vanish any off-diagonals, resulting in a linear concentration model. Thus, the MLE exists and is unique, so ML-degree$(K_m) = 1$. Then, for a chordal graph G which is the clique sum of complete graphs,

$$\text{ML-degree}(G) = 1$$

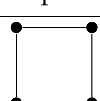
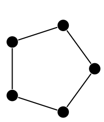
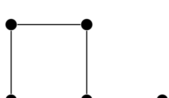
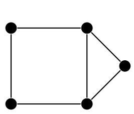
□

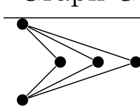
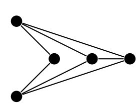
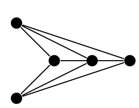
The following is the converse of the above. The authors used conjecture 2.2 which did not have a proof, so I have changed their theorem to a conjecture.

Conjecture 2.5. *If* ML-degree$(G) = 1$, *then G is chordal.*

If a graph G is not chordal, then it contains an m-hole for $m \geq 4$. The sufficient statistics defining the m-hole also define G, so the ML degree of G is bounded below by the ML degree of any of its induced subgraphs. If conjecture 2.2 is true, then ML-degree$(G) \geq$ ML-degree$(C_m) > 1$ gives the contrapositive.

2.3. **Small Graphs.** The authors computed the ML degree of all non-chordal graphs on at most 5 vertices, given below:

Graph G	ML-degree(G)
	5
	17
	5
	5

Graph G	ML-degree(G)
	7
	5
	5

2.4. **Suspension Graphs and Wheels.** Let $G = (V, E)$ be a graph. The *suspension graph* of G is $G^* = (V^*, E^*)$ where $V^* = V \cup \{0\}$, and $E^* = E \cup \{(0, v) : v \in V\}$. That is, to obtain the suspension graph of G, add a new vertex 0, and make it adjacent to all vertices.

Lemma 2.6. *[Inversion Formula by Schur Complement] Let*

$$M = \begin{bmatrix} A & B \\ C & D \end{bmatrix}$$

If M is invertible, and $A \succ 0$, then

$$M^{-1} = \begin{bmatrix} A^{-1} + A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1} & -A^{-1}B(D - CA^{-1}B)^{-1} \\ -(D - CA^{-1}B)^{-1}CA^{-1} & (D - CA^{-1}B)^{-1} \end{bmatrix}$$

I will not prove theorem 2.6 but it follows from Gaussian Elimination and the Schur Complement.

Theorem 2.7. *Let G be a graph. Then,*

$$\text{ML-degree}(G) = \text{ML-degree}(G^*)$$

Proof. Let $G = ([m], E)$ be a graph. Let $S^* \in \mathbb{S}^{m+1}_{++}$ be a sample covariance matrix on G^*. Let

$$S^* = \begin{bmatrix} S^*_{00} & v^\top \\ v & S' \end{bmatrix}$$

By the Schur complement,

$$S := S' - \frac{1}{S^*_{00}} vv^\top \succ 0$$

so S is a sample covariance matrix on G. Let $\hat{\Sigma}$ be the MLE for S on the undirected Gaussian graphical model on G. They claim that

$$\hat{\Sigma}^* := \begin{bmatrix} S^*_{00} & v^\top \\ v & \hat{\Sigma} + S' - S \end{bmatrix}$$

is the MLE for S^* on G^*.

Note that $S^*_{00} > 0$, so by the Schur complement,

$$\hat{\Sigma}^* \succ 0 \iff \hat{\Sigma} + S' - S - \frac{1}{S^*_{00}} vv^\top \succ 0 \iff \hat{\Sigma} \succ 0$$

which is true because $\hat{\Sigma}$ is the MLE.

Then, for any edge $ij \in E$ with $i = 0$ or $j = 0$, clearly $\hat{\Sigma}^*_{ij} = S^*_{ij}$. Suppose $i \neq 0, j \neq 0$. Then, since $\hat{\Sigma}$ is the MLE for S on G, $\hat{\Sigma}_{ij} = S_{ij}$, so

$$\hat{\Sigma}^*_{ij} = (\hat{\Sigma} + S' - S)_{ij} = S'_{ij} = S^*_{ij}$$

Next, applying the inversion formula in theorem 2.6,

$$(\hat{\Sigma}^*)^{-1} = \left[\begin{array}{c|c} \frac{1}{S^*_{00}} + v^\top \hat{\Sigma}^{-1} v & \frac{1}{S^*_{00}} \hat{\Sigma}^{-1} v \\ \hline \frac{1}{S^*_{00}} v^\top \hat{\Sigma}^{-1} & \hat{\Sigma}^{-1} \end{array}\right]$$

hence $(\hat{\Sigma}^*)^{-1}_{ij} = (\hat{\Sigma})^{-1}_{ij} = 0$ for all $ij \notin E$. Thus, $\hat{\Sigma}^*$ is the MLE for S^* on G^*.

Then, $\hat{\Sigma}^*$ is a rational function of $\hat{\Sigma}$, so each solution for $\hat{\Sigma}$ maps to a solution for $\hat{\Sigma}^*$. Then,

$$\text{ML-degree}(G^*) \leq \text{ML-degree}(G)$$

Conversely, for any $S \in \mathbb{S}^m_{++}$, computing the MLE on G^* for $S^* := \begin{bmatrix} 1 & 0 \\ 0 & S \end{bmatrix} \in \mathbb{S}^{m+1}_{++}$ gives the MLE for G once projected back to $\mathbb{S}^m_{++}$. This gives the MLE for S on G as a rational function of the MLE for S^* on G^*, thus giving

$$\text{ML-degree}(G^*) \geq \text{ML-degree}(G)$$

And hence we conclude that

$$\text{ML-degree}(G^*) = \text{ML-degree}(G)$$

□ □

The wheel graph W_k on $k+1$ vertices is the suspension graph of the chordless k-cycle: $W_k = (C_k)^*$. Thus, for $k \geq 4$,

$$\text{ML-degree}(W_k) = \text{ML-degree}(C_k) > 1$$

The authors also explored the equivalence of the ML degree in the linear concentration model to the beta-invariant $\beta(\mathcal{L})$ of the corresponding matroid.

3. Potential Applications, New Experiments, and Relations to CO671

3.1. **Low ML degree models.** The ML degree of a model describes the degree of the algebraic function which gives the MLE. A model with a lower ML degree would be much easier to compute, so it would be useful to be able to characterize models with lower ML degree.

This was one of the authors' leading questions; namely, to characterize the linear subspaces $\mathcal{L}$ which yielded an ML degree of 1. They claimed that in the undirected Gaussian graphical model that the Chordal graphs characterized the models with ML degree 1. I am interested in knowing if there are graphs which yield ML degrees of 2, 3, or 4. From Section 2.3, I can see that there are no graphs on ≤ 5 vertices which have ML degree less than 5.

3.2. **A callback to vertex-transitive graphs.** Recall Conjecture 2:

Conjecture 2. Let C_m be the induced m-cycle. Then, ML-degree$(C_m) > 1$.

The authors offered a proof sketch. Take $m = 7$. For $x_1, \ldots, x_7 \in [-1, 1]$, we wish to find a positive-definite matrix completion of

$$\Sigma_7 = \begin{bmatrix} 1 & x_1 & ? & ? & ? & ? & x_7 \\ x_1 & 1 & x_2 & ? & ? & ? & ? \\ ? & x_2 & 1 & x_3 & ? & ? & ? \\ ? & ? & x_3 & 1 & x_4 & ? & ? \\ ? & ? & ? & x_4 & 1 & x_5 & ? \\ ? & ? & ? & ? & x_5 & 1 & x_6 \\ x_7 & ? & ? & ? & ? & x_6 & 1 \end{bmatrix}$$

which maximizes the log-likelihood function, which is the same as maximizing $\det(\Sigma_7)$.

The authors suggest taking $x := x_1 = \cdots = x_7$. Then, since the set of MLEs is convex, and the solution set is symmetric under the dihedral group, we can write the variables as

$$\Sigma_7 = \begin{bmatrix} 1 & x & s_1 & s_2 & s_2 & s_1 & x_7 \\ x & 1 & x & s_1 & s_2 & s_2 & s_1 \\ s_1 & x & 1 & x & s_1 & s_2 & s_2 \\ s_2 & s_1 & x & 1 & x & s_1 & s_2 \\ s_2 & s_2 & s_1 & x & 1 & x & s_1 \\ s_1 & s_2 & s_2 & s_1 & x & 1 & x \\ x_7 & s_1 & s_2 & s_2 & s_1 & x & 1 \end{bmatrix}$$

Then, Σ_7 is a circulant matrix, and we can try to simplify the algebraic formula for x, s_1, s_2 in $\det(\Sigma_7)$. Using computational software (Singular), the authors were able to show that the solutions are a degree 3 algebraic function in x, giving that ML-degree$(C_7) \geq 3$.

They were able to use this method to show that ML-degree$(C_m) > 1$ for all of $4 \leq m \leq 12$.

Although this method isn't practical for showing the result for all of $m \geq 4$, as the dependency on computational software has an obvious limitation, the method they used to arrive at the circulant matrix is the same method we used for solve some of the SDPs in CO671 (namely, the computation of the maximum cardinality version of MaxCut for C_5 and C_7 as well as for computing $\theta(G)$ for the Petersen graph).

I wonder if there are more tricks used in the computation of MaxCut for C_{2n-1} that could also be leveraged to complete this proof. However, the diffculty seems to arise in the high non-linearity of the determinant function, making it very hard to compute with a closed formula.

4. Scientific Criticism

The authors claimed that ML-degree$(C_m) > 1$ for $m \geq 4$ and that the proof would appear in the dissertation by the second author, but such a proof does not exist in the dissertation. The author proved the result for $4 \leq m \leq 12$, but left the proof open for $m > 12$ [6].

The remaining remarks I have are about typos and inconsistencies through the paper.

- The authors describe various sets as being spectrahedra. For instance, in their definition

$$\text{fiber}_{\mathcal{L}}(S) := \{\Sigma \in \mathbb{S}^m_{++} : \langle \Sigma, K \rangle = \langle S, K \rangle, \text{ for all } K \in \mathcal{L}\}$$

  they describe $\text{fiber}_{\mathcal{L}}(S)$ as a spectrahedron. This does not align with our definition of a spectrahedron due to the positive-definite constraint, but they also do not offer their definition of a spectrahedron.

- On page 4, the authors write $\mathcal{K}_{\mathcal{L}}$ to refer to both $\mathcal{L} \cap \mathbb{S}^m_{++}$ and the cone in $\mathbb{R}^d$ to which $\mathcal{L} \cap \mathbb{S}^m_{++}$ is isomorphic. I think that it would be a little less confusing to use a slightly different notation in spite of the obvious isomorphism. They do the same for $\mathcal{C}_{\mathcal{L}}$, the cone of sufficient statistics.

- On page 18, the paper writes $S^m_{\succ 0}$ instead of $\mathbb{S}^m_{\succ 0}$

- When discussing MLE for a sample covariance matrix for a undirected Gaussian graphical model, the authors switch between writing "MLE for [sample covariance matrix] on [graph]" and "MLE for [graph] on [sample covariance matrix]". This is never an issue due to the abundance of context, but it is inconsistent.
- On page 11, in their definition of the undirected Gaussian graphical model, the authors fix a graph $G = ([m], E)$, and write that "a basis for $\mathcal{L}$ is the set $\{K_{ij} : ij \in E\}$ of matrices K_{ij} with a single 1-entry in position (i, j), and 0 elsewhere". Since $\mathcal{L}$ is a subspace of $\mathbb{S}^m$, each element must be a symmetric matrix, so I believe they meant to write "in position (i, j) and (j, i)".

References

[1] Anderson, T. W. (1970). Estimation of covariance matrices which are linear combinations or whose inverses are linear combinations of given matrices. In R. C. Bose, I. M. Chakravati, P. C. Mahalanobis, C. R. Rao and K. J. C. Smith (Ed.), Essays in Probability and Statistics (pp.1-24). Chapel Hill: Univ. North Carolina Press.

[2] Brown, L. D. (1986). Fundamentals of Statistical Exponential Families. IMS-monographs, Vol. IX. Hayward: Institute of Mathematical Statistics.

[3] Drton, M., Sturmfels, B., Sullivant, S. (2009). Lectures on Algebraic Statistics. Oberwolfach Seminars, Vol. 40. Basel: Birkhäuser.

[4] Højsgaard, S., Lauritzen, S. (2008). Graphical Gaussian models with edge and vertex symmetries. J. Roy. Stat. Soc., 70, 1005-1027.

[5] Sturmfels, B., Uhler, C.: Multivariate Gaussians, Semidefinite Matrix Completion, and Convex Algebraic Geometry. Ann. Inst. Statist. Math. 62 (2010), no. 4, 603-638.

[6] Uhler, C. (2011). Geometry of maximum likelihood estimation in Gaussian graphical models. UC Berkeley. ProQuest ID: Uhler_berkeley_0028E_11729. Merritt ID: ark:/13030/m5f193k1. Retrieved from https://escholarship.org/uc/item/6pz47470

Faculty of Combinatorics and Optimization, University of Waterloo

Email address: frank.wang@uwaterloo.ca