Continuum Babington-Smith model

Sangjun Moon[1] and Jong-June Jeon[1]

November 19, 2019

[1]Department of Statistics
University of Seoul

Introduction

Figure 1: Formula 1, the highest class of single-seater auto racing

Three types of ranking problems

- Who will win? (Top-K ranking problem)
- What will be the final ranking? (Learning to ranks)
- What is the probability for each rank? (probability model for ranks)

- The probability model for ranking is actually the probability model on permutations (ranking model)
- Since the ranking model is the discrete probability model assigning probability on each permutation, it requires numerous parameters when the number of ranked items is large.
- For example, suppose that we are considering the ranking model with 10 players. Then, there are about 3,620,000 parameters in the model. Thus, it is necessary to develop the parametric model effectively accounting for the patterns of observed rankings.

Conventional Ranking Models

- The early development of ranking model is motivated by the following question: 'How can we estimate the entire ranks of item only with pairwise comparison experiments?'
- Bradley-Terry (BT) model is one of the most famous ranking model. When BT model (1952) was first proposed, the model focused on the estimation of (the most probable) ranking itself, not the probability model for ranking.
- Bradley-Terry model has p parameters for p items and likelihood method is widely used for estimation of the model.

- Babington Smith (1950) proposed the ranking model with $p(p-1)/2$ parameters based on the pairwise comparison experiments for p items.
- However, the estimation of Babington-Smith (BS) model is computationally difficult at that time for large p, because it contains complex normalized term depending on the model parameters.
- Mallows (1957) proposed a ranking model simplifying Babington Smith model. Interestingly Mallows explains the result of pairwise comparisons as a distance from the unimodal ranking. Mallows model has just two parameters associated with location and dispersion.

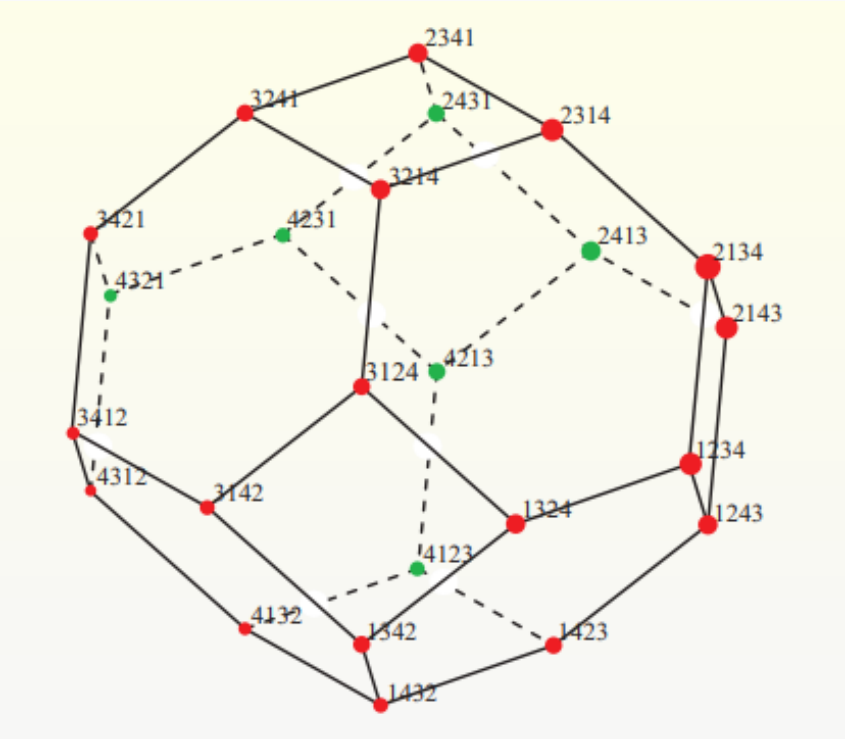

Figure 2: Mallows model and the permutation polytope

- From generalization of Mallows model, the likelihood of BT model induces the Bradley-Terry-Mallows (BTM) model, the probability model for ranking. The BTM model has p parameters.
- Family of BS model: Mallows model $\subset$ BTM $\subset$ BS model.
- Can we construct various models between each developed model?

- From generalization of Mallows model, the likelihood of BT model induces the Bradley-Terry-Mallows (BTM) model, the probability model for ranking. The BTM model has p parameters.
- Family of BS model: Mallows model $\subset$ BTM $\subset$ BS model.
- Can we construct various models between each developed model?

By adopting the idea of LARS, Regularize the BS model!

Goal

- The aim of this paper is to construct a continuum of probability models between the BS and BTM model by regularization.
- In addition, we propose a computational algorithm to obtain the the penalized likelihood estimator.
- We use l_1 regularization and obtain the penalized likelihood estimator based on the alternating direction methods of multipliers (ADMM) method.

Contributions

- Development of new models as variants of BS model.
- Development of computation algorithm for BS and BTM model.
- Investigation of theoretical property of the proposed algorithm.

Continuum of Barbington-Smith Model

Babington Smith Model

- π is a random permutation of $[p] = \{1, \cdots, p\}$ and π_j denotes the rank of item j.
- Let $I_{jk}(\pi)$ be the indicator function that indicates the j precedes k in the ranking π.
- Let $I(\pi) = (I_{jk}(\pi), 1 \leq j < k \leq p)^\top \in \mathbb{R}^{\tilde{p}}$ with $\tilde{p} = p(p-1)/2$, which represents the random permutation, the rank π.

- Let $\alpha_{jk} = \Pr(I_{jk}(\pi) = 1)$ be the probability that j precedes k in ranking π.
- The probability of the ranking π in the BS model is defined as follows:

$$\Pr(\pi; \boldsymbol{\alpha}) = K \prod_{j<k} [\alpha_{jk}]^{I_{jk}(\pi)} [1 - \alpha_{jk}]^{1 - I_{jk}(\pi)}$$

  where the normalizing term K is

$$K = \sum_{\pi \in S_p} \prod_{j<k} \alpha_{jk}^{I_{jk}(\pi)} (1 - \alpha_{jk})^{1 - I_{jk}(\pi)}.$$

- For example let $p = 3$ and $\pi = (1, 3, 2)$, i.e. $(1 \to 3 \to 2)$, then

$$\Pr(\pi) = K\alpha_{13}\alpha_{12}(1 - \alpha_{23})$$

where

$$K = \alpha_{12}\alpha_{13}\alpha_{23} + \cdots + (1 - \alpha_{12})(1 - \alpha_{13})(1 - \alpha_{23})$$

Family of BS model

- BTM model: Let

$$\alpha_{jk} = \frac{u_j}{u_j + u_k},$$

  where $\mathbf{u} = (u_1, \cdots, u_p) \in \mathbb{R}^p$, $u_j > 0$ for all $j \in [p]$.

- Mallows model: Re-define $I_{jk}(\pi)$ be the indicator of concordance pair (j, k) to σ (location parameter) and let

$$\alpha_{jk} = \phi \quad \text{(scale parameter)}$$

  for all $j < k$.

Continuum of BS Model

- We try to make a continuous ranking model that can represent all models between the BS and BT models according to the degree of regularization.
- We model the probability with parameters $\boldsymbol{\alpha} = (\alpha_j, j = 1, \cdots, p)^\top \in \mathbb{R}^p$ and $\boldsymbol{\gamma} = (\gamma_{jk}, 1 \leq j < k \leq p)^\top \in \mathbb{R}^{\tilde{p}}$.
- In addition, denote a vector excluding first p elements in $\mathbf{x}$ by $\mathbf{x}_{-(1:p)}$. Conversely, a vector including first p elements in $\mathbf{x}$ by $\mathbf{x}_{(1:p)}$.

- We model the α_{jk} as follows.

$$\alpha_{jk} = \frac{\exp(\alpha_j - \alpha_k + \gamma_{jk})}{\exp(\alpha_j - \alpha_k + \gamma_{jk}) + 1}, \quad j < k$$

  subject to $\sum_{j=1}^{p} \alpha_j = 0$, $\sum_{j=1}^{k-1} \gamma_{jk} = 0$ for each $k = 2, \cdots, p$ for identifiability.

- We can obtain the continuous ranking model by applying a regularization method to $\boldsymbol{\gamma}$ since if $\gamma_{jk} = 0$ for all $j < k$ then our model is same as the BTM model and if $\gamma_{jk} \neq 0$ for any $j < k$ then our model is same as the BS model.

Likelihood

- Since $\Pr(\pi; \boldsymbol{\alpha}, \boldsymbol{\gamma}) = \Pr(I(\pi); \boldsymbol{\alpha}, \boldsymbol{\gamma})$,

$$
\begin{aligned}
\log \Pr(\pi; \boldsymbol{\alpha}, \boldsymbol{\gamma}) &= \log K(\boldsymbol{\alpha}, \boldsymbol{\gamma}) + \\
&\quad \sum_{j<k} I_{jk}(\pi) \log\left(\frac{\alpha_{jk}}{1-\alpha_{jk}}\right) + \log(1-\alpha_{jk}) \\
&= \log K(\boldsymbol{\alpha}, \boldsymbol{\gamma}) + \\
&\quad \sum_{j<k} I_{jk}(\pi)\theta_{jk} - \log(1+\exp(\theta_{jk})) \\
&= \sum_{j<k} I_{jk}(\pi)\theta_{jk} + \log\left(\frac{K(\boldsymbol{\alpha}, \boldsymbol{\gamma})}{\prod_{j<k}(1+\exp(\theta_{jk})}\right)
\end{aligned}
$$

where $\boldsymbol{\Theta} = (\theta_{jk}, 1 \le j < k \le p)^\top \in \mathbb{R}^{\tilde{p}}$, $\theta_{jk} = \alpha_j - \alpha_k + \gamma_{jk}$, $j < k$.

- Let π^i be the i-th observation of π. To write the loglikelihood as canonical form of exponential family, we introduce a vector $\mathbf{x}^i = (I_{jk}(\pi^i), 1 \leq j < k \leq q)^\top \in \mathcal{X} := \{0,1\}^{\tilde{p}}$ for π^i
- Then, the log likelihood function of our model is written by

$$\begin{aligned} L(\boldsymbol{\Theta}|\mathcal{X}) &= \sum_{i=1}^{n} \log \Pr(\pi^i; \boldsymbol{\Theta}) \\ &= \sum_{i=1}^{n} \left(\sum_{j<k} I_{jk}(\pi^i)\theta_{jk} - \log Z(\boldsymbol{\Theta}) \right) \\ &= \sum_{i=1}^{n} \boldsymbol{\Theta}^\top \mathbf{x}^i - n \log Z(\boldsymbol{\Theta}), \end{aligned}$$

where $Z(\boldsymbol{\Theta}) = \frac{\prod_{j<k} 1+\exp(\theta_{jk})}{K(\boldsymbol{\alpha},\boldsymbol{\gamma})}$.

Parameter Constraints

- We wish to shrink γ_{jk} not θ_{jk} by regularization. In addition we have constraints for $\{\alpha_j\}$ and $\{\gamma_{jk}\}$ for identifiability.
- For example, when $p = 3$, the design matrix A and $\boldsymbol{\beta}$ can be represented as follows.

$$
\begin{aligned}
A &= \begin{pmatrix}
1 & -1 & 0 & 1 & 0 & 0 \\
1 & 0 & -1 & 0 & 1 & 0 \\
0 & 1 & -1 & 0 & 0 & 1 \\
1 & 1 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 1
\end{pmatrix} \\
\boldsymbol{\beta} &= (\alpha_1, \cdots, \alpha_3, \gamma_{12}, \cdots, \gamma_{23})^\top \in \mathbb{R}^6.
\end{aligned}
$$

- Let

$$D = \begin{pmatrix} I_{\tilde{p}\times\tilde{p}} \\ 0_{p\times\tilde{p}} \end{pmatrix} \in \mathbb{R}^{(p+\tilde{p})\times\tilde{p}}.$$

  Then,

$$A\beta = (\theta_{12}, \theta_{13}, \theta_{23}, 0, 0, 0)^\top = D\mathbf{\Theta},$$

- Since A is invertible,

$$\beta = A^{-1}D\mathbf{\Theta}$$

We use l_1-penalization with $\lambda \sum_{jk} |\gamma_{jk}|$, which is written by $\lambda \|(A^{-1}D\mathbf{\Theta})_{-(1:3)}\|_1$ in terms of θ.

Continuum of BS Model

We can set A for general $p \geq 3$. Then, the continuum BS model is estimated by the following minimization problem.

$$\min \qquad -\mathcal{L}(\boldsymbol{\Theta}|\mathcal{X}) + \lambda \|(A^{-1} D_\theta \boldsymbol{\Theta})_{-(1:p)}\|_1$$

Computational Issue

- The object function contains non-differentiable penalty.
- The computation of $Z(\mathbf{\Theta})$ and $\partial Z(\mathbf{\Theta})/\partial \mathbf{\Theta}$ is intractable due to the summations over all permutations when p is large.

Computation

- The object function contains non-differentiable penalty $\rightarrow$ ADMM
- The computation of $Z(\mathbf{\Theta})$ and $\partial Z(\mathbf{\Theta})$ is intractable due to the summations over all permutations when p is large. $\rightarrow$ Contrastive Divergence algorithm

- Our objective function can be solved through ADMM.

$$\begin{aligned} \min \quad & -\mathcal{L}(\mathbf{\Theta}|\mathcal{X}) + \lambda \|\mathbf{z}_{-(1:p)}\|_1 \\ \text{subject to} \quad & \mathbf{z} = A^{-1} D \mathbf{\Theta} \end{aligned}$$

- The process is divided into three steps:

$$\begin{aligned} \mathbf{\Theta}^{(k+1)} &= \underset{\theta}{\text{argmin}} \quad -\mathcal{L}(\mathbf{\Theta}|\mathcal{X}) + \frac{\rho}{2} \|\mathbf{z}^{(k)} - A^{-1} D \mathbf{\Theta} + \mathbf{u}^{(k)}\|_2^2 \\ \mathbf{z}^{(k+1)} &= \underset{z}{\text{argmin}} \quad \lambda \|\mathbf{z}_{-(1:p)}\|_1 + \frac{\rho}{2} \|\mathbf{z} - A^{-1} D \mathbf{\Theta}^{(k+1)} + \mathbf{u}^{(k)}\|_2^2 \\ \mathbf{u}^{(k+1)} &= \mathbf{u}^{(k)} + \mathbf{z}^{(k+1)} - A^{-1} D \mathbf{\Theta}^{(k+1)} \end{aligned}$$

Update of $\boldsymbol{\Theta}$ (Contrastive Divergence Algorithm)

- Consider a (proposal) distribution

$$\Pr(\mathbf{x}; \boldsymbol{\Theta}_0) = \exp(\boldsymbol{\Theta}_0^T \mathbf{x}) / Z(\boldsymbol{\Theta}_0).$$

  and let $\tilde{\mathbf{x}}^i \sim_{iid} \Pr(\mathbf{x}; \boldsymbol{\Theta}_0)$.

- Then,

$$\frac{\partial \mathcal{L}(\boldsymbol{\Theta}|\mathcal{X})}{\partial \boldsymbol{\Theta}} \simeq \sum_{i=1}^{n} \mathbf{x}^i - \frac{\sum_{l=1}^{m} \tilde{\mathbf{x}}^l h_l(\boldsymbol{\Theta}, \tilde{\mathbf{x}}^l; \boldsymbol{\Theta}_0)}{\sum_{l=1}^{m} h_l(\boldsymbol{\Theta}, \tilde{\mathbf{x}}^l; \boldsymbol{\Theta}_0)} := g(\boldsymbol{\Theta}; \boldsymbol{\Theta}_0, m)$$

  where $h_l = \exp((\boldsymbol{\Theta} - \boldsymbol{\Theta}_0)^T \tilde{\mathbf{x}}^l)$.

- Then, we update the gradient decent method with $\frac{\partial \mathcal{L}(\boldsymbol{\Theta}|\mathcal{X})}{\partial \boldsymbol{\Theta}}$.

Proposal distribution

- Mallows ϕ model is a special case of BS model such that the model can be written by the same type of the BS model.
- The probability function of Mallows ϕ model is

$$\Pr(\pi; \sigma, \phi) = \frac{1}{Z(\phi)} \phi^{d(\pi,\sigma)}, \quad \phi \in (0, 1]$$

where $d(\pi, \sigma) = \sum_{j<k} I(\sigma(k) \to \sigma(j) \in r)$ and $Z(\phi) = (1 + \phi)(1 + \phi + \phi^2) \cdots (1 + \cdots + \phi^{p-1})$.

Assumption

L is λ-strongly convex. Also, there exists $G > 0$ such that

$$\left\| \frac{\partial}{\partial \mathbf{\Theta}} L(\mathbf{\Theta}|\mathcal{X}) \right\|_2 \leq G, \quad \forall \mathbf{\Theta} \in \mathcal{C}$$

Theorem (Large deviation bound for approximation) Let $0 < \gamma < \sqrt{2\lambda}$. Then there exists a constant $d > 0$ only depending on $\epsilon > 0$ such that

$$\Pr_{\theta_0} \left(\left\| \frac{\partial}{\partial \mathbf{\Theta}} \hat{L}_m(\mathbf{\Theta}|\mathcal{X}) - \frac{\partial}{\partial \mathbf{\Theta}} L(\mathbf{\Theta}|\mathcal{X}) \right\|_\infty > \gamma\epsilon \right) < \exp\left(-\gamma^2\epsilon^2 dm\right).$$

for a fixed $\mathbf{\Theta} \in \mathcal{C}$.

Theorem implies the gradient $\frac{\partial}{\partial \mathbf{\Theta}} \hat{L}_m(\mathbf{\Theta}|\mathcal{X})$ is approximated to the gradient (asymptotically unbiased estimator of the gradient). Thus, we can apply the theoretical properties of stochastic gradient algorithm.

Theorem

Let $0 < \gamma < \sqrt{2\lambda}$. *Consider diminishing learning rate* $\eta_k = \frac{2}{(2\lambda - \gamma^2)k}$ *and fix* $\epsilon, \xi \in (0, 1)$ *arbitrarily. Then for sufficiently large* m

$$\Pr_{\theta_0}\left(\|\boldsymbol{\Theta}^{(k)} - \boldsymbol{\Theta}^*\|_2^2 \leq \frac{L}{k} + \frac{\epsilon^2}{2\lambda - \gamma^2} \right) \geq 1 - \xi$$

where

$$L \quad = \quad \max\left(\|\boldsymbol{\Theta}^{(1)} - \boldsymbol{\Theta}^*\|_2^2, \frac{4(\gamma^2\epsilon^2 + G^2)}{(2\lambda - \gamma^2)^2} \right)$$

Theorem implies the convergence of solution at the first step in ADMM.

Elections of the American Psychological Association ($p = 5$, $n = 5328, 1777, 1776$(traning, validation, test))

	BS	BTM	CBS(aic)	CBS(bic)	CBS(best)
KL	0.138	0.163	0.139	0.137	0.135
TV	0.204	0.235	0.205	0.205	0.201

Table 1: APA 2008

	BS	BTM	CBS(aic)	CBS(bic)	CBS(best)
KL	0.232	0.227	0.256	0.233	0.201
TV	0.274	0.272	0.289	0.276	0.254

Table 2: APA 2009

Concluding Remarks

Concluding remarks

- We develop new models as variants of BS model called of continuum Babington Smith model/
- We propose computation algorithm for BS and BTM model.
- We prove theoretical property of the proposed computational algorithm partly.
- We are proving the convergence of our ADMM algorithm and are studying estimation method with missing ranks.

Thank you!