MLE Voting Rules via Bregman Divergence

June 6, 2013

1 Results I: Positional Scoring Rules

1.1 Additional Notation

- Let $A = \{1, \ldots, m\}$ denote the set of alternatives. Let $\mathcal{L}(A)$ denote the set of all linear orders over A (votes). We view a ranking $\sigma \in \mathcal{L}(A)$ as $\sigma : A \to \{1, \ldots, m\}$. Thus, $|\mathcal{L}(A)| = m!$. Finally, we use $\pi \in \mathcal{L}(A)^n$ to denote a profile of n votes.

- Let $\alpha \in \mathbb{R}^m_{\geq 0}$ denote a score vector, where $\alpha_i \geq \alpha_{i+1}$ for $i \geq 1$. We assume that $\alpha_i > \alpha_{i+1}$ for some i, otherwise the rule is meaningless.

- Given score vector α, there is a natural positional scoring rule which gives appropriate scores to alternatives and ranks them according to their total score. We denote it by SC^α.

- For any score vector α, let ϕ^α be the representation such that for any $\sigma \in \mathcal{L}(A)$, $\phi^\alpha_i(\sigma) = \alpha_{\sigma(i)}$ for all i.

- Given a representation $\phi : \mathcal{L}(A) \to \mathbb{R}^k$, let MM^ϕ denote MLE-MODE method with representation ϕ, i.e., the voting rule that first finds the MLE parameter of exponential family over rankings with representation ϕ, and then returns the mode ranking of the exponential distribution given by the MLE parameter.

- Finally, let $\mathrm{SORT} : \mathbb{R}^m \to \mathcal{L}(A)$ denote the function that takes an m-dimensional vector, and returns the sorted order of indices. That is, for any $v \in \mathbb{R}^m$, alternative i is mapped to position j in $\mathrm{SORT}(v)$ if there are $j-1$ coordinates that have value greater than value of coordinate i. We break ties arbitrarily, as they do not matter for our results.

Let $\hat{\mu}$ and $\hat{\theta}$ denote the MLE mean and natural parameters respectively.

1.2 Recovering Positional Scoring Rules

Theorem 1. *For any score vector α, the MM method with representation ϕ^α reduces to the scoring rule SC^α, irrespective of the selection of MLE parameter.*

Proof. Fix arbitrary score vector α. Consider any profile $\pi = (\sigma_1, \ldots, \sigma_n)$. We want to show that $\mathrm{MM}^{\phi^\alpha}(\pi) = \mathrm{SC}^\alpha(\pi)$. First, using the famous result on exponential families, we have that the MLE mean parameter $\hat{\mu} = 1/n \cdot \sum_{i=1}^n \phi^\alpha(\sigma_i)$. Further, note that $\hat{\mu}$ is the vector of average scores of candidates in profile π according to score vector α. Hence, it is clear that $\mathrm{SC}^\alpha(\pi) = \mathrm{SORT}(\hat{\mu})$.

On the other hand, if $\hat{\theta}$ denotes the MLE natural parameter, then the mode ranking is given by $\arg\max_\sigma \langle \hat{\theta}, \phi^\alpha(\sigma) \rangle$. Note that since $\phi(\sigma)$ is just a re-ordering of the terms of α, by Chebyshev's inequality, the dot product is maximized when both vectors have value sorted in the same order. That is, the dot product is maximized by $\sigma = \mathrm{SORT}(\hat{\theta})$. Hence, $\mathrm{MM}^{\phi^\alpha}(\pi) = \mathrm{SORT}(\hat{\theta})$.

Since $\mathrm{SC}^\alpha(\pi) = \mathrm{SORT}(\hat{\mu})$ and $\mathrm{MM}^{\phi^\alpha}(\pi) = \mathrm{SORT}(\hat{\theta})$, we just need to show that $\mathrm{SORT}(\hat{\mu}) = \mathrm{SORT}(\hat{\theta})$. From a well-known result in exponential family, we know that $\hat{\theta} = (\nabla_\theta A)^{-1}(\hat{\mu})$. First, we need to show that the inverse exists. For this, it is important to note (and it is easy to check) that $\sum_i \hat{\mu}_i = \sum_i \alpha_i$. Second, there may be multiple (and in fact for scoring rules there are infinitely many) $\hat{\theta}$ for each $\hat{\mu}$. Thus, we prove the following two results.

Lemma 1. *If $\sum_i \mu_i = \sum_i \alpha_i$, then there exists a θ such that $\nabla_\theta A(\theta) = \mu$.*

Proof. **TODO** ☐ (Proof of Lemma 1)

Lemma 2. *Let $\hat{\mu}$ denote the MLE mean parameter. Let $\hat{\theta}$ be* any *MLE natural parameter that maps to $\hat{\mu}$. Then, SORT($\hat{\theta}$) = SORT($\hat{\mu}$).*

Proof. We know that $\nabla_\theta A(\hat{\theta}) = \hat{\mu}$. Note that

$$A(\theta) = \log \sum_{\sigma \in \mathcal{L}(A)} \exp\{\langle \theta, \phi^\alpha(\sigma) \rangle\}.$$

Hence,

$$\hat{\mu}_i = (\nabla_\theta A)_i(\hat{\theta}) = \frac{\sum_{\sigma \in \mathcal{L}(A)} \exp\{\langle \theta, \phi^\alpha(\sigma) \rangle\} \cdot \phi_i^\alpha(\sigma)}{\sum_{\sigma \in \mathcal{L}(A)} \exp\{\langle \theta, \phi^\alpha(\sigma) \rangle\}}.$$

To prove that $\mathrm{SORT}(\hat{\theta}) = \mathrm{SORT}(\hat{\mu})$, it is enough to show that for any i, j, $\hat{\theta}_i > \hat{\theta}_j$ implies $\hat{\mu}_i > \hat{\mu}_j$. Assume for some i and j, we have $\hat{\theta}_i > \hat{\theta}_j$. We want to show that $\hat{\mu}_i > \hat{\mu}_j$, i.e., $\hat{\mu}_i - \hat{\mu}_j > 0$. Now,

$$\hat{\mu}_i - \hat{\mu}_j = \frac{\sum_{\sigma \in \mathcal{L}(A)} \exp\{\langle \theta, \phi^\alpha(\sigma) \rangle\} \cdot (\phi_i^\alpha(\sigma) - \phi_j^\alpha(\sigma))}{\sum_{\sigma \in \mathcal{L}(A)} \exp\{\langle \theta, \phi^\alpha(\sigma) \rangle\}}. \tag{1}$$

Thus, $\hat{\mu}_i - \hat{\mu}_j > 0$ if and only if the numerator in Equation (1) is positive. For any ranking σ, let $\sigma_{i\leftrightarrow j}$ denote the ranking which is obtained by swapping alternatives i and j in σ. Similarly, for any natural parameter θ, let $\theta_{i\leftrightarrow j}$ denote the vector obtained by swapping the i^{th} and j^{th} coordinates of θ. Now,

$$
\begin{aligned}
&\sum_{\sigma\in\mathcal{L}(A)} e^{\langle\theta,\phi^\alpha(\sigma)\rangle}\cdot(\phi_i^\alpha(\sigma)-\phi_j^\alpha(\sigma)) \\
&= \sum_{1\le l<k\le m}\ \sum_{\substack{\sigma\in\mathcal{L}(A)\text{ s.t.}\\ \sigma(i)=l,\\ \sigma(j)=k}} \left(e^{\langle\theta,\phi^\alpha(\sigma)\rangle}\cdot(\alpha_l-\alpha_k)+e^{\langle\theta,\phi^\alpha(\sigma_{i\leftrightarrow j})\rangle}\cdot(\alpha_k-\alpha_l)\right) \\
&= \sum_{1\le l<k\le m}\ \sum_{\substack{\sigma\in\mathcal{L}(A)\text{ s.t.}\\ \sigma(i)=l,\\ \sigma(j)=k}} \left(e^{\langle\theta,\phi^\alpha(\sigma)\rangle}-e^{\langle\theta,\phi^\alpha(\sigma_{i\leftrightarrow j})\rangle}\right)\cdot(\alpha_l-\alpha_k) \\
&= \sum_{1\le l<k\le m}\ \sum_{\substack{\sigma\in\mathcal{L}(A)\text{ s.t.}\\ \sigma(i)=l,\\ \sigma(j)=k}} \left(e^{\langle\theta,\phi^\alpha(\sigma)\rangle}-e^{\langle\theta_{i\leftrightarrow j},\phi^\alpha(\sigma)\rangle}\right)\cdot(\alpha_l-\alpha_k) \\
&= \sum_{1\le l<k\le m}\ \sum_{\substack{\sigma\in\mathcal{L}(A)\text{ s.t.}\\ \sigma(i)=l,\\ \sigma(j)=k}} e^{\langle\theta_{-\{i,j\}},\phi^\alpha_{-\{i,j\}}(\sigma)\rangle}\cdot\left(e^{\theta_i\cdot\alpha_l+\theta_j\cdot\alpha_k}-e^{\theta_i\cdot\alpha_k+\theta_j\cdot\alpha_l}\right)\cdot(\alpha_l-\alpha_k) \\
&> 0.
\end{aligned}
$$

Here, the first transition follows by conditioning on the positions of alternatives i and j. The third transition follows since swapping alternatives i and j swaps the i^{th} and j^{th} coordinates in $\phi^\alpha(\sigma)$, which is further equivalent to swapping the i^{th} and j^{th} coordinates in θ (this retains the dot product intact). The fourth transition follows by taking all terms of the dot product except those from coordinates i and j out in common. Finally, the last transition follows since we have the following three conditions.

1. $\hat{\theta}_i > \hat{\theta}_j$.
2. $\alpha_l \ge \alpha_k$ for all $l < k$.
3. $\alpha_l > \alpha_k$ for some $l < k$.

Note that the first two conditions imply that $\theta_i\cdot\alpha_l+\theta_j\cdot\alpha_k \ge \theta_i\cdot\alpha_k+\theta_j\cdot\alpha_l$ for all $l<k$, and the first and the third conditions together imply that $\theta_i\cdot\alpha_l+\theta_j\cdot\alpha_k > \theta_i\cdot\alpha_k+\theta_j\cdot\alpha_l$ for some $l<k$.

Thus, we have $\mathrm{SORT}(\hat{\theta}) = \mathrm{SORT}(\hat{\mu})$, as required. $\square$ (Proof of Lemma 2)

With Lemma 2, we conclude that $\mathrm{MM}^{\phi^\alpha}(\pi) = \mathrm{SC}^\alpha(\pi)$. Since this holds for all profiles π, we have that $\mathrm{MM}^{\phi^\alpha} = \mathrm{SC}^\alpha$. $\square$ (Proof of Theorem 1)

2 Questions

1. First, it is easy to see that all positional scoring rules as well as the Kemeny rule has the form $\arg\max_\sigma \langle \sum_{i=1}^n \phi(\sigma_i), \phi(\sigma) \rangle$, with very natural ϕ.

   (a) What other rules can be represented in this form?

   (b) This is highly reminiscent of GSRs. In fact, all rules of this form are GSR with $f = \phi$ and $g = \arg\max_\sigma \langle \cdot, f(\sigma) \rangle$. Thus, they satisfy the properties that GSRs satisfy - anonymity and finite local consistency. Is there anything else that this intuitive family shares?

2. Why can't we represent Mallows' with varying dispersion parameter by completely incorporating that inside θ? Essentially, if we take $\theta = \lambda \cdot \phi(\sigma^*)$, with ϕ being the pairwise comparison representation, then we recover Mallows' with ground truth σ^* and dispersion parameter λ. Now we can learn them together.

3. Note that both pairwise comparison representation and scoring rules representation has overcomplete representation. In fact, specifically, they have the property that $\sum_i \phi_i(\sigma)$ is constant. Is this something inherent to voting? Do the representations for other rules satisfy this? In this particular case, is there a way of converting it to minimal representation?

4. It does not seem easy to do so with positional scoring rules, when you don't know which score vector is to be applied. That is, there does not seem to be any easy way of incorporating the scores inside θ as well. But can we still do something more general than a specific scoring rule? One way is simple: go to m^2 dimensional representation where there is a binary coordinate for each alternative in each positiion. This is highly overcomplete representation.

5. When does $\arg\max_\sigma \langle \hat{\theta}, \phi(\sigma) \rangle = \arg\max_\sigma \langle \hat{\mu}, \phi(\sigma) \rangle$? This is particularly more important if $\arg\max_\sigma \langle \hat{\theta}, \phi(\sigma) \rangle = \mathrm{SORT}(\hat{\theta})$ since in that case we can just show $\hat{\theta}_i > \hat{\theta}_j$ implies $\hat{\mu}_i > \hat{\mu}_j$ as we did for positional scoring rules.