MLE Voting Rules via Bregman Divergence

June 14, 2013

1 Exponential Families

To begin we have

- $\mathcal{X}$ is the set of outcomes (in general, can be discrete or continuous).
- ν is a base measure over the outcomes (usually counting or Lebesgue).
- $\phi : \mathcal{X} \to \mathcal{R}^k$ is a "sufficient statistic"—a vector-valued random variable.

An *exponential family* is a collection of probability densities over $\mathcal{X}$, with respect to ν, that take the form

$$p(x;\theta) = \exp\{\langle \theta, \phi(x) \rangle - A(\theta)\}. \tag{1}$$

Here $\theta \in \mathcal{R}^k$ is the *natural parameter*. The *cumulant* function is a normalizer:

$$A(\theta) = \log \int_{\mathcal{X}} \exp\{\langle \theta, \phi(x) \rangle\} \, d\nu(x). \tag{2}$$

The domain of A is $\Theta = \{\theta \in \mathcal{R}^k : A(\theta) < +\infty\}$. Both A and its domain Θ are convex. It is well-known that exponential families correspond to maximum entropy distributions. The entropy of a density is

$$H(p) = \int_{\mathcal{X}} p(x) \log p(x) \, d\nu(x) \tag{3}$$

Now suppose we maximize entropy subject to the constraint that the average sufficient statistic is $\mu \in \mathcal{R}^k$:

$$\int_{\mathcal{X}} \phi(x) \, p(x) d\nu(x) = \mu. \tag{4}$$

Then letting $\theta \in \mathcal{R}^k$ be the Lagrange multiplier vector for constraint (4), the solution takes the form (1). This shows that an exponential family has two parametrizations: the *natural* parametrization in terms of θ, and the *mean* parametrization in terms of μ.

2 Bregman Divergence

Associated with the convex function A is its *convex conjugate*

$$A^*(\mu) = \sup_{\theta' \in \Theta} \{\langle \theta', \mu \rangle - A(\theta')\}, \tag{5}$$

which is also a convex function. The *Bregman divergence* associated with a differentiable convex function f is $D(y, z) = f(y) - f(z) - \langle \nabla f(z), y - z \rangle$. Note that this distance function is convex in its first argument (but not necessarily its second), non-negative, and is zero only if $y = z$ when f is strictly convex. It is not necessarily symmetric, but it does satisfy some generalized version of the triangle inequality.

Now suppose θ and μ and the natural and mean parameters for a density from an exponential family. Assuming some unrestrictive conditions on ϕ (that it is a *minimal* statistic), it turns out we have $\mu = \nabla A(\theta)$ and $\theta = \nabla A^*(\mu)$. In particular the maximum in (5) is achieved at $\theta' = \theta$. Therefore, we have

$$\begin{aligned} \log p(x; \theta) &= \langle \theta, \phi(x) \rangle - A(\theta) \\ &= \langle \theta, \phi(x) - \mu \rangle + \langle \theta, \mu \rangle - A(\theta) \\ &= \langle \nabla A^*(\mu), \phi(x) - \mu \rangle + A^*(\mu) \\ &= -D_{A^*}(\phi(x), \mu) + A^*(\phi(x)) \end{aligned}$$

Therefore, the mean parametrization of the density takes the form

$$p(x; \mu) = \exp\{-D_{A^*}(\phi(x), \mu) + A^*(\phi(x)).\} \tag{6}$$

A voting rule according to this kind of parametrized distribution would proceed by first finding the MLE of μ (or θ) given the agents' reported rankings, and then computing the mode of the associated distribution.

3 MLE-MODE Voting

Recall the Mallows model for the distribution over permutations given a mode permutation π_0 and scalar parameter θ.

$$P(\sigma; \lambda, \sigma^*) = \exp\{-\lambda d(\sigma, \sigma^*)\}/\psi(\lambda, \sigma^*). \tag{7}$$

Here d is a distance function, which should satisfy $d(\pi, \sigma) \geq 0$ with equality if and only if $\pi = \sigma$. The remaining properties assumed of d vary in the literature. Sometimes it satisfies symmetry and the triangle inequality (so that it becomes a metric), but not always. Given this distribution, votes from agents are treated as sample points and the associated voting rule selects the pi_0 that is an MLE with respect to the sample.

There are many similarities in comparing (7) with (1) and (6) but also important differences. In (7) λ (a scalar "dispersion" parameter) and σ^* (the true

ranking ranking) are the underlying parameteres, while the natural underlying parameter θ in (1) is a vector parameter. Although Mallows' model can be recovered (if the distance function can be represented as a dot product) by incorporating both λ and σ^* inside θ, this only covers a sub-space of the distributions given by (1). Note that the functions $\psi(\lambda, \sigma^*)$ and $A(\theta)$ both serve as normalizers and are basically equivalent up to a log transformation.

A key difference, as pointed out above, is that the Mallows model (7) has discrete parameter space for its parameter σ^* while (1) and (6) have convex parameter sets. This is crucial because for an exponential family, the MLE of the μ parameter given a set of samples $x_1, \ldots, x_n$ is just $\hat{\mu} = \sum_{i=1}^{n} \phi(x_i)$. Deriving the natural parameter from the mean parameter is a nontrivial computational problem. However, the natural parametrization is convenient for then computing the mode, because this reduces to the linear programming problem of maximizing $\langle \theta, y \rangle$ where $y \in \mathcal{M} = \text{co}\{\phi(x) : x \in X\}$, where here X is the set of permutations of the candidates. There is a very large amount of work in the machine learning literature on graphical models on methods to efficiently compute the natural parametrization from the mean parametrization, and vice-versa, which we could draw on to implement our version of voting with exponential families and Bregman divergences. This method involves first computing the MLE natural parameter and then the mode ranking of the distribution given by the MLE natural parameter; we refer to this as the MLE-MODE method.

Another important difference is that (1) and (6) are both defined with respect to a certain representation $\phi(x)$ of rankings in Euclidean space $\mathcal{R}^k$, while (7) is defined directly on rankings. The exponential family approach opens the door to associating properties of the voting rule with the structure of ϕ.

4 Notation

- Let $C = \{1, \ldots, m\}$ denote the set of alternatives and $\mathcal{L}(C)$ denote the set of all linear orders over C (votes). Thus, $|\mathcal{L}(C)| = m!$.
- We view a ranking $\sigma \in \mathcal{L}(C)$ as $\sigma : C \to \{1, \ldots, m\}$. We use $\pi \in \mathcal{L}(C)^n$ to denote a profile of n votes.
- Given a representation $\phi : \mathcal{L}(C) \to \mathbb{R}^k$, 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.
- Unless mentioned otherwise, the performance of the rules would not depend on the tie-breaking for the choice of the MLE parameter, or for the choice of the mode ranking.
- Let $\hat{\mu}$ and $\hat{\theta}$ denote the MLE mean and natural parameters respectively.

5 Scoring Representation

For working with scoring rules, we introduce some additional notation.

- 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 $\alpha_1 > \alpha_m$, otherwise the rule is meaningless.
- We denote the scoring rule associated with score vector α by SC^α.
- For any score vector α, let $\phi^\alpha : \mathcal{L}(C) \to \mathbb{R}^m$ be the representation such that $\phi^\alpha_i(\sigma) = \alpha_{\sigma(i)}$ for all σ and i.
- Finally, let $\mathrm{SORT} : \mathbb{R}^m \to \mathcal{L}(C)$ denote the function that takes an m-dimensional vector, and returns the sorted order of indices. That is, alternative i is mapped to position j in $\mathrm{SORT}(v)$ if $|\{k | v_k > v_i\}| = j - 1$. We break ties arbitrarily, as they do not matter for our results.

5.1 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 the MLE parameter and the tie-breaking for the mode ranking.*

Proof. Fix arbitrary score vector α and any profile $\pi = (\sigma_1, \ldots, \sigma_n)$. We want to show that $\mathrm{MM}^{\phi^\alpha}(\pi) = \mathrm{SC}^\alpha(\pi)$. First, 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 the score vector α. Hence, we clearly have $\mathrm{SC}^\alpha(\pi) = \mathrm{SORT}(\hat{\mu})$.

On the other hand, if $\hat{\theta}$ denotes some MLE natural parameter, then its 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 coordinates 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 $\mathrm{SORT}(\hat{\mu}) = \mathrm{SORT}(\hat{\theta})$. Our proof in fact shows that $\hat{\mu}_i = \hat{\mu}_j$ if and only if $\hat{\theta}_i = \hat{\theta}_j$, which takes care of the tie-breaking issues in SORT.

Now, 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. *There exists a $\hat{\theta}$ such that $\nabla_\theta A(\hat{\theta}) = \hat{\mu}$.*

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

Proof of Lemma 1. **TODO** □ (Proof of Lemma 1)

Proof of Lemma 2. We know that $\nabla_\theta A(\hat{\theta}) = \hat{\mu}$. Note that

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

Hence,

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

To prove that $\text{SORT}(\hat{\theta}) = \text{SORT}(\hat{\mu})$, it is enough to show that for any i, j, $\hat{\theta}_i > \hat{\theta}_j$ implies $\hat{\mu}_i > \hat{\mu}_j$. By symmetry, this implies $\hat{\theta}_i > \hat{\theta}_j$ if and only if $\hat{\mu}_i > \hat{\mu}_j$, which also implies that $\hat{\theta}_i = \hat{\theta}_j$ if and only if $\hat{\mu}_i = \hat{\mu}_j$. While the former is useful for the result to be independent of MLE natural parameter selection, the latter is useful for the result to be independent of the tie-breaking in mode selection.

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,

$$\hat{\mu}_i - \hat{\mu}_j = \frac{\sum_{\sigma \in \mathcal{L}(C)} \exp\{\langle \theta, \phi^\alpha(\sigma) \rangle\} \cdot (\phi_i^\alpha(\sigma) - \phi_j^\alpha(\sigma))}{\sum_{\sigma \in \mathcal{L}(C)} \exp\{\langle \theta, \phi^\alpha(\sigma) \rangle\}}. \tag{8}$$

Thus, $\hat{\mu}_i - \hat{\mu}_j > 0$ if and only if the numerator in Equation (8) 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 values of the i^{th} and j^{th} coordinates of θ. Now,

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

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. For the final transition, note that we have $\hat{\theta}_i > \hat{\theta}_j$. Now, for a weak inequality note that for all $l < k$, we have $\alpha_l \geq \alpha_k$. Thus, $\theta_i \cdot \alpha_l + \theta_j \cdot \alpha_k \geq \theta_i \cdot \alpha_k + \theta_j \cdot \alpha_l$. This equality becomes strict for $l = 1$ and $k = m$ since $\alpha_1 > \alpha_m$, proving the final transition.

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

TODO: To be completely formal, $\text{SORT}(x)$ needs to be the set of all tied rankings, and then we should show set equality. I will change this later.

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

6 Pairwise Comparison Representation

No inclusion relation between the sets of tied rankings returned by Kemeny and MM+PC methods.

With bad tie-breaking, violates monotonicity, IIA, and Majority $\rightarrow$ thus Condorcet consistency as well $\rightarrow$ **lose the bad tie-breaking part!!**

Violates PM-c irrespective of tie-breaking

Definitely anonymous, neutral

consistency?

7 Properties of MM rules

Always anonymous.

Consistency? Monotonicity? PM-c?

Condorcet consistency? Majority? $\rightarrow$ Although these two are not appropriate for SWF.

7.1 Neutrality and Linear Representations

Linear representation $\rightarrow$ Left invariance of Bregman divergence $\rightarrow$ Neutrality of distribution $\rightarrow$ Neutrality of MM method

8 Distance Function in Euclidean Embeddings

$d(\sigma_1, \sigma_2) = C - \langle \phi(\sigma_1), \phi(\sigma_2) \rangle$, where $C = \langle \phi(\sigma), \phi(\sigma) \rangle$ is a constant for all σ. If ϕ satisfies this, then it forms a distance metric. Essentially, we have $d(\sigma_1, \sigma_2) = (1/2) \cdot ||\phi(x) - \phi(y)||$. Also, if ϕ is a linear representation, then d is left-invariant too.

9 Questions

1. Consistency, monotonicity, and PM-c of MM method?
2. What about Condorcet consistency and Majority? Although, they are not suited for SWFs.
3. Kemeny $\in$ MM+PC? Also, show comparison lemma for MM+PC to prove that it is PM-c.
4. Investigate MM+PC counter example to Kemeny and why only the comaprison between a and c matters.
5. Investigate the distribution of MLE natural parameter as a randomized voting rule. That is, the distribution we get right before taking mode.
6. Find motivation for exponential family. Subjective preference aggregation -¿ must have nice properties. Ground truth discovery -¿ must model properly what the ground truth is, and how our method is finding it. Even showing that given infinite samples, our method would find the ground truth is enough.
   (a) Can we show something like the following? "If the votes are drawn from any distribution out of a family $\mathcal{F}$ of distributions, then fitting Mallows' model and finding MLE ground truth will find the actual ground truth with probability 1 given infinite samples."
   (b) What about exponential family?
7. Mean affine voting rules and connection to the exponential distribution with distance function given in Section 8.
   (a) Also, their connection to GSRs. Taking $f = \phi$ and $g = \arg\max_\sigma \langle \cdot, f(\sigma) \rangle$ recovers a mean affine voting rule. But g doesn't always depend on pairwise comparisons in this case.
   (b) Notice that GSRs nicely combine scoring and comparing.
8. Including a family of voting rules
   (a) All Mallows': Pairwise comparison representation and $\theta = \lambda \cdot \phi(\sigma^*)$.
   (b) Scoring rules: Need to move to m^2-dimensional representation where there is a binary coordinate for each alternative in each positiion. In this case, the $\hat{\theta} = [\alpha; \alpha; \ldots; \alpha]$ for the score vector α. This set is convex. Can we use exponential family literature to learn the scoring rule?
9. What conditions on ϕ ensure $\arg\max_\sigma \langle \hat{\theta}, \phi(\sigma) \rangle = \arg\max_\sigma \langle \hat{\mu}, \phi(\sigma) \rangle$?
   (a) In particular, it would be nice if $\arg\max_\sigma \langle \hat{\theta}, \phi(\sigma) \rangle = \text{SORT}(\hat{\theta})$ (and same for $\hat{\mu}$) since in that case we can show comparison lemma as we did for scoring rules.

10. Structure of ϕ to get IIA, and alternate proof of Arrow's theorem for this restricted class of voting rules.

A Generalized Entropy

The concepts above were specifically linked to the entropy function (3), but they can be generalized to any convex function over the set of probability densities, which is then construed as a generalized (negative) entropy function.

Let $\mathcal{P}$ denote the set of probability densities over $\mathcal{X}$ (with respect to base measure ν). Let H be a strictly convex function over the convex domain $\mathcal{P}$. To this convex function we can associate a *scoring rule* (not to be confused with position scores in a voting context) defined over $\mathcal{X} \times \mathcal{P}$:

$$S(x, p) = H(p) + \langle \nabla H(p), \mathbf{1}_x - p \rangle,$$

where $\mathbf{1}_x$ is the probability density that puts mass 1 on outcome x. It can be shown that if an agent is compensated with $S(x, q)$ when it reports $q \in \mathcal{P}$ and $x \in \mathcal{X}$ occurs, then its optimal strategy is to report its actual belief (i.e., the actual probability density it has in mind). Hence the term "scoring rule". For instance, if H is negative entropy then $S(x, p) = \log p(x)$.

To formulate the generalized exponential family, we then take the family of densities that satisfy

$$S(x, p) = \alpha + \langle \theta, \phi(x) \rangle. \tag{9}$$

Namely, there are $\alpha \in \mathcal{R}$ and $\theta \in \mathcal{R}^k$ so that the score $S(x, p)$ is a linear function of $\phi(x)$ as above. This gives the "natural" parametrization of p. Taking H as negative entropy, this condition becomes $\log p(x) = \alpha + \langle \theta, \phi(x) \rangle$, so that p takes the form of an exponential family.

The density p has an alternate characterization as the distribution that minimizes $H(p)$, subject to the constraint that the expectation of $\phi(x)$ under p is μ. This is the "mean" parametrization. More explicitly, let p_μ be the density with mean μ that minimizes $H(p)$, and define the convex function $F(\mu) = H(p_\mu)$ over $\mathcal{M} = \text{co}\{\phi(x) : x \in \mathcal{X}\}$, where co denotes the convex hull. We take the family of densities that satisfy

$$S(x, p) = -D_F(\phi(x), \mu) + F(\phi(x)). \tag{10}$$

Note: still need to prove that (9) and (10) are equivalent characterizations. But I'm not sure that's true for any convex H.