Axioms for Learning from Pairwise Comparisons

Ritesh Noothigattu
Carnegie Mellon University
riteshn@cmu.edu

Dominik Peters
Carnegie Mellon University
dominikp@cs.cmu.edu

Ariel D. Procaccia
Harvard University
arielpro@seas.harvard.edu

Abstract

To be well-behaved, systems that process preference data must satisfy certain conditions identified by economic decision theory and by social choice theory. In ML, preferences and rankings are commonly learned by fitting a probabilistic model to noisy preference data. The behavior of this learning process from the view of economic theory has previously been studied for the case where the data consists of rankings. In practice, it is more common to have only pairwise comparison data, and the formal properties of the associated learning problem are more challenging to analyze. We show that a large class of random utility models (including the Thurstone–Mosteller Model), when estimated using the MLE, satisfy a Pareto efficiency condition. These models also satisfy a strong monotonicity property, which implies that the learning process is responsive to input data. On the other hand, we show that these models fail certain other consistency conditions from social choice theory, and in particular do not always follow the majority opinion. Our results inform existing and future applications of random utility models for societal decision making.

1 Introduction

More than two centuries ago, the marquis de Condorcet [5] suggested a statistical interpretation of voting. Each vote, Condorcet argued, can be seen as a noisy estimate of a ground-truth ranking of the alternatives. A voting rule should aggregate votes into a ranking that is most likely to coincide with the ground truth. Although Condorcet put forward a specific noise model, his reasoning applies to any *random noise model*, which is a distribution over votes parameterized by a ground truth ranking [6, 4].

Until NeurIPS 2014, this statistical approach to voting was studied in parallel to the more common normative approach, which evaluates voting rules based on axiomatic properties. But the two approaches converged in a paper by Azari Soufiani et al. [1], whose key idea was to determine whether maximum likelihood estimators (MLEs) for two noise models satisfy basic axiomatic properties. Their results were sharpened and extended by Xia [18].

Our point of departure is that instead of random noise models we consider *random utility models*, where each alternative x has a utility β_x, and the probability of drawing a pairwise comparison that puts x above y depends on β_x and β_y. For example, under the well-known Thurstone–Mosteller Model [17, 12], this pairwise comparison would be generated by sampling $u(x)$ and $u(y)$ from normal distributions with means β_x and β_y, respectively, and the same variance.

Our research question, then, is this:

> *For a given random utility model, consider the aggregation rule that takes pairwise comparisons between alternatives as input and returns the ranking over alternatives defined by the MLE; which axioms does it satisfy?*

1.1 Why Is the Research Question Important?

The MLE, as an aggregation rule, is *statistically* well-motivated. From a voter's perspective, though, it is not immediately clear the the MLE is a good rule that adequately aggregates preferences. In particular, in case the statistical assumptions of a random utility fail to capture reality, the MLE may give a bad result. However, if we can show that the MLE satisfies standard axioms from voting theory, this implies a certain degree of robustness. It also provides reassurance that the statistical process cannot fall prey to pathological behavior in edge cases.

A string of recent papers [7, 15, 10, 11] proposes a sequence of systems for automated societal decision making through social choice and machine learning. They all aggregate pairwise comparisons, by fitting them to random utility models. As these systems are being deployed to support decisions in non-profits and government, it becomes crucial to understand normative properties of this framework.[1]

The work of Freedman et al. [7] provides a concrete illustration. The paper deals with prioritization of patients in kidney exchange. They asked workers on Amazon Mechanical Turk to decide which of each given pair of patients with chronic kidney disease (defined by their medical profiles) should receive a kidney, and computed the MLE utilities assuming the pairwise comparisons were generated by a Bradley–Terry model [3]. The resulting ranking over profiles was used to help a kidney exchange algorithm prioritize some patients over others. As is common, the authors pooled pairwise comparisons reported by many different voters, and fit a single random utility model to the pool, as opposed to fitting models to individual voters. This move is usually done to improve statistical accuracy, but, to an extent, it invalidates the underlying motivation of the noise model, which imagines a single decision maker with imprecise perception of utilities. Pooling assumes that a group of agents can be captured by the same model. Given this leap of faith, we believe that a normative analysis of the process becomes especially important.

Other papers apply an emerging approach called *virtual democracy* [10] to automate decisions in two domains, autonomous vehicles [15] and food allocation [11]. Lee et al. [11] asked stakeholders in a nonprofit food rescue organization to report which of each given pair of recipient organizations should be allocated an incoming food donation. Unlike Freedman et al. [7], they fit a random utility model (Thurstone–Mosteller) to the pairwise comparisons provided by each stakeholder *individually*, and used the Borda voting rule to aggregate the predictions given by each of the individual models. On the one hand, our axiomatic results may justify a move to the pooled approach of Freedman et al. [7], which could improve accuracy. On the other hand, even when learning individual models, axiomatics can convince voters that their preferences are learned using a sensible method.

1.2 Our Results

We examine four axiomatic properties, suitably adapted to our setting. Informally, they are:

- *Pareto efficiency:* If x dominates y in the input dataset, x should be above y in the MLE ranking.
- *Monotonicity:* Adding $a \succ b$ comparisons to the input dataset can only help a and harm b in the MLE ranking.
- *Pairwise majority consistency:* If the input dataset is consistent with a ranking over the alternatives, that ranking must coincide with the MLE ranking.
- *Separability:* If a is preferred to b in the MLE rankings of two different datasets, a must also be preferred to b in the MLE ranking of the combined dataset.

The first two properties, Pareto efficiency and monotonicity, have immediate appeal and seem crucial: a system violating these is not faithful to input preferences. In Sections 3 and 4 we show that both properties are satisfied by a large class of random utility models when fitted using MLEs. For monotonicity, our main result, the proof is surprisingly involved, since we need to reason about the optimum utility values of all alternatives simultaneously. (In contrast, for random noise models, monotonicity is a simple consequence of the definition [1, 18].)

The latter two properties are *not* satisfied by MLEs, for all random utility models satisfying mild conditions. In a way, these negative results illuminate the behavior of random utility models: The

[1] Previous work [1, 18] does not apply as it focuses on the aggregation of input *rankings* (a special case of pairwise comparisons) through random noise models.

case of pairwise majority consistency illustrates a trade-off, where random utility models ensure that a strong preference is respected, even if this leads them to override a majority preference elsewhere. While negative, we do not see the counterexamples in Sections 5 and 6 as pathological, though they may suggest contexts in which the use of random utility models is not appropriate.

2 Model

Let $\mathcal{X}$ be a finite set of alternatives. For notational simplicity, we let $\mathcal{X}^2 = \{(x, y) : x, y \in \mathcal{X}, x \neq y\}$ denote the set of *distinct* pairs of alternatives. Let $\# : \mathcal{X}^2 \to \mathbb{N}$ be a *dataset* of pairwise comparisons between alternatives: For $x, y \in \mathcal{X}$, $\#\{x \succ y\}$ is the number of times x beat y in the dataset.

The (pairwise) *comparison graph* $\mathcal{G}_\# = (\mathcal{X}, E)$ with respect to dataset $\#$ is the directed graph with the alternatives $\mathcal{X}$ as the vertices, and edges E such that there exists a directed edge $(u, v) \in E$ iff $\#\{u \succ v\} > 0$. We say that $\mathcal{G}_\#$ is *connected* if its undirected form is connected, and we call it *strongly connected* if for all $(x, y) \in \mathcal{X}^2$, there is a directed path from x to y in $\mathcal{G}_\#$.

Given a dataset, our goal is to learn a random utility model (RUM). A random utility model specifies, for any two distinct alternatives $x, y \in \mathcal{X}$, the probability that when asking the decision maker to compare x and y, the answer will be $x > y$. (Due to noise, when repeatedly querying the same pair, we may see different answers.) For us, a random utility model is parameterized by a vector $\beta \in \mathbb{R}^{\mathcal{X}}$, where β_x is an unknown utility value for $x \in \mathcal{X}$. When we ask for a comparison between two alternatives $x, y \in \mathcal{X}$, we model the decision maker as sampling noisy utilities $u(x)$ and $u(y)$ from distributions parameterized by (and typically centered at) β_x and β_y. Then, the decision maker reports the comparison $x > y$ iff $u(x) > u(y)$.

In this paper, we focus on random utility models with i.i.d. noise, so that $u(x) = \beta_x + \zeta(x)$, where $\zeta(x) \sim \mathcal{P}$ is i.i.d. across all alternatives. Let F be the CDF of a random variable which is the difference between two independent random variables with distribution $\mathcal{P}$. Then the probability that alternative x beats y when they are compared is[2]

$$\Pr(x \succ y) = \Pr(u(x) > u(y)) = \Pr(\zeta(y) - \zeta(x) < \beta_x - \beta_y) = F(\beta_x - \beta_y). \tag{1}$$

We derived Equation (1) from a specific noise model, but it makes sense for any function $F : \mathbb{R} \to [0, 1]$ with CDF-like properties, even if it does not correspond to a noise distribution $\mathcal{P}$. Indeed, we can take any F which is non-decreasing, satisfies $F(\Delta u) + F(-\Delta u) = 1$ for all $\Delta u \in \mathbb{R}$, and is such that $\lim_{\Delta u \to -\infty} F(\Delta u) = 0$ and $\lim_{\Delta u \to \infty} F(\Delta u) = 1$. We adopt Equation (1) as the general definition of a random utility model for our technical results.

Two of the most common random utility models are

- the *Thurstone–Mosteller (TM) model*: We sample utility as $u(x) = \beta_x + \zeta(x)$, with i.i.d. noise $\zeta(x) \sim \mathcal{N}(0, 1/2)$. This is equivalent to Equation (1) with F as the Gaussian CDF Φ.
- the *Bradley–Terry model* (equivalent to the *Plackett–Luce* model restricted to pairwise comparisons), where $\Pr(x \succ y) = \frac{e^{u(x)}}{e^{u(x)} + e^{u(y)}}$. This is Equation (1) with F as the logistic function.

We usually assume that F is strictly log-concave, and that it is strictly monotonic and continuous,[3] so that F has an inverse on $(0, 1)$. These conditions hold for Thurstone–Mosteller and Bradley–Terry.

For a random utility model, given a dataset $\#$, our goal is to find parameters $(\beta_x)_{x \in \mathcal{X}}$ that best fit $\#$. We find these parameters by maximum likelihood estimation. The log-likelihood is given by

$$\mathcal{L}(\beta) = \sum_{(x,y) \in \mathcal{X}^2} \#\{x \succ y\} \log F(\beta_x - \beta_y).$$

When the dataset $\#$ is clear from the context, we write $\hat{\beta} \in \mathbb{R}^{\mathcal{X}}$ for a parameter vector that maximizes log-likelihood, and say that $\hat{\beta}$ is the *MLE*. Note that if $c \in \mathbb{R}$ is a scalar, then $\mathcal{L}(\beta) = \mathcal{L}(\beta + c)$ for all $\beta \in \mathbb{R}^{\mathcal{X}}$ (since $\Pr(x \succ y)$ depends only on the difference $\beta_x - \beta_y$), so the MLE is only defined up to an additive shift. For concreteness, we pick some $r \in \mathcal{X}$, call it the *reference alternative*, and fix $\beta_r = 0$; then, we maximize $\mathcal{L}$ over $\mathcal{D} = \{\beta \in \mathbb{R}^{\mathcal{X}} : \beta_r = 0\}$.

[2] We assume $\mathcal{P}$ to be a continuous distribution, and so we do not have to worry about ties.

[3] Continuity of F is guaranteed when the corresponding noise distribution $\mathcal{P}$ is continuous.

A random utility model is particularly appropriate when the dataset $\#$ consists of pairwise comparisons which are all reported by a single decision maker. However, in many cases the dataset is obtained by pooling reports from many agents, for instance to minimize the labeling effort of each individual agent, or if we have the explicit aim to aggregate preferences from different agents. Some of the axioms we study are explicitly motivated by cases where $\# = \sum_{i \in \mathcal{R}} \#_i$, i.e., the dataset is obtained by pooling individual datasets, where $\mathcal{R}$ is the set of agents. It then seems natural to assume that each agent behaves in accordance with some random utility model with unknown parameters β^i and unknown CDF-like function F_i. Then the dataset $\#_i$ is generated by repeatedly querying the agent's random utility model for a comparison.

2.1 Existence and Boundedness of MLE

Before turning to our main results, we briefly state conditions that guarantee the existence of a finite MLE, and that guarantee uniqueness (up to a shift).

In some scenarios, no finite β maximizes likelihood, and thus the MLE may not exist. For instance, if some alternative a beats other alternatives, but is not beaten even a single time in the dataset, the likelihood can always be strictly increased by increasing β_a (when F is strictly monotonic). Lemma 2.1 states a condition under which an MLE exists (i.e. $\mathcal{L}(\beta)$ has a maximizer). Its proof also provides a weak bound on one such maximizer. The proofs of the results in this section are in Appendix A.

Lemma 2.1 (MLE exists)**.** *Suppose F is strictly monotonic and continuous. Then the MLE exists if and only if every connected component of the comparison graph $\mathcal{G}_\#$ is strongly connected.*

For alternative $x, y \in \mathcal{X}$, we define the *perfect-fit distance* between x and y as

$$\delta(x, y) := F^{-1}\left(\frac{\#\{x \succ y\}}{\#\{x \succ y\} + \#\{y \succ x\}}\right).$$

This is the difference in utilities of x and y required for the model to exactly match the observed frequencies of $\#\{x \succ y\}$ and $\#\{y \succ x\}$ in the data. We can check that the MLE will respect this perfect-fit distance when an alternative has only a single neighbor in the comparison graph.

Lemma 2.2. *Let F be strictly monotonic and continuous. Suppose that for alternative a there is exactly one alternative b for which $\#\{a \succ b\} + \#\{b \succ a\} > 0$. If both $\#\{a \succ b\} > 0$ and $\#\{b \succ a\} > 0$, then any MLE $\hat{\beta}$ satisfies $\hat{\beta}_a - \hat{\beta}_b = \delta(a, b)$.*

We can use this result to provide a stronger bound on the MLE than the one from Lemma 2.1, which holds under slightly stronger conditions.

Lemma 2.3. *Suppose that $\#\{x \succ y\} > 0$ and $\#\{y \succ x\} > 0$ for all x and y, and that F is continuous and strictly monotonic. Then for every MLE $\hat{\beta}$ we have the bound*

$$\|\hat{\beta}\|_\infty \leq |\mathcal{X}| \cdot \max_{(x,y) \in \mathcal{X}^2} \delta(x, y).$$

2.2 Uniqueness of MLE

Under mild conditions on the function F and the comparison graph $\mathcal{G}_\#$, we have seen that a bounded MLE exists. When is the MLE unique? Note that if F is a strictly log-concave, this implies that the log-likelihood $\mathcal{L}(\beta)$ is concave. If we additionally require that the comparison graph $\mathcal{G}_\#$ is connected, then $\mathcal{L}(\beta)$ is in fact strictly concave, and thus the MLE is unique, as we prove in Appendix B.

Lemma 2.4. *Suppose that F is strictly log-concave. Then $\mathcal{L}(\beta)$ is strictly concave and the MLE is unique (assuming it exists), if and only if the comparison graph $\mathcal{G}_\#$ is connected.*

3 Pareto Efficiency

A minimal requirement in economic theory is *Pareto efficiency*: if all agents prefer a to b, then in aggregate, a should be preferred to b. A first attempt at defining this notion for the environment of pairwise comparisons would be to say that if $\#\{a \succ b\} > 0$ but $\#\{b \succ a\} = 0$, then the MLE should satisfy $\hat{\beta}_a \geq \hat{\beta}_b$. However, this property is too restrictive. Consider a dataset with

$$\#\{a \succ b\} = 100, \#\{b \succ c\} = 1, \#\{c \succ a\} = 1,$$

and all other comparisons 0. To satisfy the mentioned property, the MLE would need to satisfy $\hat{\beta}_a \geq \hat{\beta}_b \geq \hat{\beta}_c \geq \hat{\beta}_a$, so they are all equal; however it seems better to have $\hat{\beta}_a > \hat{\beta}_b$.

A more sensible version of Pareto efficiency is motivated by the multi-agent setting described in Section 2, where $\# = \sum_{i \in \mathcal{R}} \#_i$, and each individual dataset $\#_i$ is generated by a random utility model with unknown parameters β^i. In this case, Pareto efficiency should say that if $\beta^i_a > \beta^i_b$ for all $i \in \mathcal{R}$, then the MLE $\hat{\beta}$ applied to dataset $\#$ should satisfy $\hat{\beta}_a > \hat{\beta}_b$ as well. Our official definition of Pareto efficiency implies this, but is phrased more generally.

Definition 3.1 (Pareto efficiency)**.** *Suppose $a, b \in \mathcal{X}$ satisfy $\#\{a \succ b\} > \#\{b \succ a\}$, and are such that for every other alternative $x \in \mathcal{X} \setminus \{a, b\}$, we have*

$$\#\{a \succ x\} > \#\{b \succ x\} \quad \textit{and} \quad \#\{x \succ a\} < \#\{x \succ b\}.$$

Then, Pareto efficiency requires that $\hat{\beta}_a \geq \hat{\beta}_b$.

To see that this definition captures the desired behavior in the multi-agent case, note that if $\beta^i_a > \beta^i_b$, then the dataset $\#_i$ satisfies the condition of Definition 3.1 with high probability as we grow the number of comparisons in $\#_i$, and similarly the condition holds for the pooled dataset $\# = \sum_{i \in \mathcal{R}} \#_i$.

This version of Pareto efficiency is feasible; in fact, it is satisfied by most random utility models.

Theorem 3.2. *Maximum likelihood estimation satisfies Pareto efficiency if F is strictly monotonic.*

The key idea behind the proof (given in Appendix C) is that if $a, b \in \mathcal{X}$ satisfy the condition of Definition 3.1 but some MLE $\hat{\beta}$ puts $\hat{\beta}_a < \hat{\beta}_b$, then the parameter vector β equal to $\hat{\beta}$ except that $\beta_a = \hat{\beta}_b$ and $\beta_b = \hat{\beta}_a$ has strictly higher log-likelihood.

4 Montonicity

If we add a pairwise comparison $a \succ b$ to a dataset, we should deduce that a is stronger and b is weaker relative to our previous estimates. It would be paradoxical if, upon seeing evidence that a is strong and b is weak, we decided to lower a's utility or increase b's utility. Monotonicity requires that this can never happen. We consider a strong form of this axiom, which requires that a is strengthened relative to *every* other alternative, and not just relative to b.

Definition 4.1 (Monotonicity)**.** *Suppose that $\#$ and $\tilde{\#}$ are two datasets with unique MLEs $\hat{\beta}$ and $\tilde{\beta}$. Suppose that $\tilde{\#}\{x \succ y\} = \#\{x \succ y\}$ for all $x, y \in \mathcal{X}$ except that $\tilde{\#}\{a \succ b\} > \#\{a \succ b\}$. Then, monotonicity requires that for all $x \in \mathcal{X}$,*

$$\tilde{\beta}_a - \tilde{\beta}_x \geq \hat{\beta}_a - \hat{\beta}_x \quad \textit{and} \quad \tilde{\beta}_b - \tilde{\beta}_x \leq \hat{\beta}_b - \hat{\beta}_x.$$

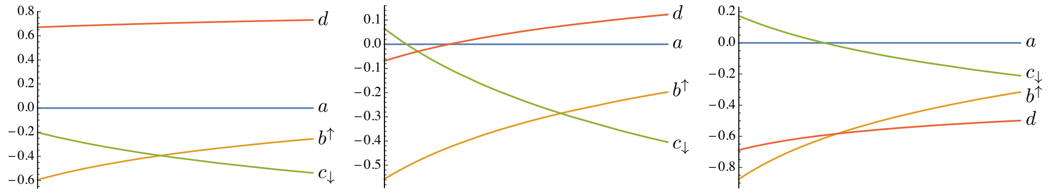
Equivalently, monotonicity requires that if $\#\{a \succ b\}$ *decreases*, then a becomes weaker relative to other alternatives, and b becomes stronger. We can interpret monotonicity as guaranteeing a kind of *participation incentive*: If we ask an agent to compare a to b, the agent is assured that the answer can only influence our inferred utilities in the desired direction.

Monotonicity is foundational to the idea of aggregating pairwise comparisons; in a sense, it encodes the proper meaning of a comparison "$a \succ b$". It may be surprising, then, that it is difficult to prove that MLEs of random utility models satisfy monotonicity.[4] While it is easy to check that the difference $\hat{\beta}_a - \hat{\beta}_b$ is increasing in $\#\{a \succ b\}$, it is much trickier to analyze the behavior of the log-likelihood for the positioning of alternatives other than a and b. However, it turns out that random utility models do satisfy the strong monotonicity axiom. Our proof depends crucially on the assumption that F is log-concave. Due to the conceptual importance of monotonicity, we consider this our main result.

Theorem 4.2. *Maximum likelihood estimation satisfies monotonicity if F is strictly monotonic, log-concave, and differentiable.*

The proof, given in Appendix D, is relatively unwieldy. For intuition, let us provide an outline of a proof for the special case of three alternatives a, b, c. Let $\#$ and $\tilde{\#}$ be datasets that are identical

[4] For the Bradley–Terry model, monotonicity is easier to check, since the first-order conditions of likelihood maximization in that model are well-behaved [8, Prop. 6.3]; that proof does not generalize to other models.



$(17, 86, 38, 56, -, 4, 19, 89, 1, 67, 83, 34)$ $(95, 74, 36, 53, -, 39, 62, 48, 1, 30, 77, 8)$ $(63, 54, 12, 19, -, 54, 57, 80, 49, 7, 50, 20)$

Figure 1: The MLE for Thurstone–Mosteller models is monotonic: with more $b \succ c$ comparisons, b's utility increases, while c's decreases. The vector shows the dataset $\#$ with $\mathcal{X}^2$ in lexic order.

except that $\#\{a \succ b\} < \tilde{\#}\{a \succ b\}$, and let $\hat{\beta}$ and $\tilde{\beta}$ be the respective MLEs, which are unique by Lemma 2.4. We take a as reference, so $\hat{\beta}_a = \tilde{\beta}_a = 0$. It is easy to see that $\hat{\beta}_b \geq \tilde{\beta}_b$, since otherwise $\hat{\beta}$ would have greater log-likelihood than $\tilde{\beta}$ for the dataset $\tilde{\#}$, as $\hat{\beta}$ performs better on the a vs b comparisons, and performs no worse on other comparisons by optimality for $\#$. To see that also $\hat{\beta}_c \geq \tilde{\beta}_c$, consider first the dataset $\#$ and associated log-likelihood $\mathcal{L}(\beta_b, \beta_c)$ (with β_a fixed to 0). Now, for $x \in \mathbb{R}$, let $\psi(x)$ denote the value of β_c that maximizes $\mathcal{L}(x, \beta_c)$, i.e., maximizes likelihood among parameters β with $\beta_a = 0$ and $\beta_b = x$. One can show that, since F is strictly log-concave, $\psi(x)$ is increasing in x.[5] Notice that the number of comparisons between a and b in a dataset does not influence the optimum position of β_c, once β_a and β_b are fixed. Hence, the function ψ is the same whether defined for $\#$ or for $\tilde{\#}$, since they only differ in a vs b comparisons. We have already seen that $\tilde{\beta}_b \leq \hat{\beta}_b$. Since ψ is increasing, we have $\tilde{\beta}_c = \psi(\tilde{\beta}_b) \leq \psi(\hat{\beta}_b) = \hat{\beta}_c$, proving monotonicity.

To visualize monotonicity, consider the three examples in Figure 1. For four alternatives, we generated random datasets by choosing $\#\{x \succ y\}$ uniformly at random between 1 and 100, and picked three examples. In each case, we let $\#\{b \succ c\}$ vary from 0 to 100 (going horizontally from left to right), and show how the MLE of the Thurstone–Mosteller model changes as the number of $b \succ c$ comparisons grows; we fix $\hat{\beta}_a = 0$ as reference. As predicted by Theorem 4.2, the orange line of $\hat{\beta}_b$ is increasing, while the green line of $\hat{\beta}_c$ is decreasing. Note that the change in $\#\{b \succ c\}$ can affect other alternatives; in the middle figure, the relative positions of a and d swap.

In the pooled setting $\# = \sum_{i \in \mathcal{R}} \#_i$ of Section 2, where each agent $i \in \mathcal{R}$ is described by a random utility model with parameters β^i that generates $\#_i$, a natural notion of monotonicity is this: Suppose we calculate the MLE $\hat{\beta}$ for $\#$ and suppose we increase the utility β^i_a for some agent i and some alternative a while keeping all other parameters fixed. Then the updated MLE $\tilde{\beta}$ should satisfy $\tilde{\beta}_a - \tilde{\beta}_x \geq \hat{\beta}_a - \hat{\beta}_x$ for all $x \in \mathcal{X}$: the learned utility of a increases relative to other alternatives. Theorem 4.2 implies that random utility models (subject to the theorem's conditions) satisfy this pooled monotonicity notion with high probability, when $\#_i$ consists of many samples and when the number of comparisons is uniform across pairs. The reason is this: with high probability, the increase of β^i_a increases the number of $a \succ x$ comparisons in $\#_i$ for all x. Assuming for now that no other dataset $\#_j$ and no other pairs in $\#_i$ are affected, then successively invoking Theorem 4.2 on $a \succ x$ pairs yields the result. Now, with some probability, other parts will be affected, but not too much. Since the MLE is continuous in $\#$ (see Appendix E), this noise will not invalidate monotonicity.

5 Pairwise Majority Consistency

Social choice theory has its root in the analysis of politics, where in many cases it is important to use aggregation rules that respect the wishes of a majority. A famous issue is that the "majority will" may not be coherent and in particular fail to be transitive. A minimal majoritarian requirement, thus, would be what we call *pairwise majority consistency (PMC)*: in cases where the majority produces a definite ranking, the aggregate should respect it.

[5] Assume that F is twice differentiable. Since $\log F$ is strictly concave, its second derivative is strictly negative. A straightforward calculation shows that then $\partial^2\mathcal{L}/\partial\beta_c\partial\beta_c < 0$ and that $\partial^2\mathcal{L}/\partial\beta_c\partial\beta_b > 0$. By definition of ψ, for each x, $(\partial\mathcal{L}/\partial\beta_c)(x, \psi(x)) = 0$. Since $\partial^2\mathcal{L}/\partial\beta_c\partial\beta_b > 0$, the function $\partial\mathcal{L}/\partial\beta_c$ is increasing in its first argument, and so $(\partial\mathcal{L}/\partial\beta_c)(x + \Delta, \psi(x)) > 0$ for all $\Delta > 0$. Since $\partial^2\mathcal{L}/\partial\beta_c\partial\beta_c < 0$, the function $\partial\mathcal{L}/\partial\beta_c$ is decreasing in its second argument, and hence $\psi(x + \Delta) > \psi(x)$, as desired.

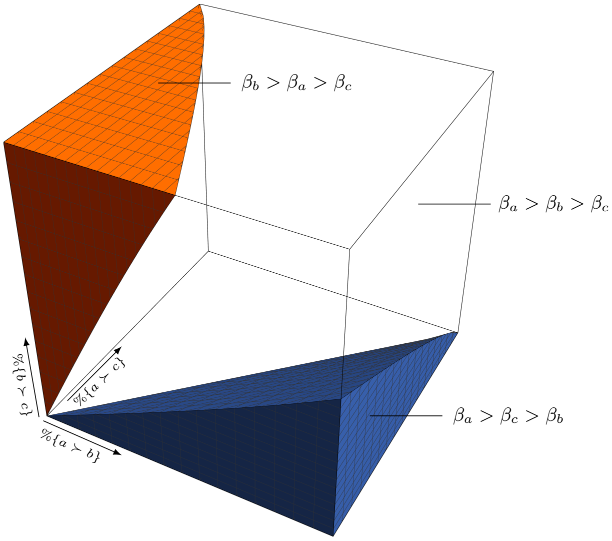


Figure 2: The cube shows all datasets in the space T, in which pairwise majority consistency requires that $\hat{\beta}_a > \hat{\beta}_b > \hat{\beta}_c$. The MLE for Thurstone-Mosteller models fails the condition in the shaded areas.

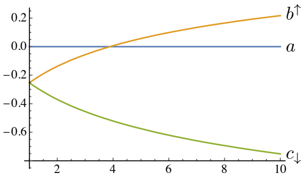
Definition 5.1. *Suppose it is possible to label alternatives as $\mathcal{X} = \{x_1, \dots, x_m\}$ such that whenever $i < j$, it holds that $\#\{x_i \succ x_j\} > \#\{x_j \succ x_i\}$. Then, pairwise majority consistency (PMC) requires that for every MLE $\hat{\beta}$, it holds that $\hat{\beta}_{x_i} \geq \hat{\beta}_{x_j}$ for all $i < j$.*

In contrast to our previous properties, PMC is violated by random utility models.

Example 5.2. Consider $\mathcal{X} = \{a, b, c\}$, and consider the dataset

$$\#\{a \succ b\} = 3, \#\{b \succ a\} = 2, \#\{a \succ c\} = 3, \#\{c \succ a\} = 2, \#\{b \succ c\} = 10, \#\{c \succ b\} = 1.$$

This dataset conforms to Definition 5.1 if we label $x_1, x_2, x_3 = a, b, c$. However, the unique MLE in the Thurstone–Mosteller model is $\hat{\beta}_a = 0$, $\hat{\beta}_b \approx 0.217$ and $\hat{\beta}_c \approx -0.751$, so that $\beta_b > \hat{\beta}_a > \hat{\beta}_c$. The same example works for Bradley–Terry, which has MLE $\hat{\beta}_a = 0$, $\hat{\beta}_b \approx 0.316$ and $\hat{\beta}_c \approx -1.256$.



Why does the MLE not respect the majority ordering on this example? If the number $\#\{b \succ c\}$ was slightly above 1, we would obtain an MLE respecting the majority ordering, with $a \succ b \succ c$. However, as $\#\{b \succ c\}$ increases, due to the monotonicity of MLEs (Theorem 4.2), we find that $\hat{\beta}_b$ increases and $\hat{\beta}_c$ decreases. When $\#\{b \succ c\}$ becomes sufficiently large, $\hat{\beta}_b$ crosses $\hat{\beta}_a$. Thus, we find that the MLE has the ordering $b \succ a \succ c$, which violates PMC. The figure on the right shows this behavior in the style of Figure 1, as $\#\{b \succ c\}$ increases from 1 to 10; we can see that PMC is violated from about 4.

This reasoning applies more generally to other random utility models beyond Thurstone–Mosteller, and we can construct similar counterexamples for a large class of such models; see Appendix E.

Theorem 5.3. *Maximum likelihood estimation violates pairwise majority consistency whenever F is strictly monotonic, strictly log-concave, and differentiable.*

How frequent are PMC violations? Write $\%\{x \succ y\} = \#\{x \succ y\}/(\#\{x \succ y\} + \#\{y \succ x\})$ for the fraction of x vs y comparisons that x wins. For $\mathcal{X} = \{a, b, c\}$, let T be the space of datasets with

$$0.5 < \%\{a \succ b\}, \%\{a \succ c\}, \%\{b \succ c\} \leq 1.$$

For all datasets in T, PMC requires that $\hat{\beta}_a > \hat{\beta}_b > \hat{\beta}_c$. In Figure 2, we draw the cube T and show the regions where the MLE for Thurstone–Mosteller fails PMC. Example 5.2, suitably normalized, falls in the upper orange region. Sampling uniformly over T, we find that Thurstone–Mosteller fails PMC in 17.8% of datasets, while Bradley–Terry fails in 16.6% of datasets.

6 Separability

We close by considering the *separability axiom* [16, 19]. It requires that when we merge two datasets, then wherever the MLE agreed on the datasets, this agreement is preserved in the combined dataset.

Definition 6.1. *Consider two datasets $\#^1$ and $\#^2$, and let $\hat{\beta}^1$ and $\hat{\beta}^2$ be MLEs. Suppose there exist two alternatives $a, b \in \mathcal{X}$ such that $\hat{\beta}^1_a > \hat{\beta}^1_b$ and $\hat{\beta}^2_a > \hat{\beta}^2_b$. Separability requires that for every MLE $\hat{\beta}$ for the pooled dataset $\# = \#^1 + \#^2$, it also holds that $\hat{\beta}_a > \hat{\beta}_b$.*

Separability is also called *consistency*, and seems particularly desirable in cases where we combine pairwise comparisons from different sources. While perhaps on first glance innocuous, separability is an extremely strong requirement, and few rules satisfy it; one can prove in general that separability constrains rules to be linear [14]. Since likelihood maximization is not linear, it is no surprise that MLEs for random utility models fail separability.

Example 6.2. Let $\mathcal{X} = \{a, b, c\}$, and consider the two datasets

$$\#^1\{a \succ c\} = 6,\ \#^1\{c \succ a\} = 4,\ \#^1\{c \succ b\} = 100, \#^1\{b \succ c\} = 1, \text{ and}$$
$$\#^2\{a \succ c\} = 6,\ \#^2\{c \succ a\} = 4,\ \#^2\{b \succ a\} = 100, \#^2\{a \succ b\} = 1,$$

with 0 counts on all unspecified pairs. The unique MLEs for Thurstone–Mosteller on $\#^1$ and $\#^2$ are

$$\hat{\beta}^1_a = 0, \hat{\beta}^1_b \approx -2.58, \hat{\beta}^1_c \approx -0.253; \quad \text{and} \quad \hat{\beta}^2_a = 0, \hat{\beta}^2_b \approx 2.330, \hat{\beta}^2_c \approx -0.253.$$

We have both $\hat{\beta}^1_a > \hat{\beta}^1_c$ and $\hat{\beta}^2_a > \hat{\beta}^2_c$. However, the unique MLE on $\# = \#^1 + \#^2$ is $\hat{\beta}_a = 0$, $\hat{\beta}_b \approx 0.987$ and $\hat{\beta}_c \approx 1.973$. Thus. $\hat{\beta}_a < \hat{\beta}_c$, and so Thurstone–Mosteller violates separability. (The same example shows that Bradley–Terry violates separability.)

Intuitively, in both $\#_1$ and $\#_2$ there is a weak tendency to rank a above c, and the MLE can implement this tendency without incurring any cost on other pairs (since Lemma 2.2 applies). However, once we combine the datasets, a strong consensus for $c \succ b \succ a$ emerges, and overriding this consensus to ensure $a \succ c$ is not worth it. While failing separability, the MLE's behavior seems perfectly sensible, and we prove in Appendix F that all random utility model do the same on this kind of example.

Theorem 6.3. *Maximum likelihood estimation violates separability whenever F is strictly monotonic, strictly log-concave, and differentiable.*

Like for PMC, we can again ask on what percentage of (pairs of) datasets the MLE fails separately. Since we sample over pairs, we might guess the answer to be of lower order than in the case of PMC, and this is borne out by the data. For $m = 3$ alternatives, sampling uniformly over the space of datasets for which each pair of distinct alternatives is compared equally often, we find that on about 1.5% of dataset pairs, Thurstone–Mosteller fails separability. This fraction increases as m increases, since there are more pairs of alternatives for which separability can be violated.

7 Discussion

To recap, we have established (under very mild assumptions) that the aggregation of pairwise comparisons via the MLE of a random utility model satisfies Pareto efficiency and monotonicity, and does not satisfy pairwise majority consistency and separability.

Our positive results deal with central properties that are required for an aggregation procedure: it does not override unanimous opinions (Pareto efficiency) and it incorporates new information (monotonicity). The latter property can be seen as a participation incentive, guaranteeing agents that each additional pairwise comparison will move the aggregate. Separability and pairwise majority consistency are not satisfied by random utility models, but arguably these properties are not as universally desirable. An analogy to the world of ranking-based voting rules is instructive, where separability characterizes a specific class of aggregators (positional scoring rules) [19], but none of them satisfies pairwise majority consistency [13].

Overall, we view our results as lending normative support to — and a more nuanced understanding of — existing and future applications of random utilities models for societal decision making.

References

[1] H. Azari Soufiani, D. C. Parkes, and L. Xia. A statistical decision-theoretic framework for social choice. In *Proceedings of the 28th Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 3185–3193, 2014.

[2] C. Berge. *Topological Spaces: including a treatment of multi-valued functions, vector spaces, and convexity*. Oliver & Boyd, 1963.

[3] R. A. Bradley. Paired comparisons: Some basic procedures and examples. In *Handbook of Statistics*, volume 4, pages 299–326. Elsevier, 1984.

[4] I. Caragiannis, A. D. Procaccia, and N. Shah. When do noisy votes reveal the truth? *ACM Transactions on Economics and Computation*, 4(3): article 15, 2016.

[5] M.-J.-A.-N. Condorcet. Essai sur l'application de l'analyse à la probabilité de décisions rendues à la pluralité de voix. Imprimerie Royal, 1785. Facsimile published in 1972 by Chelsea Publishing Company, New York.

[6] V. Conitzer and T. Sandholm. Common voting rules as maximum likelihood estimators. In *Proceedings of the 21st Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 145–152, 2005.

[7] R. Freedman, J. Schaich Borg, W. Sinnott-Armstrong, J. P. Dickerson, and V. Conitzer. Adapting a kidney exchange algorithm to align with human values. *Artificial Intelligence*, 283, 2020.

[8] Julio González-Díaz, Ruud Hendrickx, and Edwin Lohmann. Paired comparisons analysis: an axiomatic approach to ranking methods. *Social Choice and Welfare*, 42(1):139–169, 2014.

[9] G. A. Jehle and P. J. Reny. *Advanced Microeconomic Theory*. Pearson, Prentice Hall, 2011.

[10] A. Kahng, M. K. Lee, R. Noothigattu, A. D. Procaccia, and C.-A. Psomas. Statistical foundations of virtual democracy. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 3173–3182, 2019.

[11] M. K. Lee, D. Kusbit, A. Kahng, J. T. Kim, X. Yuan, A. Chan, R. Noothigattu, D. See, S. Lee, C.-A. Psomas, and A. D. Procaccia. WeBuildAI: Participatory framework for fair and efficient algorithmic governance. In *Proceedings of the 22nd ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW)*, article 181, 2019.

[12] F. Mosteller. Remarks on the method of paired comparisons: I. the least squares solution assuming equal standard deviations and equal correlations. *Psychometrika*, 16(1):3–9, 1951.

[13] H. Moulin. *The Strategy of Social Choice*, volume 18 of *Advanced Textbooks in Economics*. North-Holland, 1983.

[14] R. B. Myerson. Axiomatic derivation of scoring rules without the ordering assumption. *Social Choice and Welfare*, 12(1):59–74, 1995.

[15] R. Noothigattu, S. S. Gaikwad, E. Awad, S. Dsouza, I. Rahwan, P. Ravikumar, and A. D. Procaccia. A voting-based system for ethical decision making. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI)*, pages 1587–1594, 2018.

[16] J. H. Smith. Aggregation of preferences with variable electorate. *Econometrica*, 41(6):1027–1041, 1973.

[17] L. L. Thurstone. A law of comparative judgment. *Psychological Review*, 34:273–286, 1927.

[18] L. Xia. Bayesian estimators as voting rules. In *Proceedings of the 32nd Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, 2016.

[19] H. P. Young. Social choice scoring functions. *SIAM Journal of Applied Mathematics*, 28(4):824–838, 1975.

A Appendix for Section 2.1

A.1 Proof of Lemma 2.1

In this proof, we also show that under the given conditions, there exists an MLE $\hat{\beta}$ satisfying

$$\|\hat{\beta}\|_\infty \leq -(|\mathcal{X}| - 1)\, F^{-1}\left(2^{-K/\eta}\right),$$

where $K = \sum_{(x,y)\in\mathcal{X}^2} \#\{x \succ y\}$ and $\eta = \min_{(x,y):\#\{x\succ y\}>0} \#\{x \succ y\}$.[6]

Suppose the comparison graph $\mathcal{G}_\#$ is such that each of its connected components is strongly connected. First, we show that moving any connected component (keeping all other distances fixed) does not change the likelihood. In particular, let C be an arbitrary connected component that does not have the reference alternative r. The likelihood function can then be rewritten as

$$\mathcal{L}(\beta) = \sum_{x,y\in C} \#\{x \succ y\} \log F(\beta_x - \beta_y) + \sum_{x,y\notin C} \#\{x \succ y\} \log F(\beta_x - \beta_y),$$

as there are no edges between C and its complement. For any vector $\beta \in \mathcal{D}$, define $\beta^\Delta \in \mathcal{D}$ for any $\Delta \in \mathbb{R}$ as follows

$$\beta_x^\Delta = \begin{cases} \beta_x + \Delta & ;\text{ if } x \in C \\ \beta_x & ;\text{ otherwise.} \end{cases}$$

That is, β^Δ is the same as β, except with utilities changed by the constant Δ for C. The likelihood at this point for any Δ is

$$\mathcal{L}(\beta^\Delta) = \sum_{x,y\in C} \#\{x \succ y\} \log F\left(\beta_x + \Delta - \beta_y - \Delta\right) + \sum_{x,y\notin C} \#\{x \succ y\} \log F(\beta_x - \beta_y) = \mathcal{L}(\beta).$$

Hence, adding any Δ to a connected component does not affect the likelihood. In particular, for any maximizer, we could set Δ such that an alternative (of choice) in the connected component C has zero beta value, giving us a new maximizer. And this holds for every connected component. Hence, we just need to consider β vectors which have a reference alternative in each of the connected components in order to find a maximizer. Let $r_1, r_2, \ldots, r_k$ denote the references we set in each of the connected components $C_1, C_2, \ldots, C_k$ respectively (where k denotes the total number of connected components).

Define

$$B := -(|\mathcal{X}| - 1)F^{-1}\left(2^{-K/\eta}\right),$$

where K and η are as defined at the beginning of the proof. Consider an arbitrary beta vector (obeying the reference alternative constraints) with $\|\beta\|_\infty > B$. Then, there exists alternative $a \notin \{r_1, r_2, \ldots, r_k\}$ such that $|\beta_a| > B$. Without loss of generality, let $\beta_a > B$. Let C_t be the connected component that a lies in, with the reference alternative r_t. Consider all alternatives in C_t whose β value lies between that of r_t and a. The total number of these alternatives (including the end points r_t and a) is at most $|\mathcal{X}|$. Hence, the number of pairwise segments encountered starting from r_t and ending at a is at most $(|\mathcal{X}| - 1)$.[7] And since all these pairwise distances make up the total distance $\beta_a - \beta_{r_t} > B$, it implies that there exists at least one pairwise distance that is strictly larger than $B/(|\mathcal{X}| - 1)$. Let (b, c) denote the ends of this pairwise segment. That is, $b, c \in C_t$ such that $\beta_c - \beta_b > \frac{B}{|\mathcal{X}|-1}$ and there is no alternative in C_t with a β value lying in the segment (β_b, β_c). Let $\mathcal{U}$ denote the set of alternatives of C_t that lie to the left of b, i.e., $\mathcal{U} = \{x \in C_t | \beta_x \leq \beta_b\}$, and $\mathcal{V}$ be the set of alternatives of C_t that lie to the right of c, i.e., $\mathcal{V} = \{x \in C_t | \beta_x \geq \beta_c\}$. Since no alternative in C_t lies in between b and c, $(\mathcal{U}, \mathcal{V})$ is a partition of C_t. Next, as every connected component is strongly connected, C_t is also strongly connected. Hence, there has to be at least one edge going from $\mathcal{U}$ to $\mathcal{V}$ (otherwise, there would be no paths from alternatives in $\mathcal{U}$ to alternatives in $\mathcal{V}$ breaking strongly

[6] That is, K denotes the total number of comparisons in the dataset, and η denotes the smallest positive comparison number in it (or equivalently, the smallest positive weight in $\mathcal{G}_\#$). Also note that, F^{-1} exists in $(0, 1)$ as F is strictly monotonic and continous.

[7] assuming all alternatives of C_t are placed on the real line according to their β values.

connectedness). Let this edge be given by $(u, v) \in \mathcal{U} \times \mathcal{V}$. This implies that $\beta_u \leq \beta_b$, $\beta_v \geq \beta_c$ and $\#\{u \succ v\} > 0$. Hence, we have

$$\beta_v - \beta_u \geq \beta_c - \beta_b > \frac{B}{|\mathcal{X}| - 1}.$$

The log-likelihood can be rewritten as

$$\begin{aligned}
\mathcal{L}(\beta) &= \#\{u \succ v\} \log F(\beta_u - \beta_v) + \sum_{(x,y) \neq (u,v)} \#\{x \succ y\} \log F(\beta_x - \beta_y) \\
&\leq \#\{u \succ v\} \log F(\beta_u - \beta_v) \\
&< \#\{u \succ v\} \log F\left(-\frac{B}{|\mathcal{X}| - 1}\right), \qquad (2)
\end{aligned}$$

where the first inequality holds because $\#\{x \succ y\} \geq 0$ and $\log F(\beta_x - \beta_y) \leq 0$ (as $F(\cdot) \leq 1$), and the second inequality holds because $\#\{u \succ v\} > 0$, $\beta_u - \beta_v < -\frac{B}{|\mathcal{X}|-1}$ and $\log F$ is strictly increasing. Next, consider the log-likelihood of the zero vector. We have,

$$\mathcal{L}(0) = \sum_{x \neq y} \#\{x \succ y\} \log F(0) = K \log F(0),$$

as K is the total number of comparisons in the dataset. Recall the definition of B, we have,

$$B = -(|\mathcal{X}| - 1) F^{-1}\left(2^{-K/\eta}\right) \implies \log F\left(-\frac{B}{|\mathcal{X}| - 1}\right) = \frac{K}{\eta} \log\left(\frac{1}{2}\right).$$

Combining this with Equation (2), we have

$$\begin{aligned}
\mathcal{L}(\beta) &< \#\{u \succ v\} \log F\left(-\frac{B}{|\mathcal{X}| - 1}\right) \\
&= \#\{u \succ v\} \frac{K}{\eta} \log\left(\frac{1}{2}\right) \\
&\leq K \log\left(\frac{1}{2}\right) \\
&= K \log F(0) = \mathcal{L}(0),
\end{aligned}$$

where the inequality holds because $\eta = \min_{(x,y):\#\{x \succ y\}>0} \#\{x \succ y\} \leq \#\{u \succ v\}$ and $\log(1/2) < 0$, and the next equality holds as $F(0) = 1/2$. Hence, this shows that $\mathcal{L}(\beta) < \mathcal{L}(0)$ for any β with $\|\beta\|_\infty > B$. In other words, such a β vector cannot be a maximizer of $\mathcal{L}$. Therefore, to maximize $\mathcal{L}(\beta)$ we just need to consider β vectors in $[-B, B]^{|\mathcal{X}|-k}$. And since this is a closed space, a maximizer always exists. Further, this maximizer satisfies $\|\hat{\beta}\|_\infty \leq B$.

Next, to prove the converse of the theorem statement, suppose there exists a connected component of $\mathcal{G}_\#$ that is not strongly connected. Denote this connected component by C. Consider all the strongly connected components of C; they form a DAG (as the condensation of a graph is always acyclic). Hence, there exists a strongly connected component in this DAG that has no incoming edge (from the rest of C). Let this strongly connected component be denoted by S. Further, as C itself is a connected component, this implies that there exists at least one edge going from S to $C \setminus S$. Putting all this together, we have strongly connected component S such that there is no (incoming) edge from $\mathcal{X} \setminus S$ to S, and there is at least one (outgoing) edge from S to $\mathcal{X} \setminus S$. Now, suppose for the sake of contradiction that $\mathcal{L}(\beta)$ has a maximizer. And, let $\hat{\beta}$ denote an MLE. The log-likelihood can be written as

$$\begin{aligned}
\mathcal{L}(\beta) = &\sum_{x,y \in S} \#\{x \succ y\} \log F(\beta_x - \beta_y) + \sum_{x \in S, y \notin S} \#\{x \succ y\} \log F(\beta_x - \beta_y) \\
&+ \sum_{x,y \notin S} \#\{x \succ y\} \log F(\beta_x - \beta_y).
\end{aligned}$$

Consider another beta vector $\tilde{\beta} \in \mathcal{D}$ that is the same as $\hat{\beta}$ except that it has beta values increased by a constant for alternatives in S.[8] For instance,

$$\tilde{\beta}_x = \begin{cases} \hat{\beta}_x + 1 & ; \text{if } x \in S \\ \hat{\beta}_x & ; \text{otherwise.} \end{cases}$$

The likelihood at this point is

$$\begin{aligned} \mathcal{L}(\tilde{\beta}) = & \sum_{x,y \in S} \#\{x \succ y\} \log F(\hat{\beta}_x + 1 - \hat{\beta}_y - 1) + \sum_{x \in S, y \notin S} \#\{x \succ y\} \log F(\hat{\beta}_x - \hat{\beta}_y + 1) \\ & + \sum_{x,y \notin S} \#\{x \succ y\} \log F(\hat{\beta}_x - \hat{\beta}_y) \\ > & \sum_{x,y \in S} \#\{x \succ y\} \log F(\hat{\beta}_x - \hat{\beta}_y) + \sum_{x \in S, y \notin S} \#\{x \succ y\} \log F(\hat{\beta}_x - \hat{\beta}_y) \\ & + \sum_{x,y \notin S} \#\{x \succ y\} \log F(\hat{\beta}_x - \hat{\beta}_y) \\ = & \mathcal{L}(\hat{\beta}), \end{aligned}$$

where the inequality holds because $\#\{x \succ y\} \geq 0$, $\log F(\hat{\beta}_x - \hat{\beta}_y + 1) > \log F(\hat{\beta}_x - \hat{\beta}_y)$ for all $(x, y) \in S \times S^C$ as $\log F$ is strictly increasing, and there exists at least one $(x, y) \in S \times S^C$ with $\#\{x \succ y\} > 0$ (because of the presence of the outgoing edge from S to S^C). This leads to a contradiction as $\tilde{\beta}$ has strictly higher likelihood than the MLE $\hat{\beta}$. Therefore, an MLE does not exist. □

A.2 Proof of Lemma 2.2

Let a be an alternative such that there is exactly one other alternative b for which $\#\{a \succ b\} + \#\{b \succ a\} > 0$. The log-likelihood function is

$$\begin{aligned} \mathcal{L}(\beta) &= \sum_{(x,y)} \#\{x \succ y\} \log F(\beta_x - \beta_y) \\ &= \left[\sum_{\substack{(x,y) \\ x \neq a, y \neq a}} \#\{x \succ y\} \log F(\beta_x - \beta_y) \right] + \#\{a \succ b\} \log F(\beta_a - \beta_b) + \#\{b \succ a\} \log F(\beta_y - \beta_x) \\ &= \mathcal{G}(\beta_{-a}) + \#\{a \succ b\} \log F(\beta_a - \beta_b) + \#\{b \succ a\} \log F(\beta_b - \beta_a), \end{aligned}$$

where $\mathcal{G}$ is the part of the likelihood function not containing β_a. Maximizing $\mathcal{L}(\beta)$ is equivalent to first maximizing with respect to β_a and then with respect to the rest, β_{-a}.[9] Hence, we maximize

$$\#\{a \succ b\} \log F(\beta_a - \beta_b) + \#\{b \succ a\} \log F(\beta_b - \beta_a) \tag{3}$$

with respect to β_a.

Claim A.1 (Coin flip likelihood). *For $h, t > 0$ and $p \in (0, 1)$, the function $f(p) = h \cdot log(p) + t \cdot log(1 - p)$ is strictly concave with the maximum uniquely attained at*

$$\hat{p} = \frac{h}{h + t}$$

Proof. $f'(p) = \frac{h}{p} - \frac{t}{1-p}$, and $f''(p) = -\frac{h}{p^2} - \frac{t}{(1-p)^2}$. Hence, $f''(p) < 0$ for all $p \in (0, 1)$ making f a strictly concave function. Further, $f'(\hat{p}) = 0$. Hence, $\hat{p}$ as defined in the claim is the point where the maximum is attained. □

[8] In the case when S has the reference alternative r, the exact effect can be achieved by instead decreasing the beta values of all alternatives in $\mathcal{X} \setminus S$ by the same constant.

[9] In the case when a was set as the reference, we could always perform this optimization by placing the reference on some other alternative, and then shifting the complete learned vector back such that a is the reference again. Observe that this does not affect the learned distance of $(\hat{\beta}_a - \hat{\beta}_b)$, for which we are proving the desired property.

Equation (3) can be rewritten as

$$\#\{a \succ b\} \log F(\beta_a - \beta_b) + \#\{b \succ a\} \log F(\beta_b - \beta_a) = f(F(\beta_a - \beta_b)),$$

where f is the function from Claim A.1 with $h = \#\{a \succ b\} > 0$ and $t = \#\{b \succ a\} > 0$, as $F(\beta_b - \beta_a) = 1 - F(\beta_a - \beta_b)$. Applying Claim A.1, we have

$$f(F(\beta_a - \beta_b)) \leq f(\hat{p}),$$

for all β_a, β_b, where $\hat{p} = \frac{\#\{a \succ b\}}{\#\{a \succ b\} + \#\{b \succ a\}}$. Further, this upper bound can be achieved by setting $F(\beta_a - \beta_b) = \hat{p}$, which is possible as F is invertible in $(0, 1)$ by strict monotonicty and continuity. Therefore, Equation (3) is uniquely maximized at $\beta_a = \beta_b + F^{-1}(\hat{p})$. And hence, every MLE satisfies

$$\hat{\beta}_a = \hat{\beta}_b + F^{-1}\left(\frac{\#\{a \succ b\}}{\#\{a \succ b\} + \#\{b \succ a\}}\right) = \hat{\beta}_b + \delta(a, b).$$

□

A.3 Proof of Lemma 2.3

The initial part of this proof is similar to the proof of Lemma 2.1. Let B denote the bound $|\mathcal{X}| \cdot \max_{(x,y)} \delta(x, y)$. And, recall that r denotes the alternative set as the reference, i.e. $\beta_r = 0$. Suppose for the sake of contradiction that there exists an MLE $\hat{\beta}$ with $\|\hat{\beta}\|_\infty > B$. This implies that there exists an alternative a such that $|\hat{\beta}_a| > B$. WLOG, suppose $\hat{\beta}_a > B$. The number of alternatives whose β value lies between that of a and the reference r (including both these points) is at most $|\mathcal{X}|$. Hence, the number of pairwise segments encountered starting from r and ending at a is at most $(|\mathcal{X}| - 1)$.[10] And since all these pairwise distances make up the total distance $\hat{\beta}_a - \hat{\beta}_r > B$, it implies that there exists at least one pairwise distance that is strictly larger than $B/(|\mathcal{X}| - 1)$. Let (b, c) denote the ends of this pairwise segment. That is, $\hat{\beta}_c - \hat{\beta}_b > \frac{B}{|\mathcal{X}|-1}$, and there is no alternative with a β value lying in the segment $(\hat{\beta}_b, \hat{\beta}_c)$. Construct a new beta vector $\tilde{\beta} \in \mathcal{D}$, such that $\tilde{\beta}$ is the same as $\hat{\beta}$ for alternatives to the left of alternative b, while is decreased by a small positive constant ϵ for all the other alternatives. That is,

$$\tilde{\beta}_x = \begin{cases} \hat{\beta}_x & ; \text{ if } \hat{\beta}_x \leq \hat{\beta}_b \\ \hat{\beta}_x - \epsilon & ; \text{ if } \hat{\beta}_x \geq \hat{\beta}_c. \end{cases}$$

In particular, choose ϵ such that the distance between b and c is still bigger than $\max_{(x,y)} \delta(x, y)$. This is possible because the original distance between b and c (i.e. $\hat{\beta}_c - \hat{\beta}_b$) is strictly larger than $\frac{B}{|\mathcal{X}|-1} = \frac{|\mathcal{X}|}{|\mathcal{X}|-1} \max_{(x,y)} \delta(x, y)$. Hence, one can choose $\epsilon > 0$ such that the new distance between b and c (i.e. $\tilde{\beta}_c - \tilde{\beta}_b$) is say the mid point of $\frac{|\mathcal{X}|}{|\mathcal{X}|-1} \max_{(x,y)} \delta(x, y)$ and $\max_{(x,y)} \delta(x, y)$. This would imply that we have

$$\tilde{\beta}_c - \tilde{\beta}_b > \max_{(x,y)} \delta(x, y). \tag{4}$$

Next, we show that in fact, $\mathcal{L}(\tilde{\beta}) > \mathcal{L}(\hat{\beta})$. The log-likelihood function is given as

$$\begin{aligned} \mathcal{L}(\beta) &= \sum_{(x,y) \in \mathcal{X}^2} \#\{x \succ y\} \log F(\beta_x - \beta_y) \\ &= \sum_{\{x,y\} \subseteq \mathcal{X}} \Big[\#\{x \succ y\} \log F(\beta_x - \beta_y) + \#\{y \succ x\} \log F(\beta_y - \beta_x)\Big] \\ &= \sum_{\{x,y\} \subseteq \mathcal{X}} f_{xy}(F(\beta_x - \beta_y)), \end{aligned}$$

where f_{xy} is the function from Claim A.1 with $h = \#\{x \succ y\} > 0$ and $t = \#\{y \succ x\} > 0$. Hence, from the claim, this function f_{xy} is strictly concave with a maximum attained at $\hat{p}_{xy} =$

[10] assuming all the alternatives are placed on the real line according to their β values.

$\frac{\#\{x \succ y\}}{\#\{x \succ y\}+\#\{y \succ x\}}$. Let's call $\mathcal{U}$ as the set of alternatives x with $\hat{\beta}_x \leq \hat{\beta}_b$ (i.e. the alternatives with β value unchanged), and $\mathcal{V}$ as the set of alternatives x with $\hat{\beta}_x \geq \hat{\beta}_c$ (i.e. the alternatives whose β value is decreased by ϵ). Observe that neither of these sets in empty, and they partition $\mathcal{X}$. Therefore, the log-likelihood at $\tilde{\beta}$ is

$$\begin{aligned}
\mathcal{L}(\tilde{\beta}) &= \sum_{\{x,y\} \subseteq \mathcal{X}} f_{xy}\left(F(\tilde{\beta}_x - \tilde{\beta}_y)\right) \\
&= \sum_{\{x,y\} \subseteq \mathcal{U}} f_{xy}\left(F(\tilde{\beta}_x - \tilde{\beta}_y)\right) + \sum_{\{x,y\} \subseteq \mathcal{V}} f_{xy}\left(F(\tilde{\beta}_x - \tilde{\beta}_y)\right) + \sum_{(v,u) \in \mathcal{V} \times \mathcal{U}} f_{vu}\left(F(\tilde{\beta}_v - \tilde{\beta}_u)\right).
\end{aligned}$$

Note that, for $x, y \in \mathcal{U}$, the distance $(\tilde{\beta}_x - \tilde{\beta}_y)$ is the same as $(\hat{\beta}_x - \hat{\beta}_y)$ as the β values are unchanged. In the case of $x, y \in \mathcal{V}$, again the distance $(\tilde{\beta}_x - \tilde{\beta}_y)$ is the same as $(\hat{\beta}_x - \hat{\beta}_y)$ as both β values (of x and y) are decreased by the same ϵ. Finally, for any pair $(v, u) \in \mathcal{V} \times \mathcal{U}$, we have $\tilde{\beta}_v - \tilde{\beta}_u = \hat{\beta}_v - \hat{\beta}_u - \epsilon$, i.e. this pairwise distance decreases by ϵ. Hence, the likelihood at $\tilde{\beta}$ becomes

$$\mathcal{L}(\tilde{\beta}) = \sum_{\{x,y\} \subseteq \mathcal{U}} f_{xy}\left(F(\hat{\beta}_x - \hat{\beta}_y)\right) + \sum_{\{x,y\} \subseteq \mathcal{V}} f_{xy}\left(F(\hat{\beta}_x - \hat{\beta}_y)\right) + \sum_{(v,u) \in \mathcal{V} \times \mathcal{U}} f_{vu}\left(F(\hat{\beta}_v - \hat{\beta}_u - \epsilon)\right).$$

Let's look at the terms $f_{vu}\left(F(\hat{\beta}_v - \hat{\beta}_u - \epsilon)\right)$ for $(v, u) \in \mathcal{V} \times \mathcal{U}$. We have

$$\hat{\beta}_v - \hat{\beta}_u > \hat{\beta}_v - \hat{\beta}_u - \epsilon = \tilde{\beta}_v - \tilde{\beta}_u \geq \tilde{\beta}_c - \tilde{\beta}_b > \max_{(x,y)} \delta(x, y) \geq \delta(v, u),$$

where the second inequality holds because $v \in \mathcal{V}$ is to the right of c while $u \in \mathcal{U}$ is to the left of b, and the third inequality holds from Equation (4). Rewriting this equation keeping only the main components, we have

$$\hat{\beta}_v - \hat{\beta}_u > \tilde{\beta}_v - \tilde{\beta}_u > \delta(v, u).$$

As F is a strictly increasing function, applying it to this equation gives us

$$F(\hat{\beta}_v - \hat{\beta}_u) > F(\tilde{\beta}_v - \tilde{\beta}_u) > F(\delta(v, u)) = \hat{p}_{vu},$$

where the equality holds by definition of the perfect-fit distance and $\hat{p}_{vu}$. Hence, by changing from $F(\hat{\beta}_v - \hat{\beta}_u)$ to $F(\tilde{\beta}_v - \tilde{\beta}_u)$, we move closer to the maxima of f_{vu} (or alternatively, $F(\tilde{\beta}_v - \tilde{\beta}_u)$ is a convex combination of $F(\hat{\beta}_v - \hat{\beta}_u)$ and the maxima $\hat{p}_{vu}$). But, as f_{vu} is strictly concave, it means that this change leads to an increase in its value. That is,

$$f_{vu}\left(F(\hat{\beta}_v - \hat{\beta}_u)\right) < f_{vu}\left(F(\tilde{\beta}_v - \tilde{\beta}_u)\right),$$

and this holds for every $(v, u) \in \mathcal{V} \times \mathcal{U}$. Hence, the log-likelihood at $\tilde{\beta}$ becomes

$$\begin{aligned}
\mathcal{L}(\tilde{\beta}) &= \sum_{\{x,y\} \subseteq \mathcal{U}} f_{xy}\left(F(\hat{\beta}_x - \hat{\beta}_y)\right) + \sum_{\{x,y\} \subseteq \mathcal{V}} f_{xy}\left(F(\hat{\beta}_x - \hat{\beta}_y)\right) + \sum_{(v,u) \in \mathcal{V} \times \mathcal{U}} f_{vu}\left(F(\tilde{\beta}_v - \tilde{\beta}_u)\right) \\
&> \sum_{\{x,y\} \subseteq \mathcal{U}} f_{xy}\left(F(\hat{\beta}_x - \hat{\beta}_y)\right) + \sum_{\{x,y\} \subseteq \mathcal{V}} f_{xy}\left(F(\hat{\beta}_x - \hat{\beta}_y)\right) + \sum_{(v,u) \in \mathcal{V} \times \mathcal{U}} f_{vu}\left(F\left(\hat{\beta}_v - \hat{\beta}_u\right)\right) \\
&= \mathcal{L}(\hat{\beta}).
\end{aligned}$$

That is, $\mathcal{L}(\tilde{\beta}) > \mathcal{L}(\hat{\beta})$, leading to a contradiction. Hence, for every MLE $\hat{\beta}$, we must have $\|\hat{\beta}\|_\infty \leq |\mathcal{X}| \cdot \max_{(x,y)} \delta(x, y)$. □

B Proof of Lemma 2.4

The log-likelihood function is given as

$$\mathcal{L}(\beta) = \sum_{(x,y) \in \mathcal{X}^2} \#\{x \succ y\} \log F(\beta_x - \beta_y).$$

Consider $\beta \neq \gamma \in \mathcal{D}$ and $\theta \in (0,1)$. Then,

$$\begin{aligned}
\mathcal{L}(\theta\beta + (1-\theta)\gamma) &= \sum_{(x,y)} \#\{x \succ y\} \log F(\theta\beta_x + (1-\theta)\gamma_x - \theta\beta_y - (1-\theta)\gamma_y) \\
&= \sum_{(x,y)} \#\{x \succ y\} \log F(\theta(\beta_x - \beta_y) + (1-\theta)(\gamma_x - \gamma_y)) \\
&\geq \sum_{(x,y)} \#\{x \succ y\} \big[\theta \log F(\beta_x - \beta_y) + (1-\theta)\log F(\gamma_x - \gamma_y)\big] \\
&= \theta \sum_{(x,y)} \#\{x \succ y\} \log F(\beta_x - \beta_y) + (1-\theta) \sum_{(x,y)} \#\{x \succ y\} \log F(\gamma_x - \gamma_y) \\
&= \theta\mathcal{L}(\beta) + (1-\theta)\mathcal{L}(\gamma),
\end{aligned}$$

where the inequality holds because $\log F$ is concave, and $\#\{x \succ y\} \geq 0$ for every $(x,y) \in \mathcal{X}^2$. Hence, $\mathcal{L}$ is a concave function.

Next, suppose the comparison graph $\mathcal{G}_\#$ is connected. Recall, r denotes the reference alternative set to zero. As $\beta \neq \gamma$, this implies that there exists an alternative $a \neq r$ such that $\beta_a \neq \gamma_a$. We know that the graph $\mathcal{G}_\#$ is connected, hence, there exists an undirected path from a to r in $\mathcal{G}_\#$. Let this (undirected) path be given as

$$a = v_0 \to v_1 \to v_2 \to \cdots \to v_t \to v_{t+1} = r.$$

As $\beta_a - \beta_r \neq \gamma_a - \gamma_r$, this implies that there exists $(l, l+1)$ such that $\beta_{v_l} - \beta_{v_{l+1}} \neq \gamma_{v_l} - \gamma_{v_{l+1}}$. Because if this difference was equal for all $l \in [0,t]$, it would imply that $\beta_a - \beta_r = \gamma_a - \gamma_r$. As there's an edge between v_l and v_{l+1}, it implies that either $\#\{v_l \succ v_{l+1}\} > 0$ or $\#\{v_{l+1} \succ v_l\} > 0$. Without loss of generality, let $\#\{v_l \succ v_{l+1}\} > 0$. The log-likelihood is then

$$\begin{aligned}
\mathcal{L}(\theta\beta + (1-\theta)\gamma) &= \#\{v_l \succ v_{l+1}\} \log F(\theta(\beta_{v_l} - \beta_{v_{l+1}}) + (1-\theta)(\gamma_{v_l} - \gamma_{v_{l+1}})) \\
&\qquad + \sum_{(x,y)\neq(v_l,v_{l+1})} \#\{x \succ y\} \log F(\theta(\beta_x - \beta_y) + (1-\theta)(\gamma_x - \gamma_y)) \\
&> \#\{v_l \succ v_{l+1}\} \big[\theta \log F(\beta_{v_l} - \beta_{v_{l+1}}) + (1-\theta)\log F(\gamma_{v_l} - \gamma_{v_{l+1}})\big] \\
&\qquad + \sum_{(x,y)\neq(v_l,v_{l+1})} \#\{x \succ y\} \left[\theta \log F(\beta_x - \beta_y) + (1-\theta)\log F(\gamma_x - \gamma_y)\right] \\
&= \theta\mathcal{L}(\beta) + (1-\theta)\mathcal{L}(\gamma)
\end{aligned}$$

where the strict inequality holds because $\#\{v_l \succ v_{l+1}\} > 0$, $\theta \in (0,1)$, $\beta_{v_l} - \beta_{v_{l+1}} \neq \gamma_{v_l} - \gamma_{v_{l+1}}$ and $\log F$ is strictly concave. Therefore, $\mathcal{L}$ is strictly concave, and, it has unique maximizers.

For the converse, suppose the comparison graph $\mathcal{G}_\#$ is not connected (in the undirected form). As there is only one reference alternative r, let C be a connected component that does not contain r. The log-likelihood can then be rewritten as

$$\mathcal{L}(\beta) = \sum_{x,y \in C} \#\{x \succ y\} \log F(\beta_x - \beta_y) + \sum_{x,y \notin C} \#\{x \succ y\} \log F(\beta_x - \beta_y),$$

as there are no edges between C and its complement. Similar to proof of Lemma 2.1, for any vector $\beta \in \mathcal{D}$, define $\beta^\Delta \in \mathcal{D}$ for any $\Delta > 0$ as follows

$$\beta^\Delta_z = \begin{cases} \beta_z + \Delta & ; \text{ if } z \in C \\ \beta_z & ; \text{ if } z \notin C. \end{cases}$$

The likelihood at this point for any Δ is

$$\mathcal{L}(\beta^\Delta) = \sum_{x,y \in C} \#\{x \succ y\} \log F(\beta_x + \Delta - \beta_y - \Delta) + \sum_{x,y \notin C} \#\{x \succ y\} \log F(\beta_x - \beta_y) = \mathcal{L}(\beta). \tag{5}$$

Consider any $\theta \in (0,1)$. Then,

$$(\theta\beta^\Delta + (1-\theta)\beta)_z = \begin{cases} \theta(\beta_z + \Delta) + (1-\theta)\beta_z &= \beta_z + \theta\Delta & ; \text{ if } z \in C \\ \theta\beta_z + (1-\theta)\beta_z &= \beta_z & ; \text{ if } z \notin C, \end{cases}$$

and hence implying that $\theta\beta^{\Delta} + (1-\theta)\beta = \beta^{\theta\Delta}$. In particular, this gives us

$$\mathcal{L}(\theta\beta^{\Delta} + (1-\theta)\beta) = \mathcal{L}(\beta^{\theta\Delta}) = \mathcal{L}(\beta) = \theta\mathcal{L}(\beta^{\Delta}) + (1-\theta)\mathcal{L}(\beta),$$

where the second equality holds because Equation (5) holds for any $\Delta > 0$ (including $\theta\Delta$). But, as $\beta^{\Delta} \neq \beta$ and $\theta \in (0,1)$, this implies that $\mathcal{L}$ is not strictly concave. Note that, this also shows that if an MLE $\hat{\beta}$ existed, it would not be unique. As, $\hat{\beta}^{\Delta}$, with say $\Delta = 1$, would have the same likelihood as $\hat{\beta}$ making it an MLE as well.

Hence, concluding the proof that $\mathcal{L}(\beta)$ is strictly concave and the MLE is unique, iff the comparison graph $\mathcal{G}_{\#}$ is connected. □

C Proof of Theorem 3.2

Suppose the dataset is such that it satisfies the properties given in Definition 3.1, i.e., $\#\{a \succ b\} > \#\{b \succ a\}$, and for every other alternative $x \in \mathcal{X} \setminus \{a, b\}$, we have

$$\#\{a \succ x\} > \#\{b \succ x\} \quad \text{and} \quad \#\{x \succ a\} < \#\{x \succ b\}.$$

Suppose for the sake of contradiction that there exists an MLE $\hat{\beta}$ such that $\hat{\beta}_a < \hat{\beta}_b$. Construct $\tilde{\beta}$ such that it is the same as $\hat{\beta}$, except with a's and b's utilities swapped.[11] That is,

$$\tilde{\beta}_x = \begin{cases} \hat{\beta}_x; \text{ if } x \notin \{a,b\} \\ \hat{\beta}_b; \text{ if } x = a \\ \hat{\beta}_a; \text{ if } x = b. \end{cases}$$

The log-likelihood at the MLE $\hat{\beta}$ is given as

$$\begin{aligned} \mathcal{L}(\beta) &= \sum_{(x,y)\in\mathcal{X}^2} \#\{x \succ y\} \log F(\beta_x - \beta_y) \\ &= \sum_{x,y\notin\{a,b\}} \#\{x \succ y\} \log F(\hat{\beta}_x - \hat{\beta}_y) \\ &\quad + \sum_{y\notin\{a,b\}} \#\{a \succ y\} \log F(\hat{\beta}_a - \hat{\beta}_y) + \sum_{y\notin\{a,b\}} \#\{b \succ y\} \log F(\hat{\beta}_b - \hat{\beta}_y) \\ &\quad + \sum_{x\notin\{a,b\}} \#\{x \succ a\} \log F(\hat{\beta}_x - \hat{\beta}_a) + \sum_{x\notin\{a,b\}} \#\{x \succ b\} \log F(\hat{\beta}_x - \hat{\beta}_b) \\ &\quad + \#\{a \succ b\} \log F(\hat{\beta}_a - \hat{\beta}_b) + \#\{b \succ a\} \log F(\hat{\beta}_b - \hat{\beta}_a). \end{aligned} \tag{6}$$

Before proceeding with the proof, we prove a simple claim.

Claim C.1. *Let $c, d, e, f > 0$ such that $c > d$ and $e > f$. Then $ce + df > cf + de$.*

Proof of Claim C.1.

$$\begin{aligned} ce + df &= c(f + (e - f)) + df \\ &= cf + c(e - f) + df \\ &> cf + d(e - f) + df \\ &= cf + de, \end{aligned}$$

where the inequality holds because $c > d$ and $(e - f) > 0$. □

[11] In case either a or b is the reference alternative, shift $\tilde{\beta}$ after swapping these two alternatives' utilities such that the reference is restored. Rest of the proof remains the same as the shifted beta vector has the same likelihood as the unshifted one.

By Claim C.1, for any $x, y \in \mathcal{X}$, we have,

$$\begin{aligned}
&\#\{a \succ y\} \log F(\hat{\beta}_a - \hat{\beta}_y) + \#\{b \succ y\} \log F(\hat{\beta}_b - \hat{\beta}_y) \\
&\qquad\qquad < \#\{a \succ y\} \log F(\hat{\beta}_b - \hat{\beta}_y) + \#\{b \succ y\} \log F(\hat{\beta}_a - \hat{\beta}_y), \\
&\#\{x \succ a\} \log F(\hat{\beta}_x - \hat{\beta}_a) + \#\{x \succ b\} \log F(\hat{\beta}_x - \hat{\beta}_b) \\
&\qquad\qquad < \#\{x \succ b\} \log F(\hat{\beta}_x - \hat{\beta}_a) + \#\{x \succ a\} \log F(\hat{\beta}_x - \hat{\beta}_b), \\
&\#\{a \succ b\} \log F(\hat{\beta}_a - \hat{\beta}_b) + \#\{b \succ a\} \log F(\hat{\beta}_b - \hat{\beta}_a) \\
&\qquad\qquad < \#\{a \succ b\} \log F(\hat{\beta}_b - \hat{\beta}_a) + \#\{b \succ a\} \log F(\hat{\beta}_a - \hat{\beta}_b),
\end{aligned}$$

using the property on the counts in the dataset, the fact that $\hat{\beta}_a < \hat{\beta}_b$ and F is strictly monotonic. Hence, using these expressions in Equation (6), we obtain

$$\begin{aligned}
\mathcal{L}(\hat{\beta}) < &\sum_{x,y \notin \{a,b\}} \#\{x \succ y\} \log F(\hat{\beta}_x - \hat{\beta}_y) \\
&+ \sum_{y \notin \{a,b\}} \#\{a \succ y\} \log F(\hat{\beta}_b - \hat{\beta}_y) + \sum_{y \notin \{a,b\}} \#\{b \succ y\} \log F(\hat{\beta}_a - \hat{\beta}_y) \\
&+ \sum_{x \notin \{a,b\}} \#\{x \succ b\} \log F(\hat{\beta}_x - \hat{\beta}_a) + \sum_{x \notin \{a,b\}} \#\{x \succ a\} \log F(\hat{\beta}_x - \hat{\beta}_b) \\
&+ \#\{a \succ b\} \log F(\hat{\beta}_b - \hat{\beta}_a) + \#\{b \succ a\} \log F(\hat{\beta}_a - \hat{\beta}_b) \\
= &\sum_{x,y \notin \{a,b\}} \#\{x \succ y\} \log F(\tilde{\beta}_x - \tilde{\beta}_y) \\
&+ \sum_{y \notin \{a,b\}} \#\{a \succ y\} \log F(\tilde{\beta}_a - \tilde{\beta}_y) + \sum_{y \notin \{a,b\}} \#\{b \succ y\} \log F(\tilde{\beta}_b - \tilde{\beta}_y) \\
&+ \sum_{x \notin \{a,b\}} \#\{x \succ b\} \log F(\tilde{\beta}_x - \tilde{\beta}_b) + \sum_{x \notin \{a,b\}} \#\{x \succ a\} \log F(\tilde{\beta}_x - \tilde{\beta}_a) \\
&+ \#\{a \succ b\} \log F(\tilde{\beta}_a - \tilde{\beta}_b) + \#\{b \succ a\} \log F(\tilde{\beta}_b - \tilde{\beta}_a) \\
= &\sum_{x \neq y} \#\{x \succ y\} \log F(\tilde{\beta}_x - \tilde{\beta}_y) \\
= &\mathcal{L}(\tilde{\beta}),
\end{aligned}$$

implying that $\tilde{\beta}$ has a strictly higher log-likelihood than the MLE $\hat{\beta}$, leading to a contradiction. Therefore, every every MLE $\hat{\beta}$ must satisfy $\hat{\beta}_a \geq \hat{\beta}_b$ under this condition. □

D Proof of Theorem 4.2

Let $\#$ and $\tilde{\#}$ be two datasets as defined in Definition 4.1, with (unique) MLEs $\hat{\beta}$ and $\tilde{\beta}$. That is, $\tilde{\#}$ is the same as $\#$, except with $\alpha > 0$ comparisons of $a \succ b$ added to it. We prove that for all alternatives $x \in \mathcal{X}$, we have

$$\tilde{\beta}_a - \tilde{\beta}_x \geq \hat{\beta}_a - \tilde{\beta}_x.$$

The proof for the b part ($\tilde{\beta}_b - \tilde{\beta}_x \leq \hat{\beta}_b - \tilde{\beta}_x$) is completely symmetric.

Let the log-likelihood function with respect to $\#$ be denoted by $\mathcal{L}$, while the log-likelihood function with respect to $\tilde{\#}$ be denoted by $\tilde{\mathcal{L}}$. Any alternative could be set as the reference, but we use a as the reference alternative in this proof for ease of exposition. As $\tilde{\#}$ is the same as $\#$, except with α additional $a \succ b$ comparisons, we have

$$\tilde{\mathcal{L}}(\beta) = \mathcal{L}(\beta) + \alpha \log F(\beta_a - \beta_b).$$

Let $\mathcal{U}$ denote the set of alternatives $u \in \mathcal{X} \setminus \{a\}$ for which $\tilde{\beta}_u - \tilde{\beta}_a \leq \hat{\beta}_u - \hat{\beta}_a$.[12] And, let $\mathcal{V}$ denote the set of alternatives $v \in \mathcal{X} \setminus \{a\}$ for which $\tilde{\beta}_v - \tilde{\beta}_a > \hat{\beta}_v - \hat{\beta}_a$. Our goal is to show that $\mathcal{U} = \mathcal{X} \setminus \{a\}$, or equivalently that $\mathcal{V} = \phi$.

[12] Even though $\tilde{\beta}_a = \hat{\beta}_a = 0$ as a is the reference, we do not omit it in some parts for better clarity.

First, we show that $b \in \mathcal{U}$. Suppose for the sake of contradiction, that $\tilde{\beta}_b - \tilde{\beta}_a > \hat{\beta}_b - \hat{\beta}_a$. Then, this implies that $\alpha \log F(\tilde{\beta}_a - \tilde{\beta}_b) < \alpha \log F(\hat{\beta}_a - \hat{\beta}_b)$ as both log and F are strictly monotonic, and $\alpha > 0$. Further, as $\hat{\beta}$ maximizes $\mathcal{L}$, we have $\mathcal{L}(\tilde{\beta}) \leq \mathcal{L}(\hat{\beta})$. This implies that $\mathcal{L}(\tilde{\beta}) + \alpha \log F(\tilde{\beta}_a - \tilde{\beta}_b) < \mathcal{L}(\hat{\beta}) + \alpha \log F(\hat{\beta}_a - \hat{\beta}_b)$. Or, $\tilde{\mathcal{L}}(\tilde{\beta}) < \tilde{\mathcal{L}}(\hat{\beta})$, which is a contradiction as $\tilde{\beta}$ is the maximizer of $\tilde{\mathcal{L}}$. This proves that $\tilde{\beta}_b - \tilde{\beta}_a \leq \hat{\beta}_b - \hat{\beta}_a$, i.e. $b \in \mathcal{U}$.

Next, suppose for the sake of contradiction that $\mathcal{V} \neq \phi$. We can rewrite the log-likelihood function $\mathcal{L}(\beta)$ as

$$\begin{aligned}\mathcal{L}(\beta) &= \sum_{(x,y)\in\mathcal{X}^2} \#\{x \succ y\} \log F(\beta_x - \beta_y) \\ &= \sum_{\{x,y\}\subseteq\mathcal{X}} \Big[\#\{x \succ y\} \log F(\beta_x - \beta_y) + \#\{y \succ x\} \log F(\beta_y - \beta_x)\Big],\end{aligned}$$

where the latter summation is over unordered pairs of alternatives $\{x, y\}$ with $x \neq y$. Denote each term in this expression by $\ell_{xy}(\beta_x - \beta_y)$, i.e.

$$\ell_{xy}(\eta) = \#\{x \succ y\} \log F(\eta) + \#\{y \succ x\} \log F(-\eta).$$

As F is log-concave, $\log F$ is a concave function. And since linear transformations, positive scalar multiplication and addition preserve concavity, each of these functions ℓ_{xy} is also concave.

Let us define operator Δ_{xy} to be such that when it is applied to a β vector, it returns the difference in β values of alternatives x and y. That is, $\Delta_{xy}\beta := \beta_x - \beta_y$. Using this notation to rewrite the log-likelihood function, we have

$$\begin{aligned}\mathcal{L}(\beta) &= \sum_{\{x,y\}} \ell_{xy}(\Delta_{xy}\beta) \\ &= \sum_{u\in\mathcal{U}} \ell_{ua}(\Delta_{ua}\beta) + \sum_{v\in\mathcal{V}} \ell_{va}(\Delta_{va}\beta) + \sum_{\{u,p\}\subseteq\mathcal{U}} \ell_{up}(\Delta_{up}\beta) \\ &\qquad + \sum_{\{v,q\}\subseteq\mathcal{V}} \ell_{vq}(\Delta_{vq}\beta) + \sum_{(u,v)\in\mathcal{U}\times\mathcal{V}} \ell_{vu}(\Delta_{vu}\beta),\end{aligned}$$

where the first two terms are the paired terms with a, the third is pairs within $\mathcal{U}$, the fourth is pairs within $\mathcal{V}$, and the last is for pairs across $\mathcal{U}$ and $\mathcal{V}$. For each $v \in \mathcal{V}$, we know $\tilde{\beta}_v - \tilde{\beta}_a > \hat{\beta}_v - \hat{\beta}_a$. Hence, we can write $\Delta_{va}\tilde{\beta} = \Delta_{va}\hat{\beta} + \delta_v$,[13] where $\delta_v > 0$ for each $v \in \mathcal{V}$. Recall, $\hat{\beta}$ is the maximizer of $\mathcal{L}$. Hence, $\mathcal{L}(\hat{\beta}_{\mathcal{U}}, \tilde{\beta}_{\mathcal{V}}) < \mathcal{L}(\hat{\beta}_{\mathcal{U}}, \hat{\beta}_{\mathcal{V}})$,[14] as the MLE $\hat{\beta}$ is unique, and $\mathcal{V} \neq \phi$. This implies that

$$\begin{aligned}&\sum_{u\in\mathcal{U}} \ell_{ua}(\Delta_{ua}\hat{\beta}) + \sum_{v\in\mathcal{V}} \ell_{va}(\Delta_{va}\tilde{\beta}) + \sum_{\{u,p\}\subseteq\mathcal{U}} \ell_{up}(\Delta_{up}\hat{\beta}) + \sum_{\{v,q\}\subseteq\mathcal{V}} \ell_{vq}(\Delta_{vq}\tilde{\beta}) + \sum_{(u,v)\in\mathcal{U}\times\mathcal{V}} \ell_{vu}(\tilde{\beta}_v - \hat{\beta}_u) \\ &< \sum_{u\in\mathcal{U}} \ell_{ua}(\Delta_{ua}\hat{\beta}) + \sum_{v\in\mathcal{V}} \ell_{va}(\Delta_{va}\hat{\beta}) + \sum_{\{u,p\}\subseteq\mathcal{U}} \ell_{up}(\Delta_{up}\hat{\beta}) + \sum_{\{v,q\}\subseteq\mathcal{V}} \ell_{vq}(\Delta_{vq}\hat{\beta}) + \sum_{(u,v)\in\mathcal{U}\times\mathcal{V}} \ell_{vu}(\hat{\beta}_v - \hat{\beta}_u).\end{aligned}$$

Cancelling terms that appear on both sides (because of the same $\hat{\beta}_{\mathcal{U}}$), and plugging in $\Delta_{va}\tilde{\beta} = \Delta_{va}\hat{\beta} + \delta_v$, we have

$$\begin{aligned}&\sum_{v\in\mathcal{V}} \ell_{va}(\Delta_{va}\hat{\beta} + \delta_v) + \sum_{\{v,q\}\subseteq\mathcal{V}} \ell_{vq}(\Delta_{vq}\hat{\beta} + \delta_v - \delta_q) + \sum_{(u,v)\in\mathcal{U}\times\mathcal{V}} \ell_{vu}(\Delta_{vu}\hat{\beta} + \delta_v) \\ &\qquad < \sum_{v\in\mathcal{V}} \ell_{va}(\Delta_{va}\hat{\beta}) + \sum_{\{v,q\}\subseteq\mathcal{V}} \ell_{vq}(\Delta_{vq}\hat{\beta}) + \sum_{(u,v)\in\mathcal{U}\times\mathcal{V}} \ell_{vu}(\Delta_{vu}\hat{\beta}).\end{aligned}$$

Or in other words,

$$\begin{aligned}&\sum_{(u,v)\in\mathcal{U}\times\mathcal{V}} \ell_{vu}(\Delta_{vu}\hat{\beta} + \delta_v) - \sum_{(u,v)\in\mathcal{U}\times\mathcal{V}} \ell_{vu}(\Delta_{vu}\hat{\beta}) \\ &\qquad < -\left[\sum_{v\in\mathcal{V}} \ell_{va}(\Delta_{va}\hat{\beta} + \delta_v) + \sum_{\{v,q\}\subseteq\mathcal{V}} \ell_{vq}(\Delta_{vq}\hat{\beta} + \delta_v - \delta_q) - \sum_{v\in\mathcal{V}} \ell_{va}(\Delta_{va}\hat{\beta}) - \sum_{\{v,q\}\subseteq\mathcal{V}} \ell_{vq}(\Delta_{vq}\hat{\beta})\right].\end{aligned} \tag{7}$$

[13] Equivalently, this could be written as $\tilde{\beta}_v = \hat{\beta}_v + \delta_v$, as $\tilde{\beta}_a = \hat{\beta}_a = 0$.

[14] Recall that alternative a has been set as the reference, and hence it zero in both these terms.

Intuitively, it says that if you increase each $\hat{\beta}_v$ by their δ_v, the increase in likelihood because of the cross terms ℓ_{vu} is less than the loss because of the exclusive v terms (or vice versa, i.e. the loss in likelihood because of ℓ_{vu} is higher than the increase because of the exclusive v terms).

For each $u \in \mathcal{U}$, we know $\tilde{\beta}_u - \tilde{\beta}_a \leq \hat{\beta}_u - \hat{\beta}_a$. Hence, we can write $\Delta_{ua}\tilde{\beta} = \Delta_{ua}\hat{\beta} - \lambda_u$,[15] where $\lambda_u \geq 0$ for each $u \in \mathcal{U}$. We now compare $\tilde{\mathcal{L}}(\tilde{\beta}_{\mathcal{U}}, \tilde{\beta}_{\mathcal{V}})$ and $\tilde{\mathcal{L}}(\tilde{\beta}_{\mathcal{U}}, \hat{\beta}_{\mathcal{V}})$.[16] In other words, we compare

$$\begin{aligned}
&\sum_{u\in\mathcal{U}} \ell_{ua}(\Delta_{ua}\tilde{\beta}) + \sum_{v\in\mathcal{V}} \ell_{va}(\Delta_{va}\tilde{\beta}) + \sum_{\{u,p\}\subseteq\mathcal{U}} \ell_{up}(\Delta_{up}\tilde{\beta}) \\
&\qquad + \sum_{\{v,q\}\subseteq\mathcal{V}} \ell_{vq}(\Delta_{vq}\tilde{\beta}) + \sum_{(u,v)\in\mathcal{U}\times\mathcal{V}} \ell_{vu}(\Delta_{vu}\tilde{\beta}) + \alpha \log F(\Delta_{ab}\tilde{\beta}) \\
&\qquad\qquad vs \\
&\sum_{u\in\mathcal{U}} \ell_{ua}(\Delta_{ua}\tilde{\beta}) + \sum_{v\in\mathcal{V}} \ell_{va}(\Delta_{va}\hat{\beta}) + \sum_{\{u,p\}\subseteq\mathcal{U}} \ell_{up}(\Delta_{up}\tilde{\beta}) \\
&\qquad + \sum_{\{v,q\}\subseteq\mathcal{V}} \ell_{vq}(\Delta_{vq}\hat{\beta}) + \sum_{(u,v)\in\mathcal{U}\times\mathcal{V}} \ell_{vu}(\hat{\beta}_v - \tilde{\beta}_u) + \alpha \log F(\Delta_{ab}\tilde{\beta}).
\end{aligned}$$

Note that the last term $\alpha \log F(\Delta_{ab}\beta)$ appears with a $\tilde{\beta}$ in both the equations because we know $b \in \mathcal{U}$. Cancelling terms that appear on both sides (because of the same $\tilde{\beta}_{\mathcal{U}}$), and plugging in $\Delta_{va}\tilde{\beta} = \Delta_{va}\hat{\beta} + \delta_v$ for $v \in \mathcal{V}$, as well as $\Delta_{ua}\tilde{\beta} = \Delta_{ua}\hat{\beta} - \lambda_u$ for $u \in \mathcal{U}$, we are comparing

$$\begin{gathered}
\sum_{v\in\mathcal{V}} \ell_{va}(\Delta_{va}\hat{\beta} + \delta_v) + \sum_{\{v,q\}\subseteq\mathcal{V}} \ell_{vq}(\Delta_{vq}\hat{\beta} + \delta_v - \delta_q) + \sum_{(u,v)\in\mathcal{U}\times\mathcal{V}} \ell_{vu}(\Delta_{vu}\hat{\beta} + \lambda_u + \delta_v) \\
vs \\
\sum_{v\in\mathcal{V}} \ell_{va}(\Delta_{va}\hat{\beta}) + \sum_{\{v,q\}\subseteq\mathcal{V}} \ell_{vq}(\Delta_{vq}\hat{\beta}) + \sum_{(u,v)\in\mathcal{U}\times\mathcal{V}} \ell_{vu}(\Delta_{vu}\hat{\beta} + \lambda_u).
\end{gathered}$$

And, rearranging this, we compare

$$\begin{gathered}
\sum_{(u,v)\in\mathcal{U}\times\mathcal{V}} \ell_{vu}(\Delta_{vu}\hat{\beta} + \lambda_u + \delta_v) - \sum_{(u,v)\in\mathcal{U}\times\mathcal{V}} \ell_{vu}(\Delta_{vu}\hat{\beta} + \lambda_u) \\
vs \\
-\left[\sum_{v\in\mathcal{V}} \ell_{va}(\Delta_{va}\hat{\beta} + \delta_v) + \sum_{\{v,q\}\subseteq\mathcal{V}} \ell_{vq}(\Delta_{vq}\hat{\beta} + \delta_v - \delta_q) - \sum_{v\in\mathcal{V}} \ell_{va}(\Delta_{va}\hat{\beta}) - \sum_{\{v,q\}\subseteq\mathcal{V}} \ell_{vq}(\Delta_{vq}\hat{\beta})\right].
\end{gathered} \tag{8}$$

If λ_u were zero, we know that the left hand side (i.e. the equation placed above in (8)) is smaller (than the one placed below) because of equation (7). But, we now show that this holds even for $\lambda_u \geq 0$ by concavity of the functions ℓ_{vu}. For each $(u, v) \in \mathcal{U} \times \mathcal{V}$, we can write

$$\ell_{vu}(\Delta_{vu}\hat{\beta} + \lambda_u + \delta_v) - \ell_{vu}(\Delta_{vu}\hat{\beta} + \lambda_u) = \int_{\Delta_{vu}\hat{\beta}+\lambda_u}^{\Delta_{vu}\hat{\beta}+\lambda_u+\delta_v} \ell'_{vu}(t)dt,$$

where ℓ'_{vu} is the derivative of ℓ_{vu}.[17] Changing the variable of intergration,

$$\ell_{vu}(\Delta_{vu}\hat{\beta} + \lambda_u + \delta_v) - \ell_{vu}(\Delta_{vu}\hat{\beta} + \lambda_u) = \int_{\Delta_{vu}\hat{\beta}}^{\Delta_{vu}\hat{\beta}+\delta_v} \ell'_{vu}(s + \lambda_u)ds.$$

[15] Equivalently, this could be written as $\tilde{\beta}_u = \hat{\beta}_u - \lambda_u$, as $\tilde{\beta}_a = \hat{\beta}_a = 0$.

[16] That is, we are again keeping the $\mathcal{U}$ part fixed, while changing the $\mathcal{V}$ part from $\tilde{\beta}_{\mathcal{V}}$ to $\hat{\beta}_{\mathcal{V}}$.

[17] which exists, as F is differentiable.

But, we know that ℓ_{vu} is a concave function, implying that ℓ'_{vu} is monotonically decreasing. Hence, $\ell'_{vu}(s+\lambda_u) \leq \ell'_{vu}(s)$ for every s, as $\lambda_u \geq 0$. This gives us

$$\begin{aligned}\ell_{vu}(\Delta_{vu}\hat{\beta}+\lambda_u+\delta_v)-\ell_{vu}(\Delta_{vu}\hat{\beta}+\lambda_u) &= \int_{\Delta_{vu}\hat{\beta}}^{\Delta_{vu}\hat{\beta}+\delta_v} \ell'_{vu}(s+\lambda_u)ds \\ &\leq \int_{\Delta_{vu}\hat{\beta}}^{\Delta_{vu}\hat{\beta}+\delta_v} \ell'_{vu}(s)ds \\ &= \ell_{vu}(\Delta_{vu}\hat{\beta}+\delta_v)-\ell_{vu}(\Delta_{vu}\hat{\beta}).\end{aligned}$$

Taking a summation of the left hand side over all $(u,v) \in \mathcal{U} \times \mathcal{V}$, shows that this summation is less than or equal to the left hand side of Equation (7). Hence, this summation is strictly smaller than the right hand side of Equation (7) (because of Equation (7) itself). This in turn implies that the equation placed above in (8) is strictly smaller than the one placed below. In other words, $\tilde{\mathcal{L}}(\tilde{\beta}) < \tilde{\mathcal{L}}(\tilde{\beta}_{\mathcal{U}}, \hat{\beta}_{\mathcal{V}})$, contradicting the fact that $\tilde{\beta}$ is the maximizer of $\tilde{\mathcal{L}}$. Hence, $V = \phi$. In other words, for each $x \in \mathcal{X} \setminus \{a\}$, $\tilde{\beta}_x - \tilde{\beta}_a \leq \hat{\beta}_x - \hat{\beta}_a$. □

E Proof of Theorem 5.3

Consider $\mathcal{X} = \{a,b,c\}$, and let the dataset be as follows. $\#\{a \succ b\} = 5+\epsilon, \#\{b \succ a\} = 5, \#\{a \succ c\} = 5+\epsilon, \#\{c \succ a\} = 5, \#\{b \succ c\} = 100$ and $\#\{c \succ b\} = 1$. Here, ϵ is a constant lying in $[0,1]$. Observe that for any $\epsilon > 0$, this dataset conforms to Definition 5.1 if we label $x_1, x_2, x_3 = a, b, c$. To show violation of PMC, we show that there exists $\epsilon_o \in (0,1]$ for which the (unique) MLE $\hat{\beta}$ violates the corresponding requirement of $\hat{\beta}_a \geq \hat{\beta}_b \geq \hat{\beta}_c$.

The log-likelihood function for this data is given by

$$\begin{aligned}\mathcal{L}_\epsilon(\beta) = (5+\epsilon)\log F(\beta_a-\beta_b) + 5\log F(\beta_b-\beta_a) + 100\log F(\beta_b-\beta_c) + \log F(\beta_c-\beta_b) \\ + (5+\epsilon)\log F(\beta_a-\beta_c) + 5\log F(\beta_c-\beta_a).\end{aligned}$$

Observe that every alternative has been compared with every other alternative, and hence, the comparison graph $\mathcal{G}_\#$ is strongly connected. Further, as F is strictly monotonic, continuous and strictly log-concave, the MLE exists and is unique for any $\epsilon \in [0,1]$ (by Lemmas 2.1 and 2.4). Further, the log-likelihood $\mathcal{L}_\epsilon(\beta)$ is a strictly concave function (for each $\epsilon \in [0,1]$). Any alternative could be set as the reference, but we use c as the reference alternative in this proof for ease of exposition. That is, our domain is $\mathcal{D} = \{\beta \in \mathbb{R}^{\mathcal{X}} : \beta_c = 0\}$. The (unique) maximum likelihood estimator is given by

$$\hat{\beta}(\epsilon) = \underset{\beta \in \mathcal{D}}{\operatorname{argmax}}\ \mathcal{L}_\epsilon(\beta).$$

We first show that $\hat{\beta}(\epsilon)$ is a continuous function of ϵ. As F is strictly monotonic and continuous, for each $\epsilon \in [0,1]$, Lemma 2.3 tells us that the MLE is bounded as

$$\|\hat{\beta}(\epsilon)\|_\infty \leq |\mathcal{X}| \cdot \max_{(x,y)\in\mathcal{X}^2} \delta_\epsilon(x,y),$$

where δ_ϵ is the perfect-fit distance, but is now dependent on ϵ. For the dataset at hand, these perfect-fit distances are given by

$$\delta_\epsilon(a,b) = F^{-1}\left(\frac{5+\epsilon}{10+\epsilon}\right), \quad \delta_\epsilon(b,c) = F^{-1}\left(\frac{100}{101}\right) \quad \text{and} \quad \delta_\epsilon(a,c) = F^{-1}\left(\frac{5+\epsilon}{10+\epsilon}\right).$$

And, $\delta_\epsilon(b,a) = -\delta_\epsilon(a,b), \delta_\epsilon(c,b) = -\delta_\epsilon(b,c)$ and $\delta_\epsilon(c,a) = -\delta_\epsilon(a,c)$, as $F^{-1}(1-x) = -F^{-1}(x)$. Further, the first three distances are non-negative (making the remaining three non-positive) as $F^{-1}(x) \geq 0$ for $x \geq \frac{1}{2}$. Hence, the bound on the MLE simplifies to

$$\|\hat{\beta}(\epsilon)\|_\infty \leq 3 \cdot \max\left(F^{-1}\left(\frac{5+\epsilon}{10+\epsilon}\right), F^{-1}\left(\frac{100}{101}\right)\right).$$

As F is strictly monotonic, it implies that F^{-1} is also strictly increasing. Applying this, we have

$$F^{-1}\left(\frac{5+\epsilon}{10+\epsilon}\right) \leq F^{-1}\left(\frac{6}{11}\right) < F^{-1}\left(\frac{100}{101}\right),$$

as $\epsilon \in [0,1]$. Therefore, the bound on the MLE further simplifies to

$$\|\hat{\beta}(\epsilon)\|_\infty \leq 3\ F^{-1}\left(\frac{100}{101}\right),$$

for any $\epsilon \in [0,1]$. Hence, the MLE optimization problem can be rewritten as

$$\hat{\beta}(\epsilon) = \underset{\beta \in \mathcal{D}: \|\beta\|_\infty \leq 3\ F^{-1}\left(\frac{100}{101}\right)}{\operatorname{argmax}} \mathcal{L}_\epsilon(\beta).$$

This shows that we are optimizing over a compact space. Hence, by the Theorem of the Maximum [2, 9], both the maximum likelihood and the corresponding maximizer $\hat{\beta}(\epsilon)$ are continuous in the parameter ϵ, for all $\epsilon \in [0,1]$.

Next, we analyze the MLE at $\epsilon = 0$. The log-likelihood function for this value of ϵ is

$$\begin{aligned}\mathcal{L}_0(\beta) = 5\log F(\beta_a - \beta_b) + 5\log F(\beta_b - \beta_a) + 100\log F(\beta_b - \beta_c) + \log F(\beta_c - \beta_b)\\ + 5\log F(\beta_a - \beta_c) + 5\log F(\beta_c - \beta_a). \qquad (9)\end{aligned}$$

For ease of exposition, we use $\beta^\dagger$ to denote the MLE when $\epsilon = 0$, i.e.

$$\beta^\dagger := \hat{\beta}(0) = \underset{\beta \in \mathcal{D}}{\operatorname{argmax}}\ \mathcal{L}_0(\beta).$$

Recall that we used c as the reference alternative, and hence, $\beta^\dagger_c = 0$. Our goal is to show that $\beta^\dagger_b > \beta^\dagger_a > \beta^\dagger_c$. To this end, we first show that $\beta^\dagger_a = \beta^\dagger_b/2$, i.e. in terms of β values, a lies at the mid-point of b and c. Suppose for the sake of contradiction that $\beta^\dagger_a \neq \beta^\dagger_b/2$. Consider another vector $\tilde{\beta} \in \mathcal{D}$ that is the same as $\beta^\dagger$, except with the distances between β values of b & a and a & c swapped. This can be achieved by setting

$$\tilde{\beta}_x = \begin{cases} \beta^\dagger_x & ; \text{ if } x \in \{b,c\} \\ \beta^\dagger_b - \beta^\dagger_a & ; \text{ if } x = a. \end{cases}$$

Then, we have $\tilde{\beta}_a - \tilde{\beta}_c = \beta^\dagger_b - \beta^\dagger_a$ and $\tilde{\beta}_b - \tilde{\beta}_a = \beta^\dagger_a - \beta^\dagger_c$. Hence, the log-likelihood at this point is given by

$$\begin{aligned}\mathcal{L}_0(\tilde{\beta}) &= 5\log F(\tilde{\beta}_a - \tilde{\beta}_b) + 5\log F(\tilde{\beta}_b - \tilde{\beta}_a) + 100\log F(\tilde{\beta}_b - \tilde{\beta}_c) + \log F(\tilde{\beta}_c - \tilde{\beta}_b)\\ &\quad + 5\log F(\tilde{\beta}_a - \tilde{\beta}_c) + 5\log F(\tilde{\beta}_c - \tilde{\beta}_a)\\ &= 5\log F(\beta^\dagger_c - \beta^\dagger_a) + 5\log F(\beta^\dagger_a - \beta^\dagger_c) + 100\log F(\beta^\dagger_b - \beta^\dagger_c) + \log F(\beta^\dagger_c - \beta^\dagger_b)\\ &\quad + 5\log F(\beta^\dagger_b - \beta^\dagger_a) + 5\log F(\beta^\dagger_a - \beta^\dagger_b)\\ &= \mathcal{L}_0(\beta^\dagger).\end{aligned}$$

That is, swapping these distances does not change the likelihood, because of symmetry. Now, consider a new vector $\bar{\beta} = (\beta^\dagger + \tilde{\beta})/2$. Note that, as $\beta^\dagger_a \neq \beta^\dagger_b/2$, it implies that $\beta^\dagger_a \neq \beta^\dagger_b - \beta^\dagger_a = \tilde{\beta}_a$. In other words, $\tilde{\beta} \neq \beta^\dagger$. Therefore, applying strict concavity of $\mathcal{L}_0$, we have

$$\mathcal{L}_0(\bar{\beta}) = \mathcal{L}_0\left(\frac{\beta^\dagger + \tilde{\beta}}{2}\right) > \frac{\mathcal{L}_0(\beta^\dagger) + \mathcal{L}_0(\tilde{\beta})}{2} = \mathcal{L}_0(\beta^\dagger),$$

which is a contradiction as $\beta^\dagger$ is the maximizer of $\mathcal{L}_0$. This proves that $\beta^\dagger_a = \beta^\dagger_b/2$. In other words, $\beta^\dagger$ is of the form $(\beta^\dagger_b/2, \beta^\dagger_b, 0)$. Hence, $\beta^\dagger$ continues to be the maximizer of $\mathcal{L}_0$ among the vectors $\mathcal{A} = \{(\alpha/2, \alpha, 0) : \alpha \in \mathbb{R}\} \subseteq \mathcal{D}$. Rewriting the log-likelihood (9) for vectors in $\mathcal{A}$, we have

$$\begin{aligned}\mathcal{L}_0((\alpha/2, \alpha, 0)) &= 5\log F\left(-\frac{\alpha}{2}\right) + 5\log F\left(\frac{\alpha}{2}\right) + 100\log F(\alpha) + \log F(-\alpha)\\ &\quad + 5\log F\left(\frac{\alpha}{2}\right) + 5\log F\left(-\frac{\alpha}{2}\right)\\ &= 10\log F\left(\frac{\alpha}{2}\right) + 10\log F\left(-\frac{\alpha}{2}\right) + 100\log F(\alpha) + \log F(-\alpha).\end{aligned}$$

Overloading notation, we denote this log-likelihood by $\mathcal{L}_0(\alpha)$, and this is maximized at $\alpha = \beta^\dagger_b$. For ease of exposition, denote the composition of log and F by G, i.e. $G := \log F$. As F is strictly monotonic, differentiable and strictly log-concave, G is also strictly monotonic and differentiable, and is strictly concave.[18] Rewriting the log-likelihood with this notation, we have

$$\mathcal{L}_0(\alpha) = 10G\left(\frac{\alpha}{2}\right) + 10G\left(-\frac{\alpha}{2}\right) + 100G(\alpha) + G(-\alpha).$$

We show that this function is not maximized at any $\alpha \le 0$. In other words, $\beta^\dagger_b > 0$. Computing the derivative of $\mathcal{L}_0$, we have

$$\mathcal{L}'_0(\alpha) = 5G'\left(\frac{\alpha}{2}\right) - 5G'\left(-\frac{\alpha}{2}\right) + 100G'(\alpha) - G'(-\alpha).$$

As G is strictly concave, it implies that G' is strictly decreasing. Hence, for $\alpha \le 0$, it implies that $G'(\frac{\alpha}{2}) \ge G'(-\frac{\alpha}{2})$ and $G'(\alpha) \ge G'(-\alpha)$. This shows that for $\alpha \le 0$, we have

$$\mathcal{L}'_0(\alpha) \ge 99G'(\alpha) > 0,$$

where the last inequality holds as G is a strictly increasing function, leading to G' being positive.[19] In other words, if $\alpha \le 0$, the log-likelihood can be strictly increased by taking an infinitesimally small step in the direction $\left[\frac{1}{2}, 1, 0\right]$. Hence, none of these points maximizes $\mathcal{L}_0$, and $\beta^\dagger_b > 0$. Also, recall that $\beta^\dagger$ was of the form $(\beta^\dagger_b/2, \beta^\dagger_b, 0)$; this proves that $\beta^\dagger_b > \beta^\dagger_a > \beta^\dagger_c$.

Finally, recall that we need $\epsilon > 0$ for the dataset to conform to Definition 5.1 with the labelling $x_1, x_2, x_3, = a, b, c$. To be able to find such a value of ϵ, we use continuity of $\hat{\beta}(\epsilon)$. By continuity, we know that for every $\gamma > 0$, there exists $\delta > 0$ such that $\|\hat{\beta}(\epsilon) - \hat{\beta}(0)\|_\infty < \gamma$ for all $|\epsilon - 0| < \delta$. Define $\theta := \beta^\dagger_b - \beta^\dagger_a > 0$. Then, choose $\gamma = \theta/3$, and let δ_o denote the corresponding value of δ. Hence, choose $\epsilon_o = \min(\delta_o/2, 1) > 0$. For this value of ϵ_o, we indeed have $\|\hat{\beta}(\epsilon_o) - \hat{\beta}(0)\|_\infty < \theta/3$. That is,

$$\hat{\beta}(\epsilon_o)_b > \beta^\dagger_b - \frac{\theta}{3} \qquad \text{and} \qquad \hat{\beta}(\epsilon_o)_a < \beta^\dagger_a + \frac{\theta}{3}.$$

Hence,

$$\hat{\beta}(\epsilon_o)_b - \hat{\beta}(\epsilon_o)_a > \beta^\dagger_b - \beta^\dagger_a - \frac{2\theta}{3} = \theta - \frac{2\theta}{3} > 0.$$

Therefore, at $\epsilon = \epsilon_o \in (0, 1]$, the MLE satisfies $\hat{\beta}(\epsilon_o)_b > \hat{\beta}(\epsilon_o)_a$. Hence, the dataset with $\epsilon = \epsilon_o$ satisfies the PMC condition, but the corresponding MLE does not conform to the corresponding ordering, proving violation of pairwise majority consistency. □

F Proof of Theorem 6.3

Consider $\mathcal{X} = \{a, b, c\}$, and let the two datasets be as follows. The first dataset is such that $\#^1\{a \succ c\} = 5 + \epsilon, \#^1\{c \succ a\} = 5 - \epsilon, \#^1\{c \succ b\} = 100, \#^1\{b \succ c\} = 1$, and has zero counts otherwise. The second dataset is such that $\#^2\{a \succ c\} = 5 + \epsilon, \#^2\{c \succ a\} = 5 - \epsilon, \#^2\{b \succ a\} = 100, \#^2\{a \succ b\} = 1$, and has zero counts otherwise. Here, ϵ is a constant lying in $(0, 1]$.

First, we analyze the MLE for the dataset $\#^1$. As the comparison graph $\mathcal{G}_{\#^1}$ is strongly connected, and F is strictly monotonic, continuous and strictly log-concave, the MLE $\hat{\beta}^1$ exists and is unique (by Lemmas 2.1 and 2.4). Further, the pair (a, c) satisfies the condition of Lemma 2.2, similarly does the pair (b, c). Applying the lemma for the pair (a, c) says that the MLE satisfies

$$\hat{\beta}^1_a = \hat{\beta}^1_c + F^{-1}\left(\frac{5+\epsilon}{10}\right) > \hat{\beta}^1_c,$$

[18] This part of the proof does not require concavity of G, but we use it nevertheless as it simplifies the proof.

[19] Strictly speaking, a function might be strictly increasing and have a derivative that is not strictly positive at every point (in particular, the derivative might be zero at stationary points). But in our case, as G' is also a strictly decreasing function, it cannot be zero at any point, because that would make it negative at larger points, violating strict monotonicity of G.

as $(5+\epsilon)/10$ is larger than $1/2$. Similarly, applying Lemma 2.2 for the pair (b, c) says that the MLE satisfies

$$\hat{\beta}^1_b = \hat{\beta}^1_c + F^{-1}\left(\frac{1}{1+100}\right) < \hat{\beta}^1_c,$$

as $1/101$ is smaller than $1/2$. Putting these equations together, we have $\hat{\beta}^1_a > \hat{\beta}^1_c > \hat{\beta}^1_b$.

Next, we analyze the MLE for the dataset $\#^2$. As the comparison graph $\mathcal{G}_{\#^2}$ is strongly connected, and F is strictly monotonic, continuous and strictly log-concave, the MLE $\hat{\beta}^2$ exists and is unique (by Lemmas 2.1 and 2.4). Further, the pair (c, a) satisfies the condition of Lemma 2.2, similarly does the pair (b, a). Applying the lemma for the pair (c, a) says that the MLE satisfies

$$\hat{\beta}^2_c = \hat{\beta}^2_a + F^{-1}\left(\frac{5-\epsilon}{10}\right) < \hat{\beta}^2_a,$$

as $(5-\epsilon)/10$ is smaller than $1/2$. Similarly, applying Lemma 2.2 for the pair (b, a) says that the MLE satisfies

$$\hat{\beta}^2_b = \hat{\beta}^2_a + F^{-1}\left(\frac{100}{1+100}\right) > \hat{\beta}^2_a,$$

as $100/101$ is larger than $1/2$. Putting these equations together, we have $\hat{\beta}^2_b > \hat{\beta}^2_a > \hat{\beta}^2_c$. Hence, both datasets $\#^1$ and $\#^2$ have MLEs $\hat{\beta}^1$ and $\hat{\beta}^2$ such that $\hat{\beta}^1_a > \hat{\beta}^1_c$ and $\hat{\beta}^2_a > \hat{\beta}^2_c$.

Finally, we analyze the MLE for the dataset $\# = \#^1 + \#^2$ obtained by pooling both datasets $\#^1$ and $\#^2$. Recall that the proof so far holds for any constant $\epsilon \in (0, 1]$; but, from this point on, we allow ϵ to take the value of zero as well, i.e. $\epsilon \in [0, 1]$. The log-likelihood function for the pooled data $\#$ is given by

$$\begin{aligned}\mathcal{L}_\epsilon(\beta) = 100 \log F(\beta_c - \beta_b) + \log F(\beta_b - \beta_c) + 100 \log F(\beta_b - \beta_a) + \log F(\beta_a - \beta_b)\\ + (10+2\epsilon)\log F(\beta_a - \beta_c) + (10-2\epsilon)\log F(\beta_c - \beta_a).\end{aligned}$$

Observe that every alternative has been compared with every other alternative, and hence, the comparison graph $\mathcal{G}_\#$ is strongly connected. Further, as F is strictly monotonic, continuous and strictly log-concave, the MLE exists and is unique for any $\epsilon \in [0, 1]$ (by Lemmas 2.1 and 2.4). Further, the log-likelihood $\mathcal{L}_\epsilon(\beta)$ is a strictly concave function (for each $\epsilon \in [0, 1]$). Any alternative could be set as the reference, but we use a as the reference alternative in this proof for ease of exposition. That is, our domain is $\mathcal{D} = \{\beta \in \mathbb{R}^\mathcal{X} : \beta_a = 0\}$. The (unique) maximum likelihood estimator is given by

$$\hat{\beta}(\epsilon) = \operatorname*{argmax}_{\beta \in \mathcal{D}} \mathcal{L}_\epsilon(\beta).$$

Similar to the proof of Theorem 5.3, we first show that $\hat{\beta}(\epsilon)$ is a continuous function of ϵ. As F is strictly monotonic and continuous, for each $\epsilon \in [0, 1]$, Lemma 2.3 tells us that the MLE is bounded as

$$\|\hat{\beta}(\epsilon)\|_\infty \leq |\mathcal{X}| \cdot \max_{(x,y)\in\mathcal{X}^2} \delta_\epsilon(x, y),$$

where δ_ϵ is the perfect-fit distance, but is now dependent on ϵ. For our pooled dataset, these perfect-fit distances are given by

$$\delta_\epsilon(b, a) = F^{-1}\left(\frac{100}{101}\right), \quad \delta_\epsilon(c, b) = F^{-1}\left(\frac{100}{101}\right) \quad \text{and} \quad \delta_\epsilon(a, c) = F^{-1}\left(\frac{10+2\epsilon}{20}\right).$$

And, $\delta_\epsilon(a, b) = -\delta_\epsilon(b, a), \delta_\epsilon(b, c) = -\delta_\epsilon(c, b)$ and $\delta_\epsilon(c, a) = -\delta_\epsilon(a, c)$, as $F^{-1}(1-x) = -F^{-1}(x)$. Further, the first three distances are non-negative (making the remaining three non-positive) as $F^{-1}(x) \geq 0$ for $x \geq \frac{1}{2}$. Hence, the bound on the MLE simplifies to

$$\|\hat{\beta}(\epsilon)\|_\infty \leq 3 \cdot \max\left(F^{-1}\left(\frac{100}{101}\right), F^{-1}\left(\frac{10+2\epsilon}{20}\right)\right).$$

As F is strictly monotonic, it implies that F^{-1} is also strictly increasing. Applying this, we have

$$F^{-1}\left(\frac{10+2\epsilon}{20}\right) \leq F^{-1}\left(\frac{12}{20}\right) < F^{-1}\left(\frac{100}{101}\right),$$

as $\epsilon \in [0,1]$. Therefore, the bound on the MLE further simplifies to

$$\|\hat{\beta}(\epsilon)\|_\infty \leq 3\ F^{-1}\left(\frac{100}{101}\right),$$

for any $\epsilon \in [0,1]$. Hence, the MLE optimization problem can be rewritten as

$$\hat{\beta}(\epsilon) = \underset{\beta \in \mathcal{D}: \|\beta\|_\infty \leq 3\ F^{-1}\left(\frac{100}{101}\right)}{\operatorname{argmax}} \mathcal{L}_\epsilon(\beta).$$

This shows that we are optimizing over a compact space. Hence, by the Theorem of the Maximum, both the maximum likelihood and the corresponding maximizer $\hat{\beta}(\epsilon)$ are continuous in the parameter ϵ, for all $\epsilon \in [0,1]$.

Next, we analyze the MLE at $\epsilon = 0$. The log-likelihood function for this value of ϵ is

$$\begin{aligned}\mathcal{L}_0(\beta) = {} & 100 \log F(\beta_c - \beta_b) + \log F(\beta_b - \beta_c) + 100 \log F(\beta_b - \beta_a) + \log F(\beta_a - \beta_b) \\ & + 10 \log F(\beta_a - \beta_c) + 10 \log F(\beta_c - \beta_a). \end{aligned} \tag{10}$$

For ease of exposition, we use $\beta^\dagger$ to denote the MLE when $\epsilon = 0$, i.e.

$$\beta^\dagger := \hat{\beta}(0) = \underset{\beta \in \mathcal{D}}{\operatorname{argmax}}\ \mathcal{L}_\epsilon(\beta).$$

Recall that we used a as the reference alternative, and hence, $\beta_a^\dagger = 0$. Our goal is to show that $\beta_c^\dagger > \beta_b^\dagger > \beta_a^\dagger$. To this end, we first show that $\beta_b^\dagger = \beta_c^\dagger/2$, i.e. in terms of β values, b lies at the mid-point of c and a. Suppose for the sake of contradiction that $\beta_b^\dagger \neq \beta_c^\dagger/2$. Consider another vector $\tilde{\beta} \in \mathcal{D}$ that is the same as $\beta^\dagger$, except with the distances between c & b and b & a swapped. This can be achieved by setting

$$\tilde{\beta}_x = \begin{cases} \beta_x^\dagger & ;\ \text{if } x \neq \{a, c\} \\ \beta_c^\dagger - \beta_b^\dagger & ;\ \text{if } x = b. \end{cases}$$

Then, we have $\tilde{\beta}_b - \tilde{\beta}_a = \beta_c^\dagger - \beta_b^\dagger$ and $\tilde{\beta}_c - \tilde{\beta}_b = \beta_b^\dagger - \beta_a^\dagger$. Hence, the likelihood at this point is given by

$$\begin{aligned}\mathcal{L}_0(\tilde{\beta}) &= 100 \log F(\tilde{\beta}_c - \tilde{\beta}_b) + \log F(\tilde{\beta}_b - \tilde{\beta}_c) + 100 \log F(\tilde{\beta}_b - \tilde{\beta}_a) + \log F(\tilde{\beta}_a - \tilde{\beta}_b) \\ &\quad + 10 \log F(\tilde{\beta}_a - \tilde{\beta}_c) + 10 \log F(\tilde{\beta}_c - \tilde{\beta}_a) \\ &= 100 \log F(\beta_b^\dagger - \beta_a^\dagger) + \log F(\beta_a^\dagger - \beta_b^\dagger) + 100 \log F(\beta_c^\dagger - \beta_b^\dagger) + \log F(\beta_b^\dagger - \beta_c^\dagger) \\ &\quad + 10 \log F(\beta_a^\dagger - \beta_c^\dagger) + 10 \log F(\beta_c^\dagger - \beta_a^\dagger) \\ &= \mathcal{L}_0(\beta^\dagger).\end{aligned}$$

That is, swapping these distances does not change the likelihood, because of symmetry. Now, consider a new vector $\bar{\beta} = (\beta^\dagger + \tilde{\beta})/2$. Note that, as $\beta_b^\dagger \neq \beta_c^\dagger/2$, it implies that $\beta_b^\dagger \neq \beta_c^\dagger - \beta_b^\dagger = \tilde{\beta}_b$. In other words, $\tilde{\beta} \neq \beta^\dagger$. Therefore, applying strict concavity of $\mathcal{L}_0$, we have

$$\mathcal{L}_0(\bar{\beta}) = \mathcal{L}_0\left(\frac{\beta^\dagger + \tilde{\beta}}{2}\right) > \frac{\mathcal{L}_0(\beta^\dagger) + \mathcal{L}_0(\tilde{\beta})}{2} = \mathcal{L}_0(\beta^\dagger),$$

which is a contradiction as $\beta^\dagger$ is the maximizer of $\mathcal{L}_0$. This proves that $\beta_b^\dagger = \beta_c^\dagger/2$. In other words, $\beta^\dagger$ is of the form $(0, \beta_c^\dagger/2, \beta_c^\dagger)$. Hence, $\beta^\dagger$ continues to be the maximizer of $\mathcal{L}_0$ among the vectors $\mathcal{A} = \{(0, \alpha/2, \alpha) : \alpha \in \mathbb{R}\} \subseteq \mathcal{D}$. Rewriting the log-likelihood (10) for vectors in $\mathcal{A}$, we have

$$\begin{aligned}\mathcal{L}_0((0, \alpha/2, \alpha)) &= 100 \log F\left(\frac{\alpha}{2}\right) + \log F\left(-\frac{\alpha}{2}\right) + 100 \log F\left(\frac{\alpha}{2}\right) + \log F\left(-\frac{\alpha}{2}\right) \\ &\quad + 10 \log F(-\alpha) + 10 \log F(\alpha) \\ &= 200 \log F\left(\frac{\alpha}{2}\right) + 2 \log F\left(-\frac{\alpha}{2}\right) + 10 \log F(\alpha) + 10 \log F(-\alpha).\end{aligned}$$

Overloading notation, we denote this log-likelihood by $\mathcal{L}_0(\alpha)$, and this is maximized at $\alpha = \beta_c^\dagger$. For ease of exposition, denote the composition of log and F by G, i.e. $G := \log F$. As F is strictly monotonic, differentiable and strictly log-concave, G is also strictly monotonic and differentiable, and is strictly concave.[20] Rewriting the log-likelihood with this notation, we have

$$\mathcal{L}_0(\alpha) = 200G\left(\frac{\alpha}{2}\right) + 2G\left(-\frac{\alpha}{2}\right) + 10G(\alpha) + 10G(-\alpha).$$

We show that this function is not maximized at any $\alpha \leq 0$. In other words, $\beta_c^\dagger > 0$. Computing the derivative of $\mathcal{L}_0$, we have

$$\mathcal{L}_0'(\alpha) = 100G'\left(\frac{\alpha}{2}\right) - G'\left(-\frac{\alpha}{2}\right) + 10G'(\alpha) - 10G'(-\alpha).$$

As G is strictly concave, it implies that G' is strictly decreasing. Hence, for $\alpha \leq 0$, it implies that $G'(\frac{\alpha}{2}) \geq G'(-\frac{\alpha}{2})$ and $G'(\alpha) \geq G'(-\alpha)$. This shows that for $\alpha \leq 0$, we have

$$\mathcal{L}_0'(\alpha) \geq 99G'\left(\frac{\alpha}{2}\right) > 0,$$

where the last inequality holds as G is a strictly increasing function, leading to G' being positive.[21] In other words, if $\alpha \leq 0$, the log-likelihood can be strictly increased by taking an infinitesimally small step in the direction $\left[0, \frac{1}{2}, 1\right]$. Hence, none of these points maximizes $\mathcal{L}_0$, and $\beta_c^\dagger > 0$. Also, recall that $\beta^\dagger$ was of the form $(0, \beta_c^\dagger/2, \beta_c^\dagger)$; this proves that $\beta_c^\dagger > \beta_b^\dagger > \beta_a^\dagger$.

Finally, recall that the initial part of the proof (analyzing the MLE for the individual datasets) works only for $0 < \epsilon \leq 1$. Hence, we need to use an ϵ value strictly larger than zero even for the pooled dataset. To be able to find such a value of ϵ, we use continuity of $\hat{\beta}(\epsilon)$. By continuity, we know that for every $\gamma > 0$, there exists $\delta > 0$ such that $\|\hat{\beta}(\epsilon) - \hat{\beta}(0)\|_\infty < \gamma$ for all $|\epsilon - 0| < \delta$. Define $\theta := \beta_c^\dagger - \beta_a^\dagger > 0$. Then, choose $\gamma = \theta/3$, and let δ_o denote the corresponding value of δ. Hence, choose $\epsilon_o = \min(\delta_o/2, 1) > 0$. For this value of ϵ_o, we indeed have $\|\hat{\beta}(\epsilon_o) - \hat{\beta}(0)\|_\infty < \theta/3$. That is,

$$\hat{\beta}(\epsilon_o)_c > \beta_c^\dagger - \frac{\theta}{3} \qquad \text{and} \qquad \hat{\beta}(\epsilon_o)_a < \beta_a^\dagger + \frac{\theta}{3}.$$

Hence,

$$\hat{\beta}(\epsilon_o)_c - \hat{\beta}(\epsilon_o)_a > \beta_c^\dagger - \beta_a^\dagger - \frac{2\theta}{3} = \theta - \frac{2\theta}{3} > 0.$$

Therefore, at $\epsilon = \epsilon_o \in (0, 1]$, the MLE (on the pooled data) satisfies $\hat{\beta}(\epsilon_o)_c > \hat{\beta}(\epsilon_o)_a$. Hence, for $\epsilon = \epsilon_o$, the two datasets $\#^1$ and $\#^2$ have MLEs $\hat{\beta}^1$ and $\hat{\beta}^2$ such that $\hat{\beta}_a^1 > \hat{\beta}_c^1$ and $\hat{\beta}_a^2 > \hat{\beta}_c^2$, but the MLE $\hat{\beta}$ on the pooled dataset $\# = \#^1 + \#^2$ satisfies $\hat{\beta}_a < \hat{\beta}_c$, proving violation of separability. □

[20] This part of the proof does not require concavity of G, but we use it nevertheless as it simplifies the proof.

[21] Strictly speaking, a function might be strictly increasing and have a derivative that is not strictly positive at every point (in particular, the derivative might be zero at stationary points). But in our case, as G' is also a strictly decreasing function, it cannot be zero at any point, because that would make it negative at larger points, violating strict monotonicity of G.