IP Scoring Rules: Foundations and Applications

Jason Konek

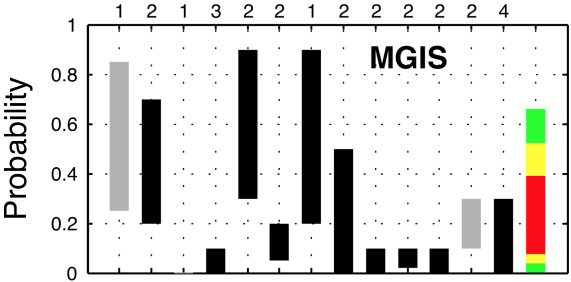
Department of Philosophy
University of Bristol

11^{th} International Symposium on Imprecise Probabilities
4 July 2019

Aggregating Interval Forecasts

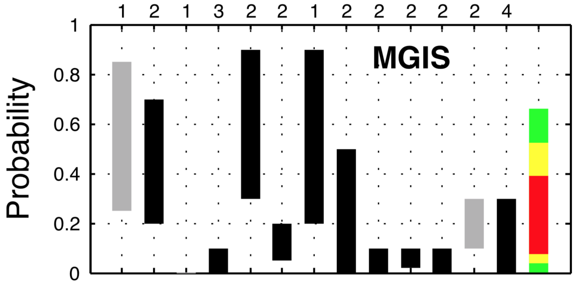
MGIS: Greenland ice sheet will melt by 2200

Aggregating Interval Forecasts



(Kriegler, 2009)

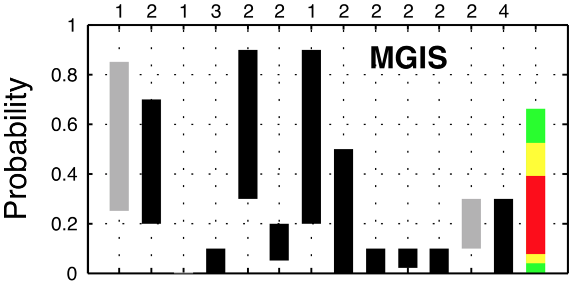
Aggregating Interval Forecasts



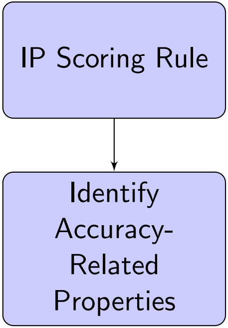
(Kriegler, 2009)

IP Scoring Rule

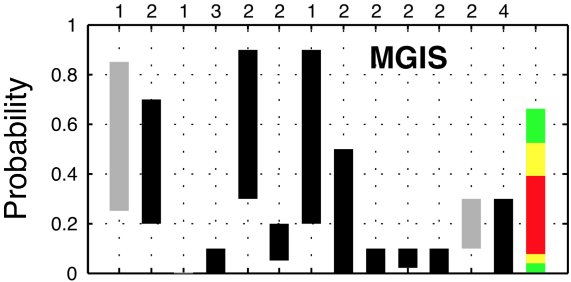
Aggregating Interval Forecasts



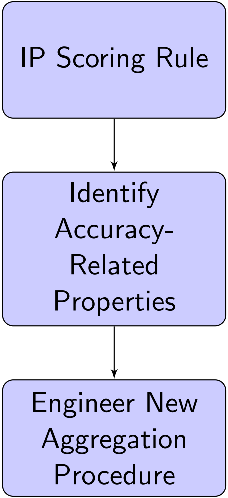
(Kriegler, 2009)



Aggregating Interval Forecasts



(Kriegler, 2009)



IP Scoring Rules: generalised type I and type II error

- An IP scoring rule is a loss function $\mathcal{I}$ which maps an IP distribution $\mathcal{C}$ (lower probabilities, lower previsions, sets of probabilities, etc.) and a world w to a non-negative real number, $\mathcal{I}(\mathcal{C}, w)$.

IP Scoring Rules: generalised type I and type II error

- An IP scoring rule is a loss function $\mathcal{I}$ which maps an IP distribution $\mathcal{C}$ (lower probabilities, lower previsions, sets of probabilities, etc.) and a world w to a non-negative real number, $\mathcal{I}(\mathcal{C}, w)$.
  - $\mathcal{I}(\mathcal{C}, w)$ measures the extent to which $\mathcal{C}$ fails to avoid various types of alethic (accuracy-related) error.

IP Scoring Rules: generalised type I and type II error

- An IP scoring rule is a loss function $\mathcal{I}$ which maps an IP distribution $\mathcal{C}$ (lower probabilities, lower previsions, sets of probabilities, etc.) and a world w to a non-negative real number, $\mathcal{I}(\mathcal{C}, w)$.
  - $\mathcal{I}(\mathcal{C}, w)$ measures the extent to which $\mathcal{C}$ fails to avoid various types of alethic (accuracy-related) error.
- Two principal types of alethic error: *generalised type I and type II error*.

IP Scoring Rules: generalised type I and type II error

- An IP scoring rule is a loss function $\mathcal{I}$ which maps an IP distribution $\mathcal{C}$ (lower probabilities, lower previsions, sets of probabilities, etc.) and a world w to a non-negative real number, $\mathcal{I}(\mathcal{C}, w)$.
  - $\mathcal{I}(\mathcal{C}, w)$ measures the extent to which $\mathcal{C}$ fails to avoid various types of alethic (accuracy-related) error.
- Two principal types of alethic error: *generalised type I and type II error*.
  - $\mathcal{C}$ *avoids generalised type I error* to the extent that it leaves open accurate previsions/estimates.

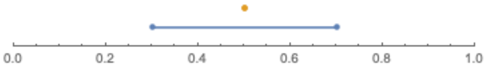
IP Scoring Rules: generalised type I and type II error

- An IP scoring rule is a loss function $\mathcal{I}$ which maps an IP distribution $\mathcal{C}$ (lower probabilities, lower previsions, sets of probabilities, etc.) and a world w to a non-negative real number, $\mathcal{I}(\mathcal{C}, w)$.
  - $\mathcal{I}(\mathcal{C}, w)$ measures the extent to which $\mathcal{C}$ fails to avoid various types of alethic (accuracy-related) error.
- Two principal types of alethic error: *generalised type I and type II error*.
  - $\mathcal{C}$ *avoids generalised type I error* to the extent that it leaves open accurate previsions/estimates.
  - $\mathcal{C}$ *avoids generalised type II error* to the extent that it rules out inaccurate previsions/estimates.

IP Scoring Rules: generalised type I and type II error

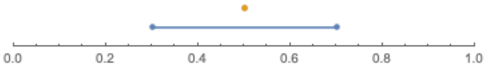
- An IP scoring rule is a loss function $\mathcal{I}$ which maps an IP distribution $\mathcal{C}$ (lower probabilities, lower previsions, sets of probabilities, etc.) and a world w to a non-negative real number, $\mathcal{I}(\mathcal{C}, w)$.
  - $\mathcal{I}(\mathcal{C}, w)$ measures the extent to which $\mathcal{C}$ fails to avoid various types of alethic (accuracy-related) error.
- Two principal types of alethic error: *generalised type I and type II error*.
  - $\mathcal{C}$ *avoids generalised type I error* to the extent that it leaves open accurate previsions/estimates.
  - $\mathcal{C}$ *avoids generalised type II error* to the extent that it rules out inaccurate previsions/estimates.
- $\mathcal{I}_\alpha(\mathcal{C}, w) = \alpha \cdot \min_{c \in \mathcal{C}} I(c, w) + (1 - \alpha) \cdot \max_{c \in \mathcal{C}} I(c, w)$

Impossibility Theorems



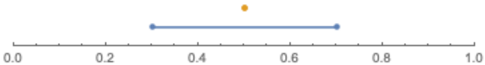
- $\mathcal{I}_{0.7}([.3, .7], 0) = \mathcal{I}_{0.7}([.3, .7], 1) = 0.21$

Impossibility Theorems



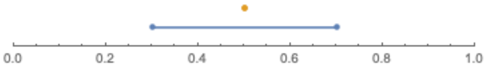
- $\mathcal{I}_{0.7}([.3, .7], 0) = \mathcal{I}_{0.7}([.3, .7], 1) = 0.21$
- $\mathcal{I}_{0.7}([.5, .5], 0) = \mathcal{I}_{0.7}([.5, .5], 1) = 0.25$

Impossibility Theorems



- $\mathcal{I}_{0.7}([.3,.7],0) = \mathcal{I}_{0.7}([.3,.7],1) = 0.21$
- $\mathcal{I}_{0.7}([.5,.5],0) = \mathcal{I}_{0.7}([.5,.5],1) = 0.25$
- Seidenfeld et al. (2012), Mayo-Wilson and Wheeler (2015), and Schoenfield (2015) show that any continuous IP scoring rule $\mathcal{I}$ renders some IP distribution dominated.

Impossibility Theorems



- $\mathcal{I}_{0.7}([.3,.7],0) = \mathcal{I}_{0.7}([.3,.7],1) = 0.21$
- $\mathcal{I}_{0.7}([.5,.5],0) = \mathcal{I}_{0.7}([.5,.5],1) = 0.25$
- Seidenfeld et al. (2012), Mayo-Wilson and Wheeler (2015), and Schoenfield (2015) show that any continuous IP scoring rule $\mathcal{I}$ renders some IP distribution dominated.
- Scepticism about using IP scoring rules to provide accuracy-centered foundations for IP, elication

Impossibility Theorems

- Choose an interval forecast $[a, b]$ of event E.

Impossibility Theorems

- Choose an interval forecast $[a, b]$ of event E.
- Though no IP scoring rule renders all interval forecasts non-dominated, each interval forecast is rendered non-dominated by some IP scoring rule

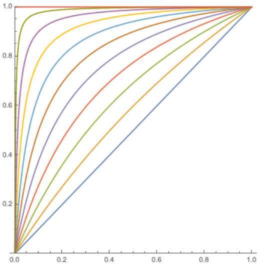
Impossibility Theorems

- Choose an interval forecast $[a, b]$ of event E.
- Though no IP scoring rule renders all interval forecasts non-dominated, each interval forecast is rendered non-dominated by some IP scoring rule
- $\mathcal{I}_\alpha$ with

$$\alpha = \frac{-b + ab + \sqrt{ab - a^2 b - ab^2 + a^2 b^2}}{a - b}$$

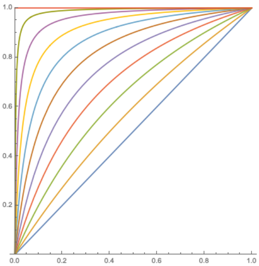
  is the *unique* IP scoring rule of the form proposed in Konek (2019) that renders $[a, b]$ non-dominated.

Impossibility Theorems



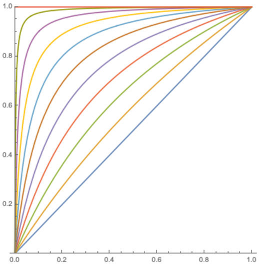
- How $\mathcal{I}_\alpha$ prioritises avoiding type I over type II error (or vice versa) fixes the stock of available interval forecasts.

Impossibility Theorems



- How $\mathcal{I}_\alpha$ prioritises avoiding type I over type II error (or vice versa) fixes the stock of available interval forecasts.
- Where you report 0.5, I report $[0.3, 0.7]$

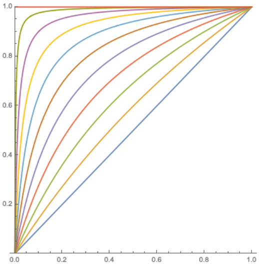
Impossibility Theorems



- How $\mathcal{I}_\alpha$ prioritises avoiding type I over type II error (or vice versa) fixes the stock of available interval forecasts.
- Where you report 0.5, I report $[0.3, 0.7]$

- Different interval forecasts, same evidence

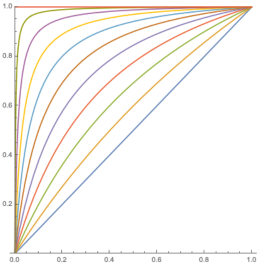
Impossibility Theorems



- How $\mathcal{I}_\alpha$ prioritises avoiding type I over type II error (or vice versa) fixes the stock of available interval forecasts.
- Where you report 0.5, I report $[0.3, 0.7]$

- Different interval forecasts, same evidence
    - Reason: different attitudes toward the comparative importance of avoiding type I and type II error

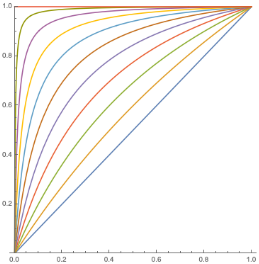
Impossibility Theorems



- How $\mathcal{I}_\alpha$ prioritises avoiding type I over type II error (or vice versa) fixes the stock of available interval forecasts.
- Where you report 0.5, I report $[0.3, 0.7]$

- Different interval forecasts, same evidence
  - Reason: different attitudes toward the comparative importance of avoiding type I and type II error
- Certain ways of prioritising type I over type II error-avoidance (captured by α) justify certain IP methods (more imprecise interval forecasts);

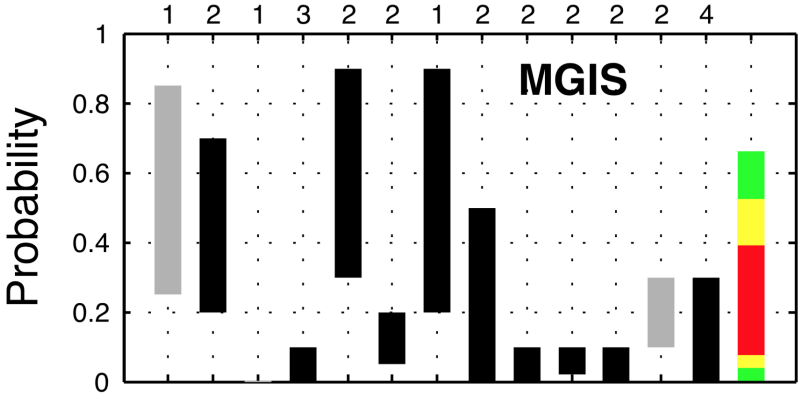
Impossibility Theorems



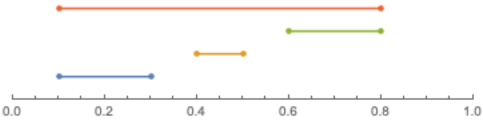
- How $\mathcal{I}_\alpha$ prioritises avoiding type I over type II error (or vice versa) fixes the stock of available interval forecasts.
- Where you report 0.5, I report $[0.3, 0.7]$

- Different interval forecasts, same evidence
  - Reason: different attitudes toward the comparative importance of avoiding type I and type II error
- Certain ways of prioritising type I over type II error-avoidance (captured by α) justify certain IP methods (more imprecise interval forecasts); others justify others (more precise forecasts).

Aggregating Interval Forecasts

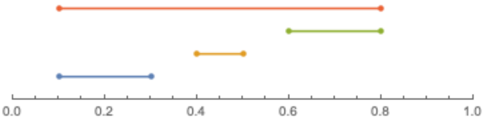


Current Approaches: Convex IP Pooling



Convex IP Pooling (Stewart and Quintana, 2018): the aggregate of $[a_1, b_1] \ldots [a_n, b_n]$ is the convex hull of their union, $conv\{\cup_i [a_i, b_i]\}$

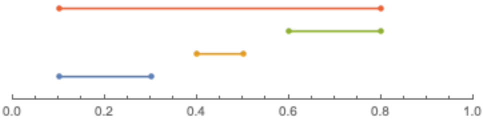
Current Approaches: Convex IP Pooling



Convex IP Pooling (Stewart and Quintana, 2018): the aggregate of $[a_1, b_1] \ldots [a_n, b_n]$ is the convex hull of their union, $conv\{\cup_i [a_i, b_i]\}$

Good:

Current Approaches: Convex IP Pooling

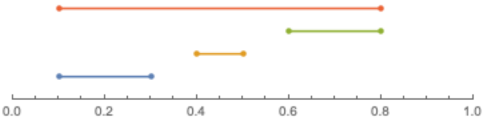


Convex IP Pooling (Stewart and Quintana, 2018): the aggregate of $[a_1, b_1] \ldots [a_n, b_n]$ is the convex hull of their union, $conv\{\cup_i [a_i, b_i]\}$

Good:

- Preserves independence judgments (in a sense)

Current Approaches: Convex IP Pooling

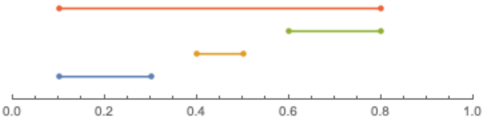


Convex IP Pooling (Stewart and Quintana, 2018): the aggregate of $[a_1, b_1] \ldots [a_n, b_n]$ is the convex hull of their union, $conv\{\cup_i [a_i, b_i]\}$

Good:

- Preserves independence judgments (in a sense)
- Commutes with Bayesian conditioning (Field conditioning, imaging)

Current Approaches: Convex IP Pooling



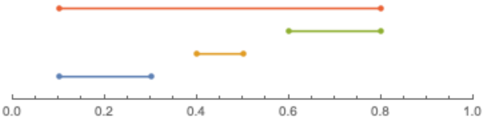
Convex IP Pooling (Stewart and Quintana, 2018): the aggregate of $[a_1, b_1] \ldots [a_n, b_n]$ is the convex hull of their union, $conv\{\cup_i [a_i, b_i]\}$

Good:

- Preserves independence judgments (in a sense)
- Commutes with Bayesian conditioning (Field conditioning, imaging)

Bad:

Current Approaches: Convex IP Pooling



Convex IP Pooling (Stewart and Quintana, 2018): the aggregate of $[a_1, b_1] \ldots [a_n, b_n]$ is the convex hull of their union, $conv\{\cup_i [a_i, b_i]\}$

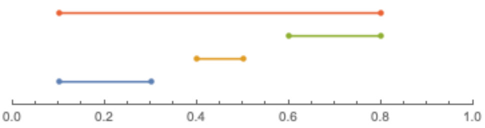
Good:

- Preserves independence judgments (in a sense)
- Commutes with Bayesian conditioning (Field conditioning, imaging)

Bad:

- Captures *consensus*, not *compromise*; washes out info distributed across community

Current Approaches: Convex IP Pooling



Convex IP Pooling (Stewart and Quintana, 2018): the aggregate of $[a_1, b_1] \ldots [a_n, b_n]$ is the convex hull of their union, $conv\{\cup_i [a_i, b_i]\}$

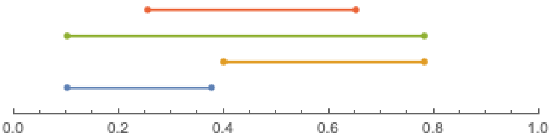
Good:

- Preserves independence judgments (in a sense)
- Commutes with Bayesian conditioning (Field conditioning, imaging)

Bad:

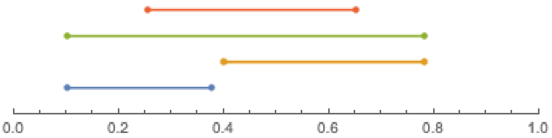
- Captures *consensus*, not *compromise*; washes out info distributed across community
- Ill-suited to inform future research, serve as an input to decision-theory.

Imprecise Credence



BIGGER PROBLEM: Convex IP pooling delivers *dominated* aggregates

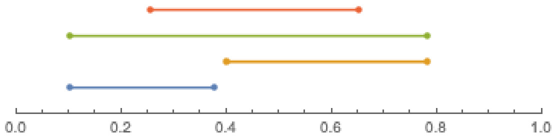
Imprecise Credence



BIGGER PROBLEM: Convex IP pooling delivers *dominated* aggregates

- My interval forecast for MGIS: $[0.1, 0.376923]$; yours: $[0.4, 0.784]$.

Imprecise Credence

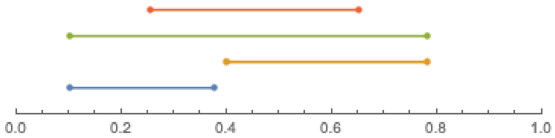


BIGGER PROBLEM: Convex IP pooling delivers *dominated* aggregates

- My interval forecast for MGIS: $[0.1, 0.376923]$; yours: $[0.4, 0.784]$.
- $\mathcal{I}_{0.7}$ uniquely renders our respective forecasts non-dominated

$$\mathcal{I}_{0.7}([a, b], w) = 0.7 \cdot \min_{x \in [a,b]} I(x, w) + 0.3 \cdot \min_{x \in [a,b]} I(x, w)$$

Imprecise Credence



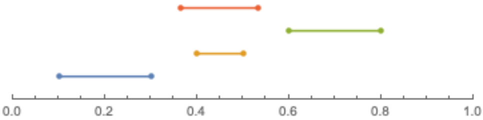
BIGGER PROBLEM: Convex IP pooling delivers *dominated* aggregates

- My interval forecast for MGIS: $[0.1, 0.376923]$; yours: $[0.4, 0.784]$.
- $\mathcal{I}_{0.7}$ uniquely renders our respective forecasts non-dominated

$$\mathcal{I}_{0.7}([a,b],w) = 0.7 \cdot \min_{x \in [a,b]} I(x,w) + 0.3 \cdot \min_{x \in [a,b]} I(x,w)$$

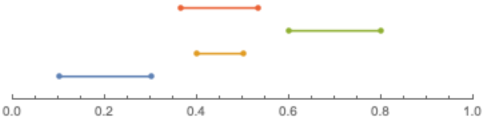
- But the convex IP pool, $[0.1, 0.784]$, is *accuracy dominated* by $[0.257, 0.652]$ according to $\mathcal{I}_{0.7}$

Current Approaches



Linear pooling of lower probabilities: the aggregate of $[x_1, y_1] \ldots [x_n, y_n]$ is $[a, b]$ with $a = \sum_i \lambda_i x_i$ and $b = \sum_i \lambda_i y_i$, where $\sum_i \lambda_i = 1$.

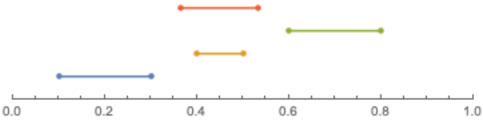
Current Approaches



Linear pooling of lower probabilities: the aggregate of $[x_1, y_1] \ldots [x_n, y_n]$ is $[a, b]$ with $a = \sum_i \lambda_i x_i$ and $b = \sum_i \lambda_i y_i$, where $\sum_i \lambda_i = 1$.

Linear pooling of Bayes risk functions (Nau, 2002): the aggregate of $[x_1, y_1] \ldots [x_n, y_n]$ is an IP distribution $\mathcal{C}$ such that betting against $\mathcal{C}$ and betting against $[x_1, y_1] \ldots [x_n, y_n]$ one by one is indistinguishable.

Current Approaches

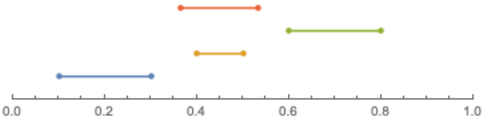


Linear pooling of lower probabilities: the aggregate of $[x_1, y_1] \ldots [x_n, y_n]$ is $[a, b]$ with $a = \sum_i \lambda_i x_i$ and $b = \sum_i \lambda_i y_i$, where $\sum_i \lambda_i = 1$.

Linear pooling of Bayes risk functions (Nau, 2002): the aggregate of $[x_1, y_1] \ldots [x_n, y_n]$ is an IP distribution $\mathcal{C}$ such that betting against $\mathcal{C}$ and betting against $[x_1, y_1] \ldots [x_n, y_n]$ one by one is indistinguishable.

- **Good**: satisfy some *prima facie* desirable axioms

Current Approaches



Linear pooling of lower probabilities: the aggregate of $[x_1, y_1] \ldots [x_n, y_n]$ is $[a, b]$ with $a = \sum_i \lambda_i x_i$ and $b = \sum_i \lambda_i y_i$, where $\sum_i \lambda_i = 1$.

Linear pooling of Bayes risk functions (Nau, 2002): the aggregate of $[x_1, y_1] \ldots [x_n, y_n]$ is an IP distribution $\mathcal{C}$ such that betting against $\mathcal{C}$ and betting against $[x_1, y_1] \ldots [x_n, y_n]$ one by one is indistinguishable.

- **Good**: satisfy some *prima facie* desirable axioms
- **Bad**: deliver *dominated* aggregates

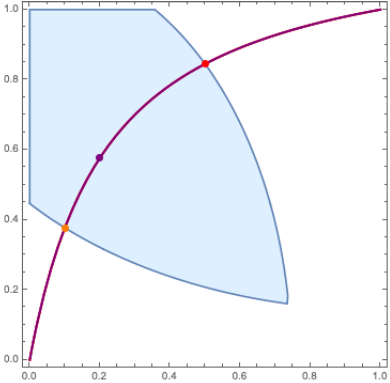
Epistemic Utility Based Aggregation

EU Aggregation (Unique EU Function): If n individuals have interval forecasts $[a_1, b_1], \ldots, [a_n, b_n]$ for E, and they all are rendered non-dominated by $\mathcal{I}_\alpha$, then any reasonable aggregate must take the following form:

$$\left[x, \frac{\alpha^2 x}{1 - 2\alpha + \alpha^2 - x + 2\alpha x}\right]$$

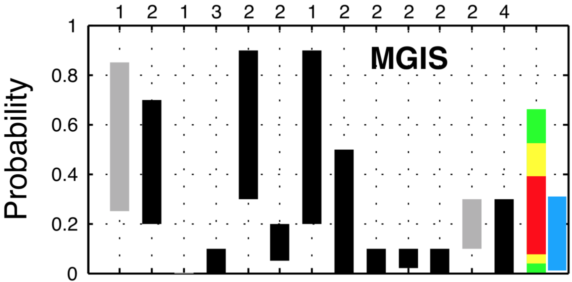
where $\min_i a_i \leq x \leq \max_i a_i$.

Epistemic Utility Based Aggregation



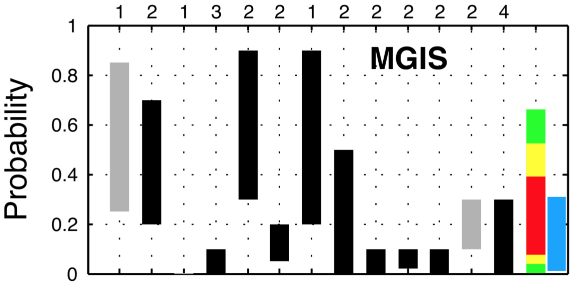
Permissible aggregates of $[0.1, 0.376923]$, $[0.2, 0.576471]$ and $[0.5, 0.844828]$

Advantages



Epistemic utility based aggregate using uniform expert weights: $[0.0153442, 0.318378]$

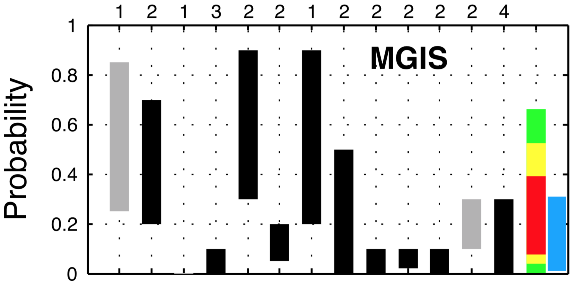
Advantages



Epistemic utility based aggregate using uniform expert weights: $[0.0153442, 0.318378]$

Non-dominated aggregates;

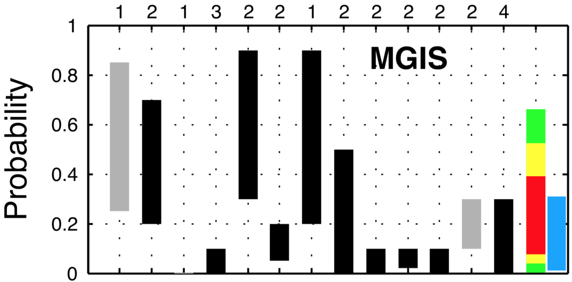
Advantages



Epistemic utility based aggregate using uniform expert weights: $[0.0153442, 0.318378]$

Non-dominated aggregates; Compromise, not consensus;

Advantages



Epistemic utility based aggregate using uniform expert weights: $[0.0153442, 0.318378]$

Non-dominated aggregates; Compromise, not consensus; Robust against outliers.

Joyce, J. (1998). A nonpragmatic vindication of probabilism. Philosophy of Science 65(4), 575–603.

Joyce, J. M. (2009). Accuracy and coherence: Prospects for an alethic epistemology of partial belief. In F. Huber and C. Schmidt-Petri (Eds.), Degrees of Belief, Volume 342. Dordrecht: Springer.

Konek, J. (2019). Epistemic conservativity and imprecise credence. Philosophy and Phenomenological Research.

Kriegler, E. (2009). Imprecise probability assessment of tipping points in the climate system. Proceedings of the National Academy of Sciences 106(13), 5041–5046.

Mayo-Wilson, C. and G. Wheeler (2015). Accuracy and imprecision: A mildly immodest proposal. Philosophy and Phenomenological Research.

Moss, S. (2011). Scoring rules and epistemic compromise. Mind 120(480), 1053–1069.

Nau, R. (2002). The aggregation of imprecise probabilities. Journal of Statistical Planning and Inference 105(1), 265–282.

Schoenfield, M. (2015). The accuracy and rationality of imprecise credences. Nous.

Seidenfeld, T., M. J. Schervish, and J. B. Kadane (2012). Forecasting with imprecise probabilities. International Journal of Approximate Reasoning 53, 1248–1261.

Stewart, R. and I. Quintana (2018). Probabilistic opinion pooling with imprecise probabilities. J Philos Logic 47, 17–45.