Supra-Bayesian Approach to Merging of Incomplete and Incompatible Data

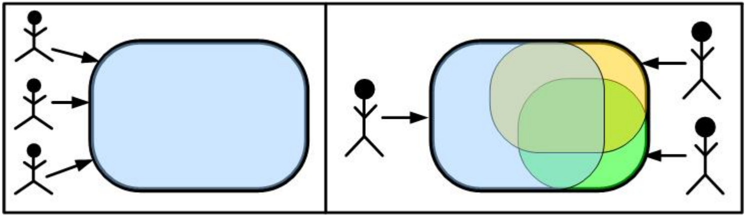
Theoretical and practical results

Vladimíra Sečkárová

Charles University in Prague
Institute of Information Theory and Automation

28.3.2011

Situation

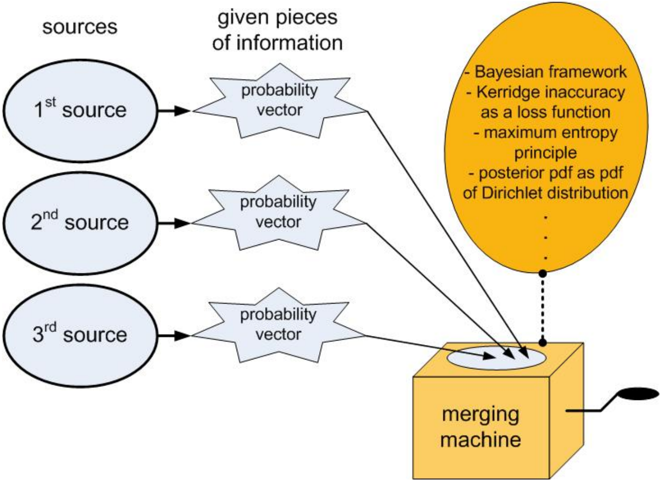


Assumptions

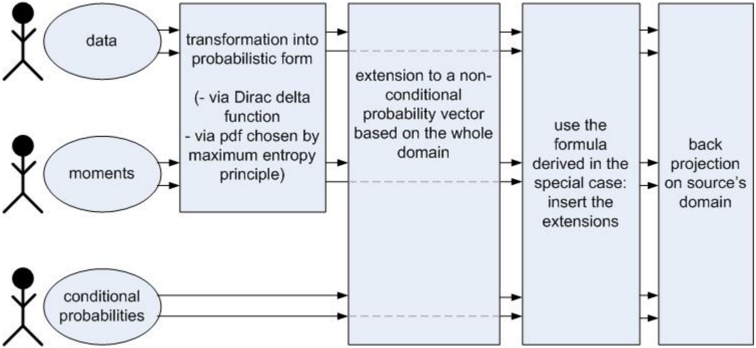
- finite number of sources
- given knowledge: probabilities, data, ...

Task: find the optimal merger of given information

Same domains, probability information



Different domains (but neighbors), different forms of given information



Final merger

$$
{}^{O}h(.) = \frac{1}{n + \sum_1^s \lambda_j(\mathrm{data})} + \frac{\sum_1^s \lambda_j(\mathrm{data}) g_j(.)}{n + \sum_1^s \lambda_j(\mathrm{data})} \qquad (1)
$$

- n - no. of realizations of a random vector described by the sources ($< \infty$)
- s - no. of sources ($< \infty$)
- λ_j Lagrange multipliers – expresses how important the information given by j^{th} source is (based on constraints – distance between given distribution and unknown distribution)

Bayes rule

- $\mathbf{X} = (X_1, \dots, X_s)$ observations - random variables
- $f(\mathbf{X}|\theta)$ - model
- assumption: X_j (conditionally) independent identically distributed, $j = 1, \dots, s$:

$$f(\mathbf{X}|\theta) = \prod_{1}^{s} f(X_j|\theta)$$

- $q(\theta)$ - a prior pdf
- $\pi(\theta|\mathbf{X})$ - a posterior pdf

$$\pi(\theta|\mathbf{X}) \propto q(\theta) f(\mathbf{X}|\theta) = q(\theta) \prod_{1}^{s} f(X_j|\theta)$$

Supra-Bayes: add an unknown parameter

- s sources, $(\mathbf{X}, \theta)$ - random vector
- assumption: θ has finite no. of realizations
- denote n^* no. of realizations of $(\mathbf{X}, \theta)$
- in this case, the Lagrange multipliers λ_j, $j = 1, \ldots, s$ do not depend on θ, because:

$$\text{Lagrangian} = \text{Entropy}(\pi(h|D)) - \int \sum_{1}^{n^*} [\ldots] + \ldots$$

- final merger:

$$^{O}h(\mathbf{X}, \theta) = \frac{1}{n^* + \sum_1^s \lambda_j} + \frac{\sum_1^s \lambda_j g_j(\mathbf{X}, \theta)}{n^* + \sum_1^s \lambda_j} \qquad (2)$$

Situation – 2 sources

- sources vs. observations: 2 conditionally independent sources
- form of given information we need from sources:
  - first step: $X_1 \quad \rightarrow \quad \delta(X_1 - x_1)$
  - second step: model + prior pdf are needed: $\delta(X_1 - x_1) f(X_1, X_2 | \Theta) q(\Theta)$
  - analogically for 2^{nd} source
- but summation over all realizations is not equal to 1 $\rightarrow$ normalization
- can we suppose $\delta(X_j - x_j) q(\theta)$ was given? under which assumptions?
- there are many possible situations, remember, sources have to be neighbors (mutual or fix one source and create its neighbors)

2 sources: mutual neighbors, same domain

$$^{O}h(X_1, X_2, \Theta) = \frac{1 + \sum_1^2 \lambda_j \delta(X_j - x_j) f(X_1, X_2|\Theta) q(\Theta)}{n^* + \lambda_1 + \lambda_2} = \bullet$$

A) if there appears a realization of X_j such that $\delta(X_j - x_j) = 1$, then evaluate:

$$\frac{(n^* + \sum_1^2 \lambda_j)^{O}h(X_1, X_2, \Theta) - 1}{\sum_1^2 \lambda_j \delta(X_j - x_j)} \approx \pi(\Theta|X_1, X_2)$$

B) if $\delta(X_j - x_j) = 0 \ \forall j$ then:

$$\bullet = f(X_1, X_2|\Theta) q(\Theta) \left(\frac{\frac{1}{f(X_1, X_2|\Theta) q(\Theta)} + 0}{n^* + \lambda_1 + \lambda_2} \right)$$

since situation B) will certainly occur, we need to bring more assumptions:

- $f(X_1, X_2|\Theta) > 0 \quad \rightarrow f(X_1|\Theta) > 0,\ f(X_2|\Theta) > 0$
- $q(\Theta) > 0$

then we get

$$= f(X_1, X_2|\Theta)q(\Theta)K(\Theta)$$

we want $K(.)$ to be independent from Θ, so we will try:

- $q(\Theta) \sim Uni(.)$
- $f(X_1, X_2|\Theta)$ properly flat

$+$ earlier assumptions:

- Θ has finite no. of realizations

Other possibilities of given information

- 1^{st} source: $\delta(X_1 - x_1) f(X_1, X_2|\Theta) q(\Theta)$,
  2^{nd} source: $\delta(X_2 - x_2) q(\Theta)$
- second one is the neighbor of the first one
- extension of information for 2^{nd} source:
  ${}^{O}h(X_1, X_2|\Theta)\delta(X_2 - x_2)q(\Theta)$

$$
{}^{O}h(X_1, X_2, \Theta) = \frac{\ldots + \lambda_2 \delta(X_2 - x_2) q(\Theta)\, {}^{O}h(X_1, X_2|\Theta) \frac{f(X_1, X_2|\Theta)}{f(X_1, X_2|\Theta)}}{n^* + \lambda_1 + \lambda_2}
$$

$$
= f(X_1, X_2|\Theta) q(\Theta) \left(\frac{\frac{1}{f(X_1, X_2|\Theta) q(\Theta)} + \lambda_1 \delta(X_1 - x_1) + \lambda_2 \delta(X_2 - x_2) c}{n^* + \lambda_1 + \lambda_2} \right)
$$

- the fraction $\frac{{}^{O}h(\Theta)}{q(\Theta)} = c^{*}$ does not depend on θ
- we also can find out that $\frac{{}^{O}h(X_1,X_2|\theta)}{f(X_1,X_2|\Theta)} = \frac{\frac{{}^{O}h(X_1,X_2,\Theta)}{{}^{O}h(\Theta)}}{f(X_1,X_2|\Theta)} = c$
- and we get the same situation as before (2 mutual neighbors)
- if no. of sources > 2 :
  - we need at least one source giving model + prior pdf,
  - others will be its/their neighbors: giving model or prior pdf or just $\delta(X_j - x_j)$

Choice of the distance between given and unknown distribution, boundary on the distance

- we needed to find the optimal posterior pdf because the optimal merger is $\mathrm{E}_{\pi(h|D)}(h|D)$, h is unknown distribution
- we used maximum entropy principle with constraints on distances between given and unknown distribution
- we considered:
  - expected Kerridge inaccuracy:
    $\beta \geq \mathrm{EK}(g_j, h) \geq \mathrm{Entropy}(g_j) \geq 0$
  - expected Kullback-Leibler divergence:
    $\beta \geq \mathrm{ED}_{KL}(g_j, h) \geq \mathrm{D}_{KL}(g_j, ^{O} h) \geq 0$
  - "reversed" expected Kerridge inaccuracy:
    $\beta \geq \mathrm{EK}(h, g_j)$

Example - a dice (first type Kerridge inaccuracy is used)

By sampling from $\{1, \ldots, 6\}$ (with particular probabilities) I got following results:

- for a fair dice:
  - works well for $\beta = 1.71$ – entropy of a fair dice
- one side is preferred $h = (7, 1, 1, 1, 1, 1)/12$:
  - min: 0.54 – ok/ok (results for small/large no. of sources)
  - max: 2.46 – ok/k.o.
  - mean: 2.16 – k.o./k.o.
- two sides are preferred $h = (4, 4, 1, 1, 1, 1)/12$:
  - min: 1.1 – k.o./k.o.
  - max: 2.48 – ok/ok
  - mean: 2.02 – ok/ok
- three sides are preferred $h = (3, 3, 3, 1, 1, 1)/12$:
  - min: 1.39 – k.o./k.o.
  - max: 2.49 – k.o./ok
  - mean: 1.94 – ok/ok

Thanks for the attention.