Robust Estimators under the Imprecise Dirichlet Model

Marcus Hutter

Istituto Dalle Molle di Studi sull'Intelligenza Artificiale
IDSIA, Galleria 2, CH-6928 Manno-Lugano, Switzerland
marcus@idsia.ch, http://www.idsia.ch/~marcus

2003

Table of Contents

Abstract

Walley's Imprecise Dirichlet Model (IDM) for categorical data overcomes several fundamental problems which other approaches to uncertainty suffer from. Yet, to be useful in practice, one needs efficient ways for computing the imprecise=robust sets or intervals. The main objective of this work is to derive exact, conservative, and approximate, robust and credible interval estimates under the IDM for a large class of statistical estimators, including the entropy and mutual information.

The Dirichlet Model

- Discrete random variables: $\imath \in \Omega := \{1,...,d\}$
- i.i.d. random process: outcome $i \in \{1,...,d\}$ with probability π_i.
- Likelihood of data $\boldsymbol{D}$ with n_i observations i and sample size $n = n_+$ $(x_+ := \sum_i x_i)$ is $p(\boldsymbol{D}|\boldsymbol{\pi}) = \prod_i \pi_i^{n_i}$.
- Initial uncertainty in $\boldsymbol{\pi}$ is modeled by a (second order) "belief" Dirichlet prior $p(\boldsymbol{\pi}) \propto \prod_i \pi_i^{n'_i - 1}$.

The Dirichlet Model (ctd.)

- Notation: Write $n'_i = s \cdot t_i$ with $s := n'_+$, hence $\boldsymbol{t} \in \Delta := \{\boldsymbol{t} \in \mathbb{R}^d : t_i \geq 0 \forall i,\ t_+ = 1\}$
- Examples of uninformed priors: $t_i = \frac{1}{d} \forall i$: Haldane ($s = 0$), Perks ($s = 1$), Jeffreys ($s = \frac{d}{2}$), Bayes/Laplace/uniform ($s = d$).
- Posterior: $p(\boldsymbol{\pi}|\boldsymbol{D}) = p(\boldsymbol{\pi}|\boldsymbol{n}) \propto \prod_i \pi_i^{n_i + s t_i - 1}$.
- Expected value: $E_{\boldsymbol{t}}[\mathcal{F}] = \int_\Delta \mathcal{F}(\boldsymbol{\pi}) p(\boldsymbol{\pi}|\boldsymbol{n}) d\boldsymbol{\pi}$
- Variance: $\mathrm{Var}_{\boldsymbol{t}}[\mathcal{F}] = E_{\boldsymbol{t}}[\mathcal{F}^2] - E_{\boldsymbol{t}}[\mathcal{F}]^2$.

The Imprecise Dirichlet Model

- Model our ignorance by considering sets of priors $p(\boldsymbol{\pi})$, often called Imprecise probabilities.
- The Imprecise Dirichlet Model (IDM) [Walley:96] considers the set of all $\boldsymbol{t} \in \Delta$, i.e. $\{p_{\boldsymbol{t}}(\boldsymbol{\pi}) : \boldsymbol{t} \in \Delta\}$.
- IDM satisfies symmetry principle and is reparametrization invariant (RIP).
- Set of priors $\Rightarrow$ set of posteriors $\Rightarrow$ set of expected vals.
- For real-valued quantities like $E_{\boldsymbol{t}}[\mathcal{F}]$ the sets are typically intervals (called robust):
  $E_{\boldsymbol{t}}[\mathcal{F}] \in [\min_{\boldsymbol{t}\in\Delta} E_{\boldsymbol{t}}[\mathcal{F}]\,, \max_{\boldsymbol{t}\in\Delta} E_{\boldsymbol{t}}[\mathcal{F}]]$

Problem Setup and Notation

$F(\boldsymbol{u}) := E_{\boldsymbol{t}}[\mathcal{F}]$ with identification $u_i^{\cdots} = \frac{n_i + s t_i^{\cdots}}{n+s}$.

Goal: Derive expressions for upper and lower F values

$$\overline{F} := \max_{\boldsymbol{u} \in \Delta'} F(\boldsymbol{u}) \quad \text{and} \quad \underline{F} := \min_{\boldsymbol{u} \in \Delta'} F(\boldsymbol{u}), \qquad \overline{\underline{F}} := [\underline{F}, \overline{F}]$$

$$\Delta' = \{\boldsymbol{u} : u_i \geq u_i^0,\ u_+ = 1\} \quad \text{with} \quad u_i^0 := \frac{n_i}{n+s}.$$

Example: $F(\boldsymbol{u}) = E_{\boldsymbol{t}}[\pi_i] = \frac{n_i + s t_i}{n+s} = u_i \Rightarrow \overline{\underline{F}} = [\frac{n_i}{n+s}, \frac{n_i+s}{n+s}]$

Exact Robust Intervals for Concave F

- Assume $F : \Delta' \to \mathbb{R}$ concave and $F(\boldsymbol{u}) = \sum_{i=1}^{d} f(u_i)$:

- F attains the the global minimum $\underline{F}$ at corner $\boldsymbol{u}^{\underline{F}}$ with $t_i^{\underline{F}} = \delta_{ii^{\underline{F}}}$ and $i^{\underline{F}} := \arg\max_i n_i$.

- F attains the global maximum $\overline{F}$ at water-filling point $\boldsymbol{u}^{\overline{F}}$ with $u_i^{\overline{F}} = \max\{u_i^0, \tilde{u}\}$, where $\tilde{u} = \min_{m \in \{1...d\}} \frac{s + \sum_{k \leq m} n_{i_k}}{m(n+s)}$, where $n_{i_1} \leq n_{i_2} \leq ... \leq n_{i_d}$.

Approximate Robust Intervals

Exact expansion of $F(\boldsymbol{u}) = \sum_i f(u_i)$ around $\boldsymbol{u}^0$.

Assume $F : \Delta' \to \mathbb{R}$ Lipschitz diff. and $\sigma := \frac{s}{n+s}$ small.

$\Rightarrow \overline{F} - \underline{F} = O(\sigma) \Rightarrow$ approximation to $\underline{F}$ should be $O(\sigma^2)$.

Notation: $F \sqsubseteq G \quad :\Leftrightarrow \quad F \leq G$ and $F = G + O(\sigma^2)$

$$F_0 + F_R^{lb} \sqsubseteq \underline{F} \leq F(\boldsymbol{u}) \leq \overline{F} \sqsubseteq F_0 + F_R^{ub}$$

$$F_0 = F(\boldsymbol{u}^0), \quad F_R^{ub} = \sigma \max_i f'(u_i^0) = \sigma f'(\tfrac{\min_i n_i}{n+s}),$$

$$F_R^{lb} = \sigma \min_i f'(u_i^0 + \sigma) = \sigma f'(\tfrac{\max_i n_i + s}{n+s}),$$

Application: Expected Entropy H

$\mathcal{H}(\boldsymbol{\pi}) := -\sum_i \pi_i \log \pi_i \Rightarrow$

$H(\boldsymbol{u}) := E_{\boldsymbol{t}}[\mathcal{H}] = \sum_i h(u_i)$ with

$h(u_i) = u_i \cdot \sum_{k=(n+s)u_i+1}^{n+s} k^{-1}$ (for integer s and $(n+s)u_i$)

General expression in terms of DiGamma function ψ.

Example (exact): For $d=2$, $n_1=3$, $n_2=6$, $s=1$ we have $\underline{H} = [0.5639..., 0.6256...]$, so $\overline{H} - \underline{H} = O(\frac{1}{10})$.

Example (approximate): $\sigma = \frac{1}{10}$,
$[H_0 + H_R^{lb}, H_0 + H_R^{ub}] = [0.5564..., 0.6404...]$, hence
$H_0 + H_R^{ub} - \overline{H} = 0.0148 = O(\frac{1}{10^2})$,
$\underline{H} - H_0 - H_R^{lb} = 0.0074... = O(\frac{1}{10^2})$.

Error Propagation

- $F := G + H$. Naive: $\overline{F} \leq \overline{G} + \overline{H}$, but $\overline{F} \not\sqsubseteq \overline{G} + \overline{H}$.
- Results: $O(\sigma^2)$ bounds ($\sqsubseteq$) for $F = G \star H$ and $\star \in \{+, -, \times, /, ...\}$.
- Every function F (w.b.c.) can be written as a sum of a concave function G and a convex function H.
- For convex and concave functions, determining bounds is particularly easy (special case on previous slides).
- Often F decomposes naturally into convex and concave parts as is the case for the mutual information: $\mathcal{I}(\boldsymbol{\pi}) \;=\; \mathcal{H}(\boldsymbol{\pi}_{i+}) + \mathcal{H}(\boldsymbol{\pi}_{+j}) - \mathcal{H}(\boldsymbol{\pi}_{ij})$

IDM for Product Spaces

- Product spaces: $\Omega = \Omega_1 \times \Omega_2 = \{1,...,d_1\} \times \{1,...,d_2\}$
- Applications: mutual inform., robust trees, Bayes nets.
- Full IDM invariant under general (non-column/row cross) groupings of elements of Ω:
  $\boldsymbol{t} \in \Delta := \{\boldsymbol{t} \in I\!R^{d_1 \times d_2} \,:\, t_{ij} \geq 0 \,\forall ij,\; t_{++} = 1\}$
- Smaller IDM, invariant only under groupings of whole columns and/or rows of Ω, makes more sense:
  $\boldsymbol{t} \in \Delta_{d_1} \otimes \Delta_{d_2} \;\subsetneq\; \Delta.$
- Result: Smaller IDM leads to $O(\sigma^2)$ smaller (=better) robust sets.

Exact Robust Credible Sets

For a probability distribution $p : \mathbb{R}^d \to [0,1]$,
a/the α-credible set is $A^{min} := \underset{A:p(A)\geq\alpha}{\arg\min}\ \mathrm{Vol}(A)$

For a set of probability distributions $\{p_t(x)\}$, a robust α-credible set is a set A which contains x with p_t-probability at least α for all t. A minimal size robust α-credible set is

$$A^{min} := \underset{A=\bigcup_t A_t : p_t(A_t)\geq\alpha \forall t\in T}{\arg\min} \mathrm{Vol}(A) \quad \neq \bigcup_t A_t^{min}$$

It is not easy to deal with the first expression, but $\bigcup_t A_t$ can be used as a conservative estimate.

Approximate Robust Credible Intervals

Shortest α-credible intervals w.r.t. a univariate $p_t(x)$:

$$\widetilde{\underset{\sim}{x}}_t := \underset{[a,b]:p_t([a,b])\geq\alpha}{\arg\min} (b-a),$$

$$\overline{\widetilde{x}} \leq \max_t \widetilde{x}_t \leq \max_t E_t[x] + \max_t[\widetilde{x}_t - E_t[x]] =$$

$$= \overline{E[x]} + \overline{\Delta\widetilde{x}} = \overline{E[x]} + \kappa\sigma_{t^*} + O(n^{-3/2}).$$

$\alpha = \mathrm{erf}(\kappa/\sqrt{2})$ and $\sigma^2_{t^*} = \mathrm{Var}_{t^*}[x]$ for some $t^* \in \Delta$, e.g. $x \in \{\mathcal{F}, \mathcal{H}, \mathcal{I}\}$ and $\mathrm{Var}_{\boldsymbol{t}^*}[\mathcal{I}]$ computed in [Hutter:02].

Non-Gaussian distributions depending on some sample size n are usually close to Gaussian for large n due to the central limit theorem.

Conclusions

- IDM has not only interesting theoretical properties, but explicit (exact/conservative/approximate) expressions for robust (credible) intervals for various quantities can and have been derived.
- The computational complexity of the derived bounds on $F = \sum_i f_i$ is very small, typically one or two evaluations of F or related functions, like its derivative.
- First applications of these results, especially the mutual information, to robust inference of trees look promising [Zaffalon&Hutter:03].