Minimum Dispair & Maximum Despair: Minimum Disparity Statistics

focustat

FOCUS DRIVEN STATISTICAL
INFERENCE WITH COMPLEX DATA

Nils Lid Hjort

Department of Mathematics, University of Oslo

Building Bridges at Bislett

May 2017, Teknologihuset

Minimum disparity, i.i.d. case

Suppose $y_1, \ldots, y_n$ i.i.d. from g, and we use model $f_\theta(\cdot) = f(\cdot, \theta)$. With $B(\cdot)$ convex and smooth on the half-line, and with $B(1) = 0$,

$$d(g, f_\theta) = \int B\Big(\frac{g}{f_\theta}\Big) f_\theta \,\mathrm{d}y$$

is a divergence (Johan Jensen inequality, 1906). Note that $\dim \geq 2$ is ok.

Hence it's a good estimation idea to use

$$\widehat\theta = \operatorname{argmin} d(\widetilde g, f_\theta),$$

where $\widetilde g$ is a 'wider' estimate of g. Can use

- $\widetilde g$ nonparametric, like the kernel density estimator [99% of all MinDisp papers use this], or e.g. Hjort–Glad $f(y, \widetilde\theta)\widetilde r(y)$;
- $\widetilde g$ via Bayesian nonparametrics (e.g. along with prior for θ);
- $\widetilde g$ via $g = (1-\varepsilon) f_\theta + \varepsilon H$ modelling;
- $\widetilde g(y) = g(y, \widetilde\theta, \widetilde\gamma)$ from a bigger parametric family.

Minimising $\int B(g/f_\theta) f_\theta \, \mathrm{d}y$ corresponds to solving

$$\int A\Big(\frac{g}{f_\theta}\Big) f_\theta u_\theta \, \mathrm{d}y = 0,$$

where $u_\theta = \partial \log f_\theta / \partial\theta$ is score function and

$$A(\rho) = \rho B'(\rho) - B(\rho)$$

is the Ratio Adjustment Function (RAF). So, $\widehat{\theta}$ solves

$$H_n(\theta) = \int A\Big(\frac{\widetilde{g}}{f_\theta}\Big) f_\theta u_\theta \, \mathrm{d}y = 0.$$

We have $A'(\rho) = B''(\rho)$, so $A(\rho)$ is increasing.

Properties of $A(\rho)$ drive robustness properties of $\widehat{\theta}$.

A few Min Dispair schemes

1. $B(\rho) = \rho \log \rho$: then $d = \mathrm{KL}(g, f_\theta)$, and $A(\rho) = \rho$. Estimator: solution to $\int \widetilde{g}(y) u(y, \theta)\, \mathrm{d}y = 0$ ... famous divergence, efficient at model, close to ML (inside and outside model), and not robust.

2. $B(\rho) = -\log \rho$: then $d = \mathrm{KL}(f_\theta, g)$, the converse of what we learn in school, and $A(\rho) = \log \rho - 1$. Estimator: solution to

$$H_n(\theta) = \int \log \frac{\widetilde{g}(y)}{f(y, \theta)} f(y, \theta) u(y, \theta)\, \mathrm{d}y = 0.$$

3. $B(\rho) = 1 - \sqrt{\rho}$: then $d = 1 - \int \sqrt{g f_\theta}\, \mathrm{d}y$, the Hellinger distance; $A(\rho) = \frac{1}{2}\sqrt{\rho} - 1$, and the estimator solves

$$H_n(\theta) = \int (\widetilde{g} f_\theta)^{1/2} u_\theta\, \mathrm{d}y = 0.$$

4. $B(\rho) = 1 - \rho^{1/p}$: then $d = 1 - \int g^{1/p} f_\theta^{1/q}\, \mathrm{d}y$, with $1/p + 1/q = 1$. Estimator maximises $\int \widetilde{g}^{1/p} f_\theta^{1/q}\, \mathrm{d}y$.

Issues to pursue

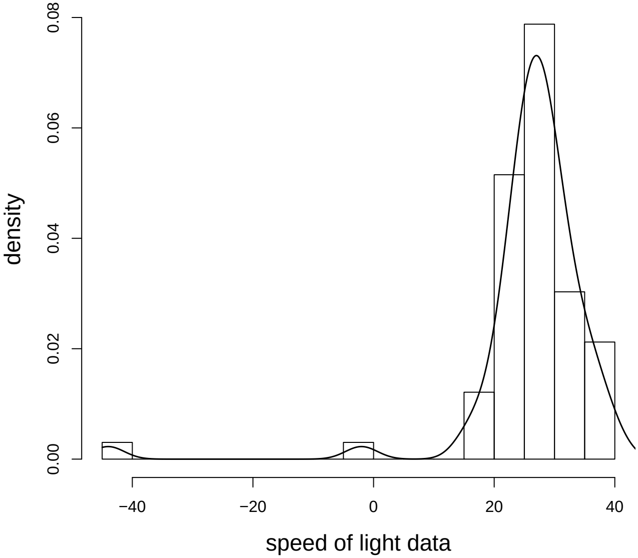
So, $\widehat\theta$ is the minimiser of $d(\widetilde g, f_\theta) = \int B(\widetilde g/f_\theta) f_\theta \,\mathrm{d}y$.

- 1 $\widetilde g$ nonparametric: properties $\widehat\theta$?
- 2 And what is really required of $\widetilde g$? Fine-tuning?
- 3 $\widehat\theta$ is often efficient at the model, with $\sqrt{n}(\widehat\theta - \widehat\theta_{\rm ML}) \to_{\rm pr} 0$; how different is it (then) from ML?
- 4 $\widetilde g$ bigger-parametric: properties $\widehat\theta$?
- 5 Which $B(\rho)$ make the procedures robust?
- 6 Choice of $B(\rho)$ function (in a class of good ones)?
- 7 Generalising from i.i.d. to regression?
- 8 Goodness of fit, based on

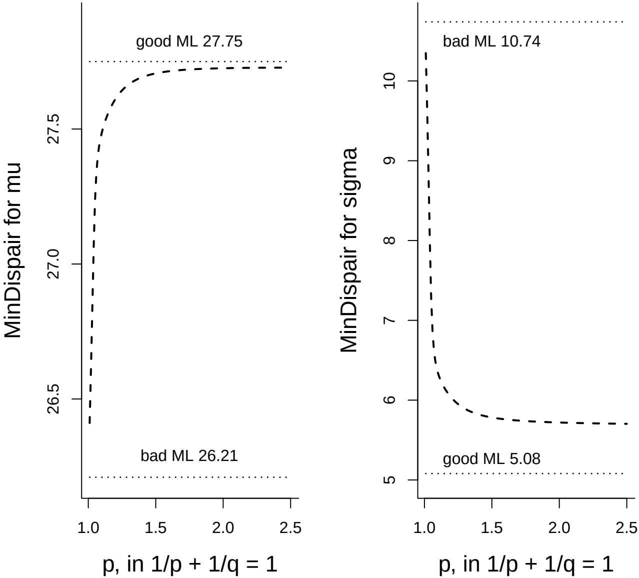
$$D_n = \min d(\widetilde g, f_\theta) = d(\widetilde g, f(\cdot, \widehat\theta))?$$

- 9 Model selection using Min Dispair? Is there a MDIC, Minimum Divergence Information Criterion?

Example: Speed of light data, $n = 66$, two outliers to the left.



Minimising $1 - \int \widetilde{g}^{1/p} f_\theta^{1/q}\,\mathrm{d}y$, for different p, where $1/p + 1/q = 1$:



Minimising d(nonparametric, parametric)

Starting with a $B(\rho)$, use $A(\rho) = \rho B'(\rho) - B(\rho)$, to define the least false θ_0 as solution to

$$\int A\Big(\frac{g}{f_\theta}\Big) f_\theta u_\theta \,\mathrm{d}y = 0$$

and the MinDisp estimator $\widehat\theta$ as solution to

$$U_n(\theta) = \int A\Big(\frac{\widetilde g}{f_\theta}\Big) f_\theta u_\theta \,\mathrm{d}y = 0.$$

Under some conditions,

$$\sqrt{n}U_n(\theta_0) = \sqrt{n}\int \Big\{A\Big(\frac{\widetilde g}{f_{\theta_0}}\Big) - A\Big(\frac{g}{f_{\theta_0}}\Big)\Big\} f_{\theta_0} u_{\theta_0} \,\mathrm{d}y \to_d U \sim \mathrm{N}_p(0, K).$$

Along with 2nd order derivative matrix to a J, and other technical details, we're led to

$$\sqrt{n}(\widehat\theta - \theta_0) \to_d J^{-1}U \sim \mathrm{N}_p(0, J^{-1}KJ^{-1}).$$

The conditions for convergence of

$$\sqrt{n}U_n(\theta_0) = \sqrt{n}\int\Big\{A\Big(\frac{\widetilde g}{f_{\theta_0}}\Big) - A\Big(\frac{g}{f_{\theta_0}}\Big)\Big\}f_{\theta_0}u_{\theta_0}\,\mathrm{d}y$$

basically amount to

$$\sqrt{n}\int A'\Big(\frac{g}{f_{\theta_0}}\Big)(\widetilde g - g)u_{\theta_0}\,\mathrm{d}y \to_d U \qquad [\text{think } \sqrt{n}h^2 \to 0]$$

and

$$\sqrt{n}\int A''\Big(\frac{g}{f_{\theta_0}}\Big)\frac{(\widetilde g - g)^2}{f_{\theta_0}}u_{\theta_0}\,\mathrm{d}y \to_{\mathrm{pr}} 0 \qquad [\text{think } \sqrt{n}h \to \infty].$$

1. The literature is far too dominated by ‘always’ using the kernel density estimator (granted, even there it’s ‘technical’).

2. At the model, with $g = f_{\theta_0}$, matters simply, both J and K are proportional to J_{fish}, and sandwich becomes J_{fish}^{-1}: the MinDisp is as efficient as ML.

3. Sam-Erik needs within-reach generalisations of these matters to $O(1/\sqrt{n})$ neighbourhoods around given models.

Minimum Dispair with Bigger-Parametric Start

Suppose $y_1, \ldots, y_n$ are i.i.d. from some g. We fit data to $g(y, \theta, \gamma)$, which contains $f(y, \theta)$ as a special case:

$$f(y, \theta) = g(y, \theta, \gamma_0) \quad \text{for a known } \gamma_0.$$

With $(\widetilde{\theta}, \widetilde{\gamma})$ the ML in bigger model, the minimum dispair estimator for θ is $\widehat{\theta}$, the minimiser of

$$d(\widetilde{g}, f_\theta) = \int B\Big(\frac{g(y, \widetilde{\theta}, \widetilde{\gamma})}{f(y, \theta)}\Big) f(y, \theta)\, \mathrm{d}y.$$

We have $\widehat{\theta} \to_{\mathrm{pr}} \theta_0$, the least false value minimising

$$\int B\Big(\frac{g(y, \theta_1, \gamma_1)}{f(y, \theta)}\Big) f(y, \theta)\, \mathrm{d}y,$$

where (θ_1, γ_1) are KL minimisers in the bigger model. If small model is correct, $\widehat{\theta}$ is consistent for the right θ_0.

Assume smaller model correct, $g(y, \theta_0, \gamma_0) = f(y, \theta_0)$.

Theorem A: MinDisp is efficient at the model:

$$\sqrt{n}(\widehat{\theta} - \theta_0) \to_d \mathrm{N}_p(0, J(\theta_0)^{-1}).$$

Theorem B: Can use

$$D_n = \min_\theta d(\widetilde{g}, f(\cdot, \theta)) = \int B\Big(\frac{g(y, \widetilde{\theta}, \widetilde{\gamma})}{f(y, \widehat{\theta})}\Big) f(y, \widehat{\theta})\,\mathrm{d}y$$

as a goodness-of-fit test statistic:

$$nD_n \to_d \tfrac{1}{2} B''(1)\chi^2_q$$

under model, where $q = \dim(\gamma)$.

Theorem C: Can work out $\sqrt{n}(\widehat{\theta} - \theta_1)$ outside model conditions.

Remarks & questions re subset of Sam-Erik's PhD

A. Why isn't everyone using Minimum Dispair? Why don't we teach minimum disparity estimators in our courses?

We can choose B and $\widetilde{g}$, and compute

$$\widehat{\theta} = \mathrm{argmin} \int B\Big(\frac{\widetilde{g}}{f_\theta}\Big) f_\theta \,\mathrm{d}y.$$

It is efficient at the model (close enough to the ML then) and robust.

- slim following (Basu et collegae; students of Lindsay; Hooker; a few other mild epicentres – though we've invented MWL and WIC).
- technicalities, fine-tuning choices, theorems are still lacking, and they're nitty-gritty demanding?
- statisticians are happy with ML and closer cousins?
- robustness isn't judged an important issue (despite propaganda in the 70ies and 80ies)?
- doesn't easily generalise to regression?

B. We need clearer conditions on $\widetilde{g}$ (and I'm bored by the default kernel method).

C. We should work out a suitable MDIC, a Minimum Disparity Information Criterion (non-trivial task).

The distance from truth to fitted model is $\int B(g/f_{\widehat\theta})f_{\widehat\theta}\,\mathrm{d}y$. What we directly observe is $D_n = \int B(\widetilde{g}/f_{\widehat\theta})f_{\widehat\theta}\,\mathrm{d}y$. Need analysis of

$$\begin{aligned}\Delta_n &= \int B\Big(\frac{\widetilde{g}}{f_{\widehat\theta}}\Big)f_{\widehat\theta}\,\mathrm{d}y - \int B\Big(\frac{g}{f_{\widehat\theta}}\Big)f_{\widehat\theta}\,\mathrm{d}y \\ &\doteq \int B'\Big(\frac{g}{f_{\widehat\theta}}\Big)(\widetilde{g}-g)\,\mathrm{d}y + \tfrac{1}{2}\int B''\Big(\frac{g}{f_{\widehat\theta}}\Big)\frac{(\widetilde{g}-g)^2}{f_{\widehat\theta}}\,\mathrm{d}y.\end{aligned}$$

For *some* $\widetilde{g}$, may show

$$(nh)^{1/2}\Delta_n \to_d \int B'\Big(\frac{g}{f_{\theta_0}}\Big)Z(y)\,\mathrm{d}y,$$

which leads to a $\mathrm{MDIC} = D_n + (nh)^{-1/2}\widehat{q}$ (it's messy, though).

D. Some of the schemes working well in $\dim = 1$ have trouble in $\dim \geq 2$ (other biases klick in). Can these matters be ameliorated?

E. Can MinDisp estimators take the place of ML estimators in FIC (the Gerda-Nils Focused Information Criterion)? I believe yes – and Sam-Erik is working on this. The root of the matter for ML is this: If data stem from $f_n(y) = f(y, \theta_0)\{1 + r(y)/\sqrt{n}\}$, then

$$\sqrt{n}(\widehat{\theta}_{\rm ML} - \theta_0) \to_d {\rm N}_p(J^{-1}b, J^{-1}), \quad b = \int f(y, \theta_0) r(y) u(y, \theta_0)\,{\rm d}y.$$

Task: Demonstrate that this holds (along with a certain list of other technical things) for classes of MinDisp.

F. Regression setting, model $f(y_i \,|\, x_i, \theta)$: choose $\widehat{\theta}$ to minimise

$$n^{-1} \sum_{i=1}^n \int B\Big(\frac{\widetilde{g}(y \,|\, x_i)}{f(y \,|\, x_i, \theta)}\Big) f(y \,|\, x_i, \theta)\,{\rm d}y.$$

G. MinDisp needs to be shown at work in clear well-motivated application stories.