Loss Function Selection and the use of Improper Models

Data-Driven Robust Bayesian Inference

Jack Jewson[1,2] and David Rossell[1,2]

ISBA World Meeting, Montreal, 29th June 2022

[1]Department of Economics, Universitat Pompeu Fabra
[2]Barcelona School of Economics

Algorithms or Models

- **Models** assign **probabilities** to observation - generally estimated using likelihood functions
- **Algorithms** are often more complicated yet deterministic functions that produce predictions - estimated using **loss-function** evaluated between predictions and observations

Breiman et al. (2001):

- “Statistical models are not realistic enough to represent reality in any useful manner
- Nor flexible enough to predict accurately complex real-world phenomena”

Efron (2020):

- “Abandoning mathematical models comes close to abandoning the historic scientific goal of understanding nature.”

Can we use the data to select between the two?

Challenge

- Models are defined using the units of probability - which must integrate/sum to 1
- But the units of an algorithm's loss can be arbitrary.
- We know how to select between probability models for data - Bayes factors, penalised likelihood (AIC, BIC ect.), ...
- But such methods fail if the loss under consideration cannot be reformulated as a normalised probability model - **non-integrability** means the scale is arbitrary

From losses to probability models

In particular,

- We can always turn a model into a loss using its log-likelihood - $\ell(x, \theta) = -\log f(x; \theta)$
- But the negative exponential of a loss $\tilde{f}(x; \theta) = \exp(-\ell(x, \theta))$ need not be an integrable probability model $\int \exp(-\ell(x, \theta)) dx = \infty$
- Hence we call this loss function selection
- Often this is not a problem e.g. Squared error loss $\mapsto$ Gaussian, Absolute error loss $\mapsto$ Laplace distribution - then we can use probability model selection
- But there are cases where this is not possible - then what?

Motivating Example

- Want to regress a response $y \in \mathbb{R}$ on some p-dimensional set of predictors X

Algorithm:

- Define a function that produces predictions for given X e.g. $\hat{y}(X, \beta) = X\beta$ - but in principle this could be more general
- Estimate parameters by minimising the loss between predictions and observations - $\hat{\beta} := \arg\min_\beta \sum_{i=1}^n \ell(y_i; \hat{y}(X_i, \beta))$

Model:

- Define a data generating density - $y|X \sim f(\cdot; X, \beta)$
- Estimate parameters by maximising the likelihood of the observations/ or conducting Bayesian updating

Least Squares

- Traditional least squares can be interpreted as a **model** or an **algorithm**
- The squared-error loss

$$\hat{\beta}_{LS} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - X_i\beta)^2$$

- Gaussian likelihood

$$\hat{\beta}_{LS} = \arg\max_{\beta} \sum_{i=1}^{n} \log \mathcal{N}\left(y_i; X_i\beta, \sigma^2\right)$$

- However, such a procedure is known not to be robust to outliers

Outlier Contamination

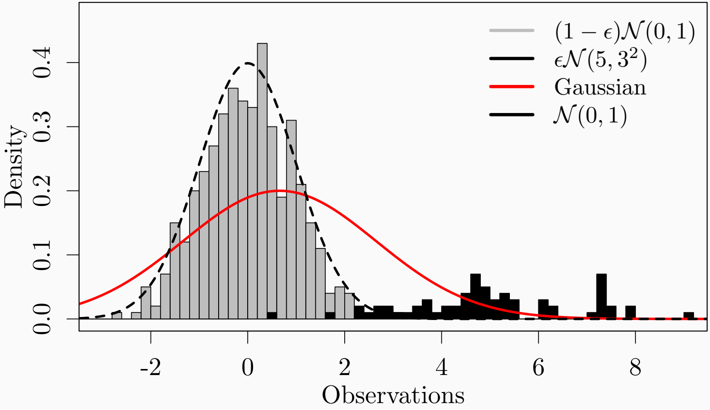

Figure 1 – Posterior predictive distribution fitting $\mathcal{N}(\mu, \sigma^2)$ to $g = 0.9\mathcal{N}(0,1) + 0.1\mathcal{N}(5, 3^2)$ using Bayes' rule

Robustification using Tukey's loss

- A traditional **algorithm** to robustify linear regression against outliers is to use Tukey's loss (Beaton & Tukey, 1974)

$$\ell_2(y_i; X_i, \theta_2, \kappa_2) = \begin{cases} \frac{(y_i - X_i\beta)^2}{2\sigma^2} - \frac{(y_i - X_i\beta)^4}{2\sigma^4\kappa_2^2} + \frac{(y_i - X_i\beta)^6}{6\sigma^6\kappa_2^4}, & \text{if } |y_i - X_i\beta| \leq \kappa_2\sigma \\ \frac{\kappa_2^2}{6}, & \text{otherwise} \end{cases}$$

$$\theta_2 = \{\beta, \sigma^2\}$$

- Hyperparameter κ_2 controls the 'robustness efficiency' trade-off
- $\kappa_2 = \infty$ recovers the Gaussian

Tukey's Loss

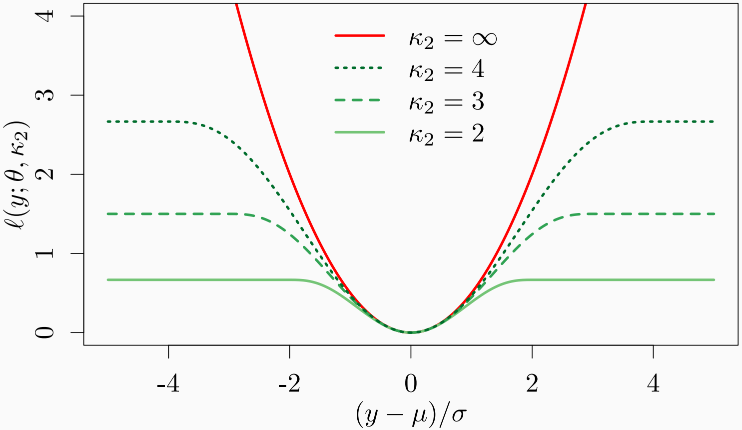

Figure 2 – Squared-error ($\kappa_2 = \infty$) (red) and Tukey's loss (green) for $\kappa_2 = 2$, 3 and 4

Robustification using Tukey's loss

However for $\kappa_2 < \infty$

$$\int \tilde{f}_2(y; X, \theta_2, \kappa_2) dy = \int \exp\left(-\ell_2(y; X, \theta_2, \kappa_2)\right) dy = \infty$$

Tukey's Improper Model

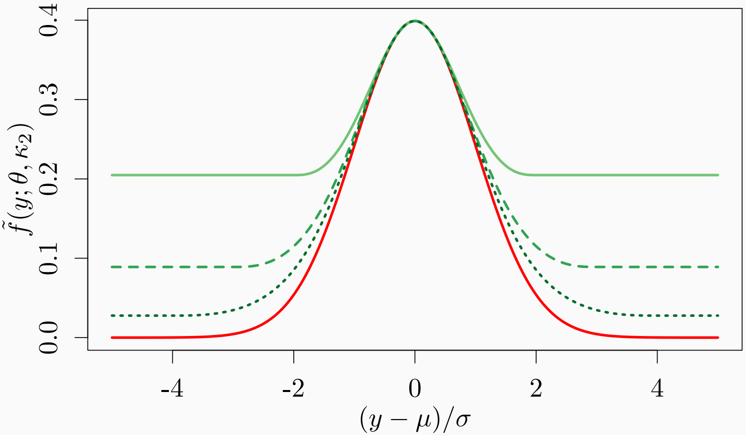

Figure 3 – Gaussian density ($\kappa_2 = \infty$) (red) and Tukey's loss improper density (green) for $\kappa_2 = 2$, 3 and 4. The improper densities for Tukey's loss are scaled to match the mode of the Gaussian density.

As a result...

$$\ell_2(y_i; X_i, \theta_2, \kappa_2) = \begin{cases} \frac{(y_i - X_i\beta)^2}{2\sigma^2} - \frac{(y_i - X_i\beta)^4}{2\sigma^4\kappa_2^2} + \frac{(y_i - X_i\beta)^6}{6\sigma^6\kappa_2^4}, & \text{if } |y_i - X_i\beta| \leq \kappa_2\sigma \\ \frac{\kappa_2^2}{6}, & \text{otherwise} \end{cases}$$

- Tukey's loss is strictly decreasing in κ_2
- Therefore, independent of the data

$$\arg\min_{\kappa_2} \ell_2(y_i; X_i, \theta_2, \kappa_2) = 0$$

- A result of the fact that κ_2 no longer indexes a probability density
- No coherent, data-driven way to set κ_2 in the literature

We want to be able to use the data to select between a Gaussian model and Tukey's loss, and if Tukey's loss is selected, estimate its hyperparameter

General Bayesian Updating

Can do we do Bayesian inference for the parameters of an algorithm?

- Bissiri, Holmes, and Walker (2016) consider inference about loss minimising parameter

$$\theta^* := \underset{\theta \in \Theta}{\arg\min} \int \ell(\theta, y) dG(y)$$

  where $G(\cdot)$ is the DGP of data y
- Given a prior $\pi(\theta)$ and observations $y_{1:n}$, a coherent and principled prior to posterior update is

$$\pi(\theta | y_{1:n}) = \frac{\pi(\theta) \exp\left(-\sum_{i=1}^{n} \ell(y_i; \theta)\right)}{\int \pi(\theta) \exp\left(-\sum_{i=1}^{n} \ell(y_i; \theta)\right) d\theta}$$

- Standard Bayesian inference recovered with $\ell(y; \theta) = -\log f(y; \theta)$

The robustness hyperparameter

- But General Bayesian Updating only ‘principled and coherent’ for loss minimising parameters.

$$\arg\min_{\kappa_2} \ell_2(y_i; X_i, \theta_2, \kappa_2) = 0$$

- We can produce ‘coherent’ posterior for $\beta|\kappa_2$, but not for κ_2 itself

Improper model, Proper DGP

- We wish to select between possibly improper models based on how well they capture the DGP
- As a result we need some notion of how an IMPROPER model can capture the PROPER DGP
- Clearly the improper model did not itself generate the data...
- But it could provide a 'better' representation of the DGP than any proper model available
- e.g Tukey's loss vs Gaussian under outliers

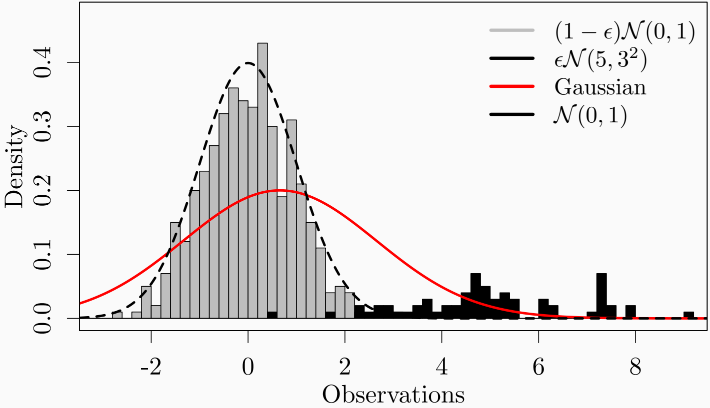

Figure 4 – Posterior predictive distribution fitting $\mathcal{N}(\mu, \sigma^2)$ to $g = 0.9\mathcal{N}(0,1) + 0.1\mathcal{N}(5,3^2)$ using Bayes' rule

Interpreting unnormalisable models

- But how can we interpret statements made by an unnormalisable pseudo density $\tilde{f}_k(y;\theta_k) \propto \exp\left(-\ell_k(y;\theta_k)\right)$
- Rather than statements about absolute probabilities, we interpret $\tilde{f}_k(y;\theta_k)$ as making statements about relative probabilities.
- e.g. Tukey's Loss
  - For any pair $|y_0 - X\beta|, |y_1 - X\beta| < \kappa_2\sigma$

    $$\frac{\tilde{f}_2(y_0;\theta_2,\kappa_2)}{\tilde{f}_2(y_1;\theta_2,\kappa_2)} \approx \frac{\mathcal{N}\left(y_0;X\beta,\sigma^2\right)}{\mathcal{N}\left(y_1;X\beta,\sigma^2\right)},$$

    these observations behave like Gaussian random variables
  - However, for $|y_0 - X\beta|, |y_1 - X\beta| > \kappa_2\sigma$

    $$\frac{\tilde{f}_2(y_0;\theta_2,\kappa_2)}{\tilde{f}_2(y_1;\theta_2,\kappa_2)} = 1$$

    all observations y with $|y - X\beta| > \kappa_2\sigma$ are equally 'likely'.

Fisher's-Divergence and the Hyvärinen-score

- Fisher's divergence can measure how well an improper model captures the relative probabilities of the DGP

$$D_F(g||\tilde{f}) := \frac{1}{2}\int ||\nabla_y \log g(y) - \nabla_y \log \tilde{f}(y)||^2 dG(y),$$
$$= \frac{1}{2}\int \left|\left|\lim_{\epsilon\to 0} \frac{\log \frac{g(y+\epsilon)}{g(y)} - \log \frac{\tilde{f}(y+\epsilon)}{\tilde{f}(y)}}{\epsilon}\right|\right|^2 g(y)dx.$$

where ∇_y is the gradient operator.

- Minimising Fisher's Divergence is the same as minimising the Hyvärinen-score (Hyvärinen, 2005) in expectation over DGP G

$$H(y;\tilde{f}_k(\cdot;\theta_k)) := 2\frac{\partial^2}{\partial y^2}\log \tilde{f}_k(y;\theta_k) + \left(\frac{\partial}{\partial y}\log \tilde{f}_k(y;\theta_k)\right)^2,$$

$\mathcal{H}$-Bayes - posterior

- Consider the general Bayesian update - applying the $\mathcal{H}$-score to $\tilde{f}_k(y;\theta_k,\kappa_k) = \exp(-\ell_k(y;\theta_k,\kappa_k)$

$$\pi^{\mathcal{H}}(\theta_k,\kappa_k|y_{1:n}) \propto \pi(\theta_k,\kappa_k)\exp(-\sum_{i=1}^{n} H(y_i;\tilde{f}_k(\cdot;\theta_k,\kappa_k))).$$

  (Giummolè, Mameli, Ruli, & Ventura, 2019) call this the $\mathcal{H}$-posterior
- Allows us to jointly produce posteriors over loss minimising parameters θ_k and loss hyperparameters κ_k (which may not desirably minimise the loss)

$\mathcal{H}$-Bayes - posterior consistency

Proposition 1
Let $y = (y_1, \ldots, y_n) \sim g$, $\tilde{\eta}_k = (\tilde{\theta}_k, \tilde{\kappa}_k)$ be $\mathcal{H}$-posterior MAP estimates, and $\eta_k^* = (\theta_k^*, \kappa_k^*)$ minimise Fisher's divergence from $f_k(\theta_k, \kappa_k)$ to g.

Under regularity conditions, as $n \to \infty$,

$$||\tilde{\eta}_k - \eta_k^*||_2 = O_p(1/\sqrt{n}).$$

where $|| \cdot ||_2$ is the L_2-norm.

$\mathcal{H}$-Bayes loss selection

Further

- Integrating out θ_k and κ_k produces scale invariant loss selection criteria

$$\mathcal{H}_k(y_{1:n}) = \int \pi(\theta_k, \kappa_k) \exp(-\sum_{i=1}^{n} H(y_i; \tilde{f}_k(\cdot; \theta_k, \kappa_k))) d\theta_k d\kappa_k.$$

- Analogue to Bayesian model selection with the marginal likelihood which has $-\log f(y_i; \theta)$ in place of the Hyvärinen-score

Laplace approximation to the $\mathcal{H}$-Bayes factor

- Loss selection can now happen through the $\mathcal{H}$-Bayes factor

$$B_{kl}^{(\mathcal{H})} := \frac{\mathcal{H}_k(y_{1:n})}{\mathcal{H}_l(y_{1:n})}$$

- We consider tractable Laplace approximations

$$\tilde{B}_{kl}^{(\mathcal{H})} := \frac{\tilde{\mathcal{H}}_k(y_{1:n})}{\tilde{\mathcal{H}}_l(y_{1:n})}$$

where $\eta_j = \{\theta_j, \kappa_j\}$, $j = k, l$

$$\tilde{\mathcal{H}}_k(y_{1:n}) := (2\pi)^{\frac{d_j}{2}} \frac{\pi\left(\tilde{\eta}_j^{(n)}\right) e^{-\sum_{i=1}^n H\left(y_i; f_k\left(\cdot; \tilde{\eta}_k^{(n)}\right)\right)}}{|A_j^{(n)}\left(\tilde{\eta}_j^{(n)}\right)|^{1/2}}$$

$$\tilde{\eta}_j^{(n)} := \arg\min_{\eta_j} \left\{ -\log \pi(\eta_j) + \sum_{i=1}^n H(y_i; f_j(\cdot; \eta_j)) \right\}.$$

Model selection consistency

Theorem 1
Under regularity conditions, and with $\#\eta_l > \#\eta_k$

1. *Suppose that* $\mathbb{E}_g[H(y; f_l(\cdot; \eta_l^*))] < \mathbb{E}_g[H(y; f_k(\cdot; \eta_k^*))]$. *Then*

$$\frac{1}{n} \log \tilde{B}_{kl}^{(\mathcal{H})} = \mathbb{E}_g[H(y; f_l(\cdot; \eta_l^*))] - \mathbb{E}_g[H(y; f_k(\cdot; \eta_k^*))] + o_p(1).$$

2. *Suppose that* $\mathbb{E}_g[H(y; f_l(\cdot; \eta_l^*))] = \mathbb{E}_g[H(y; f_k(\cdot; \eta_k^*))]$. *Then*

$$\log \tilde{B}_{kl}^{(\mathcal{H})} = \frac{d_l - d_k}{2} \log(n) + O_p(1).$$

- $\tilde{B}_{kl}^{(\mathcal{H})}$ provides consistent model selection to the model minimising the Fisher's divergence.
    1. selects more complicated l at exponential rate
    2. selects simpler k at polynomial rate

Two-stage Inference

Our consistency results allow us to advocate a two-stage inference procedure

- Use $\mathcal{H}_k(y_{1:n})$ to select between available models and algorithms/losses
- If a proper probability model was selected - revert to ordinary Bayesian inference
- If an improper model was selected -
    - Consider joint inference on θ_k and κ_k using the $\mathcal{H}$-posterior, or ...
    - Estimate κ_k using the $\mathcal{H}$-posterior and produce a general Bayesian posterior for $\theta_k|\kappa_k$

Proof of concept - Marginal $\mathcal{H}$-score

- For illustrative purposes, consider the marginal $\mathcal{H}$-score in κ_2

$$\mathcal{H}_2(y_{1:n};\kappa_2) = \int \pi(\theta_2) \exp(-\sum_{i=1}^{n} H(y_i; f_k(\cdot;\theta_2,\kappa_2)))d\theta_k.$$

- We simulate $n = 500$ observations from the ϵ-contamination model $g(x) = 0.9\mathcal{N}(0,1) + 0.1\mathcal{N}(5,3)$

Tukey's loss outlier detection

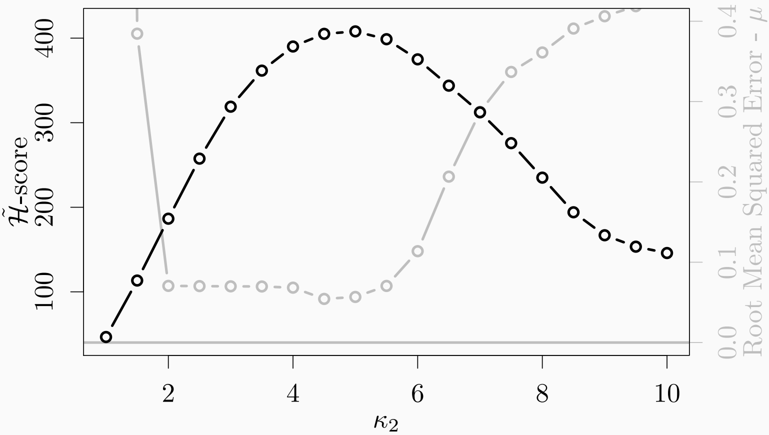

Figure 5 – Marginal $\mathcal{H}$-score $\mathcal{H}_2(y; \kappa_2)$ (black) and asymptotic approximation to the RMSE of $\hat{\beta}(\kappa_2)$ (grey) for several κ_2.

Tukey's loss outlier detection

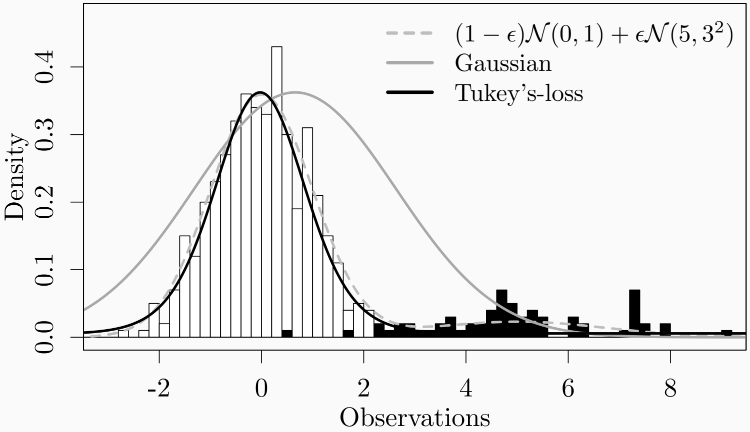

Figure 6 – Tukey's-loss vs the Gaussian model for ϵ-contaminated data

Summary

- Previously difficult to ‘learn’ parameters of unnormalisable pseudo-probability models e.g. Tukey’s loss
- and select between probability models and algorithms whose losses define unnormalisable probability models
- Interpreting them as providing relative, rather than absolute probabilities, naturally leads us to the Hyvärinen-score
- Allows for parameter estimation and loss function selection that is invariant to possibly infinite normalising constants
- ‘Data-driven robustness’

Arxiv Preprint

“General Bayesian Loss Function Selection and the use of Improper Models” - Jewson and Rossell (2021)
arXiv preprint arXiv:2106.01214

- Consider non-local prior (NLPs) (Johnson & Rossell, 2010, 2012) to improve the rate of selecting simpler (...proper) model
- Robust regression examples
- Applications to Bayesian Kernel Density Estimation

References

Beaton, A. E., & Tukey, J. W. (1974). The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. *Technometrics*, *16*(2), 147–185.

Bissiri, P., Holmes, C., & Walker, S. G. (2016). A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

Breiman, L., et al. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, *16*(3), 199–231.

Efron, B. (2020). Prediction, estimation, and attribution. *Journal of the American Statistical Association*, *115*(530), 636–655.

Giummolè, F., Mameli, V., Ruli, E., & Ventura, L. (2019). Objective bayesian inference with proper scoring rules. *Test*, *28*(3), 728–755.

Hyvärinen, A. (2005). Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, *6*(Apr), 695–709.

Johnson, V. E., & Rossell, D. (2010). On the use of non-local prior densities in bayesian hypothesis tests. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *72*(2), 143–170.

Johnson, V. E., & Rossell, D. (2012). Bayesian model selection in high-dimensional settings. *Journal of the American Statistical Association*, *107*(498), 649–660.