Improper models for data analysis

David Rossell[1,2] [work led by Jack Jewson[1,2]]

OBayes, Santa Cruz, Sep 2022

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

Models vs. loss functions

Problem: use probability model or loss function? What model/loss?

- **Models** facilitate interpretation & assign probabilities. Model assumptions can be checked
- **Losses** produce estimates and predictions. Often defined to attain desirable properties, e.g. robustness

Given data $y = (y_1, \ldots, y_n)$, a model k, the likelihood $f_k(y; \theta_k)$ defines a loss

$$\ell_k(y; \theta_k) = -\log f_k(y; \theta_k)$$

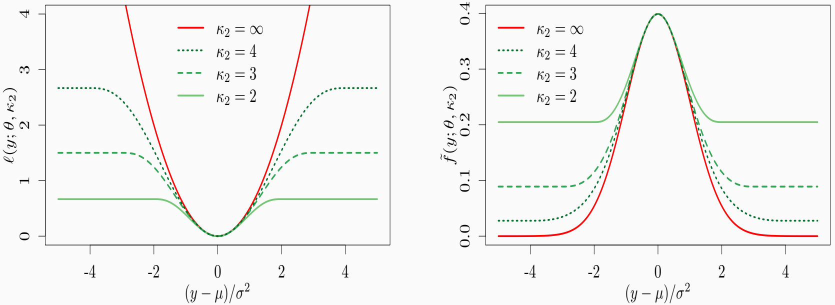
Example. $y_i \sim N(x_i^T \theta, \sigma^2 I)$ defines least-squares loss $\sum_{i=1}^n (y_i - x_i^T \theta)^2$

Given a loss $\ell_k(y; \theta_k)$, $f_k(y; \theta_k) = \exp\{-\ell_k(y; \theta_k)\}$ may not be proper wrt y

Key: to assess which loss is "best" for y, assess the (possibly improper) model that would've implied each loss

Example. Tukey's Loss (Beaton & Tukey, 1974)

Least-squares and Tukey's loss with cut-off parameter $\kappa = 2, 3, 4$



$\kappa = \infty$ gives least-squares loss. For $\kappa < \infty$

$$\int \exp\{-\ell(y; \theta, \kappa)\}\, dy = \infty \Rightarrow \text{Improper model}$$

The debate

On one hand

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

  Breiman et al. (2001)

- Losses often used in machine learning, robust statistics etc.

On the other hand

- “All models are wrong but some are useful”

  G. E. P. Box

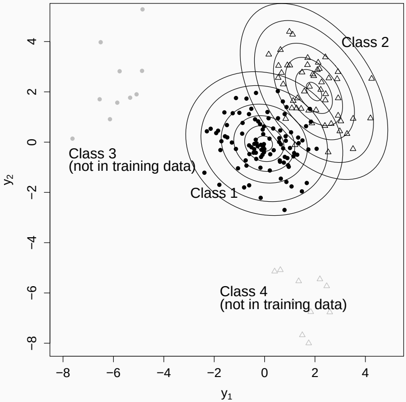
- “Abandoning mathematical models comes close to abandoning the historic scientific goal of understanding nature”

  Efron (2020)

- Models help interpret the phenomenon under study. Probabilistic forecasts portray uncertainty

Further examples

Open-set classification



Improper priors. $y_i \mid \theta \sim N(\theta, 1), p(\theta) \propto 1 \Rightarrow p(y_i) \propto 1$

Theorem. Any improper sigma-finite $p(y_i \mid \theta)$ can be represented as mixture of proper measure wrt improper prior

$$p(y_i \mid \theta) = \int \underbrace{p(y_i \mid \theta, \xi)}_{\textit{proper}} \, d \underbrace{P(\xi)}_{\textit{improper}}$$

Incorporating losses in Bayesian thinking

[PAC-Bayes (McAllester, 1999), Gibbs posteriors, safe Bayes (Grünwald, 2012), generalized Bayes (Bissiri, Holmes, & Walker, 2016)]

One may define a posterior distribution on θ using losses

$$p(\theta \mid y) \propto \exp\{-\kappa \ell(y;\theta)\}\, p(\theta)$$

where $p(\theta)$ is some prior, and $\kappa > 0$ given

- If loss defines proper model on y, back to standard Bayes
- Else, how to interpret implied predictive model on future y's?
- How to choose κ (learning rate)? Similar hyper-parameter issues: Tukey's cutoff, kernel density bandwidth...

Example: $\ell(y,\theta) = \kappa \sum_{i=1}^{n} (y_i - x_i^T \theta)^2$. We can use associated proper model to learn κ (Normal precision) fitting data "best". What if model is improper?

Learning hyper-parameters

Given κ and data-generating G, PAC-Bayes et al target

$$\theta^* := \arg\min_{\theta \in \Theta} \int \ell(y; \theta, \kappa) dG(y)$$

One obtains a 'coherent' posterior for $\theta \mid \kappa$, but not for κ

Example. For Tukey's loss, consider

$$p(\theta, \kappa \mid y) \propto \exp\{-\ell(y; \theta, \kappa)\} p(\theta, \kappa)$$

Tukey's loss is strictly decreasing in κ. Hence, regardless of y

$$\arg\min_{\kappa \geq 0} \ell(y; \theta, \kappa) = 0$$

Same for the marginal "likelihood"

$$\arg\max_{\kappa \geq 0} \int \exp\{-\ell(y; \theta, \kappa)\} p(\theta, \kappa) d\theta = 0$$

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

Goal

Since we observed y, it was generated by some proper distribution $G(y)$

Goal. Choose model or loss (and its hyper-parameters) that best approximates G

- How to define “best”?
- Cross-validation & other standard tools not applicable (which loss should one cross-validate?)
- Parsimony is key: choose smaller model when it provides a good approximation, e.g. Gaussian over Tukey’s

Many tools available if all losses define a proper model. Otherwise, unclear what to do

Methodology

Interpreting an improper model

How to interpret an improper density

$$f(y; \theta, \kappa) \propto \exp\left\{-\ell(y; \theta, \kappa)\right\}$$

Rather than giving absolute probabilities, we interpret $f(y; \theta, \kappa)$ as making statements about "relative probabilities"

$$\frac{f(y_0; \theta, \kappa)}{f(y_1; \theta, \kappa)}$$

describes how much more likely is it to observe $y = y_0$ than $y = y_1$

Example. Tukey's loss. If both $|y_0 - \theta|, |y_1 - \theta| < \kappa\sigma$

$$\frac{f(y_0; \theta, \kappa)}{f(y_1; \theta, \kappa)} \approx \frac{\mathcal{N}\left(y_0; \theta, \sigma^2\right)}{\mathcal{N}\left(y_1; \theta, \sigma^2\right)},$$

However, for $|y_0 - \theta|, |y_1 - \theta| > \kappa\sigma$

$$\frac{f(y_0, \theta, \kappa)}{f(y_1, \theta, \kappa)} = 1$$

all observations $|y - \theta| > \kappa\sigma$ are equally 'likely'

Fisher's-Divergence

If we interpret $f(y; \theta, \kappa)$ via "relative probabilities", then (θ, κ) should be set to accurately capture the "relative probabilities" of $G(y)$, the DGP

Fisher's divergence

$$D_F(g||f) := \frac{1}{2} \int ||\nabla_y \log g(y) - \nabla_y \log f(y; \theta, \kappa)||^2 g(y) dy,$$

$$= \frac{1}{2} \int \left|\left| \lim_{\epsilon \to 0} \frac{\log \frac{g(y+\epsilon)}{g(y)} - \log \frac{f(y+\epsilon; \theta, \kappa)}{f(y; \theta, \kappa)}}{\epsilon} \right|\right|^2 g(y) dy$$

Compares f's infinitesimal "relative probabilities" to g's

Key: invariant to normalizing constant. If $\tilde{f}(y; \theta, \kappa) = \frac{f(y; \theta, \kappa)}{Z(\theta, \kappa)}$ then

$$\nabla_y \log \tilde{f}(y; \theta, \kappa) = \nabla_y \log f(y; \theta, \kappa)$$

- FD (and generalizations) allow working with improper models
- Methods for intractable, but finite, normalization constants don't work (contrastive divergence, minimum probability flow etc.)

The Hyvärinen score

Minimizing Fisher's Divergence equivalent to minimizing the Hyvärinen-score (Hyvärinen, 2005) in expectation over G (under minimal tail conditions)

$$\arg\min_{\theta,\kappa} D_F(g||f) = \arg\min_{\theta,\kappa} \mathbb{E}_G\left[H(y; f(\cdot;\theta,\kappa))\right]$$

where for univariate y

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

Since $y_1, \ldots, y_n \sim G$, the loss $n^{-1}\sum_i H(y_i; f(\cdot,\theta,\kappa)) \approx \mathbb{E}_G\left[H(y; f(\cdot;\theta,\kappa))\right]$

$\mathcal{H}$-posterior

(Giummolè, Mameli, Ruli, & Ventura, 2019)

Since $\sum_i H(y_i; f(\cdot, \theta, \kappa))$ defines a loss, where $f(y; \theta, \kappa) \propto \exp\{-\ell(y; \theta, \kappa)\}$, consider the general Bayes posterior

$$p(\theta, \kappa \mid y) \propto p(\theta, \kappa) \exp\left\{-\sum_{i=1}^{n} H(y_i; f(\cdot; \theta, \kappa))\right\}$$

The $\mathcal{H}$-posterior gives joint inference on θ and hyperparameters κ

- Learning rate in PAC-Bayes / general Bayes
- Cutoff parameter in Tukey's loss
- Bandwidth parameter in kernel density estimation

$\mathcal{H}$-posterior consistency

Theorem 1. Let $y_i \sim g$ iid, $(\tilde{\theta}, \tilde{\kappa})$ be the mode of the $\mathcal{H}$-posterior, and (θ^*, κ^*) minimize Fisher's divergence from $f(y; \theta, \kappa)$ to $g(y)$.

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

$$\left|\left|(\tilde{\theta}, \tilde{\kappa}) - (\theta^*, \kappa^*)\right|\right|_2 = O_p(1/\sqrt{n})$$

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

- Even for improper models, learn the FD-optimal parameter values at the usual $\sqrt{n}$ rate
- Similar result to Dawid, Musio, and Ventura (2016), but we allow for κ^* at the boundary, e.g. in Tukey's loss $1/\kappa^* = 0$ gives the Gaussian model

Integrated $\mathcal{H}$-score for model selection

We also want a method to choose among several models $f_1, \ldots, f_K$

Analagously to the marginal likelihood in Bayesian model selection, consider

$$\mathcal{H}_k(y) = \int \exp\left\{ -\sum_{i=1}^{n} H(y_i; f_k(\cdot; \theta_k, \kappa_k)) \right\} p_k(\theta, \kappa) d\theta_k d\kappa_k$$

For analytical & computational tractability, we use Laplace approximations

Select model with highest $\mathcal{H}_k$. Equivalently, the $\mathcal{H}$-Bayes factor

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

Model selection consistency

Consider two models k, l of dimension d_k, d_l.

Theorem 2. Under regularity conditions, as $n \to \infty$

1. If model k closer to g in Fisher's div,

$$\log B_{kl}^{(\mathcal{H})} = n(\underbrace{\mathbb{E}_g[H(y; f_l(\cdot; \eta_l^*))] - \mathbb{E}_g[H(y; f_k(\cdot; \eta_k^*))]}_{>0} + o_p(1))$$

2. If both models have same Fisher's div to g (nested models)

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

Standard Bayesian model selection rates, based on Fisher's div rather than Kullback-Leibler

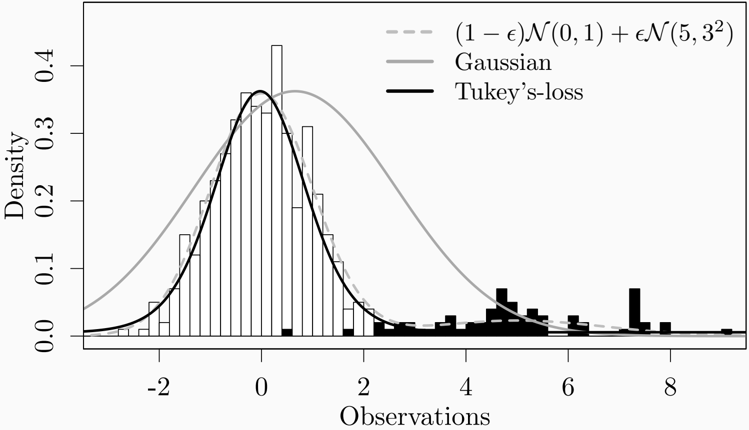
- If both models equally good, choose smaller one, e.g. Gaussian over Tukey's
- If mode occurs at the boundary, Theorem 2 may not hold
- Non-local priors (Johnson & Rossell, 2010) improve rates for Part 2 and allow for mode at the boundary

Experiments

Proof of concept. Robustness-efficienty trade-off

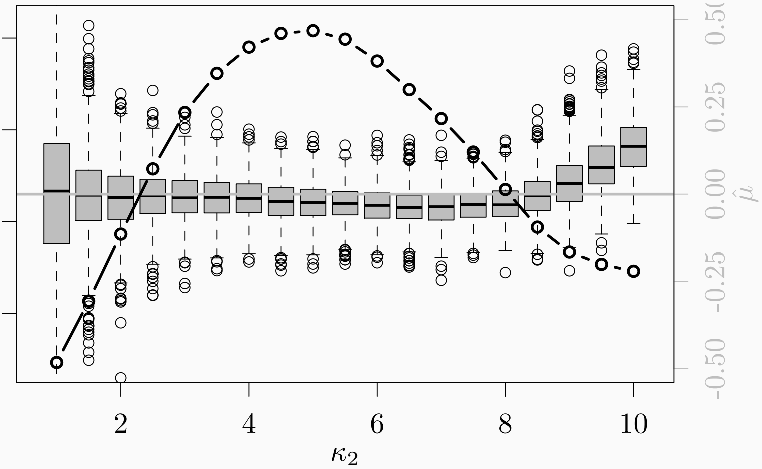
Simulate $n = 500$ observations from $g(y) = 0.9\mathcal{N}(y; 0, 1) + 0.1\mathcal{N}(y; 5, 3^2)$

- Tukey with small κ: very robust, less efficient (if y near-Normal)
- Tukey with large κ: less robust, more efficient (if y near-Normal)



Bias and variance vs. κ

- Box plots of $\hat{\mu}(\kappa)$ across 1,000 simulations
- Grey line: true mean of uncontaminated component
- Black: marginal $\mathcal{H}$-score $\mathcal{H}(y; \kappa)$



Model selection consistency

Simulate data with $n = 100, 1000, 10^4$ and 10^5 from

$$y_i \sim \mathcal{N}(x_i^T \beta, 1)$$

- 5 covariates $x_i \sim N(0, \Sigma)$ with unit variances and 0.5 correlation
- $\beta = (0, 0.5, 1, 1.5, 0, 0)$ (including the intercept)

Goal: select Gaussian vs. Tukey's model

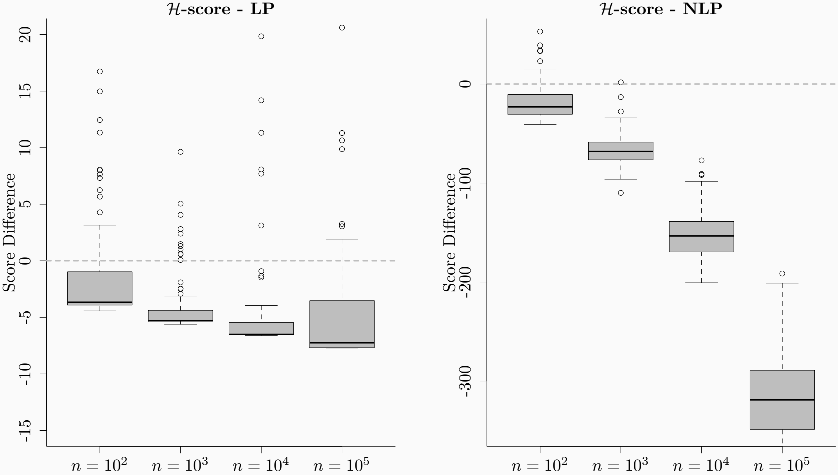
- Same priors on (β, σ^2) under both models
- Local prior: half Gaussian prior on $\nu = \frac{1}{\kappa^2}$
- Non-local prior: inverse-Gamma prior on ν
- Prior parameters assign 0.95 probability to $\kappa \in (1, 3)$

Local vs non-local prior on Tukey's cutoff

log $\mathcal{H}$-Bayes Factor. Negative values correctly select Gaussian model

Local prior

Non-local prior

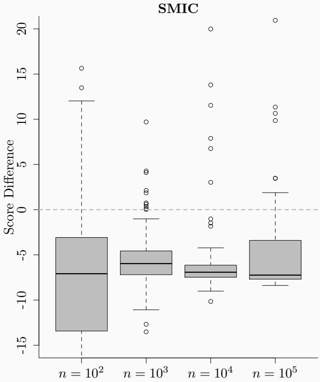


SMIC (Matsuda et al., 2019)

Score matching information criteria (Matsuda, Uehara, & Hyvarinen, 2019)

- Estimate Fisher's div. by correcting bias of in-sample Hyvärinen score
- Predictive criteria similar to the AIC (no consistent model selection)
- Improper models not considered (but feasible, in principle)

$SMIC_1(y) - SMIC_2(y)$. Negative values correctly select the Gaussian model



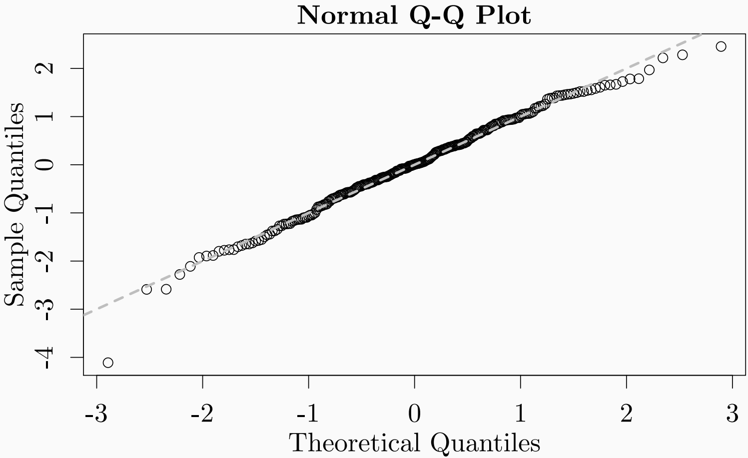
TGF-β data (Calon et al., 2012)

- Gene expression data for $n = 262$ colon cancer patients
- TGF-β is an important gene for colon cancer metastasis
- We regress TGF-β on the 7 genes in the 'TGF-β 1 pathway'

Results

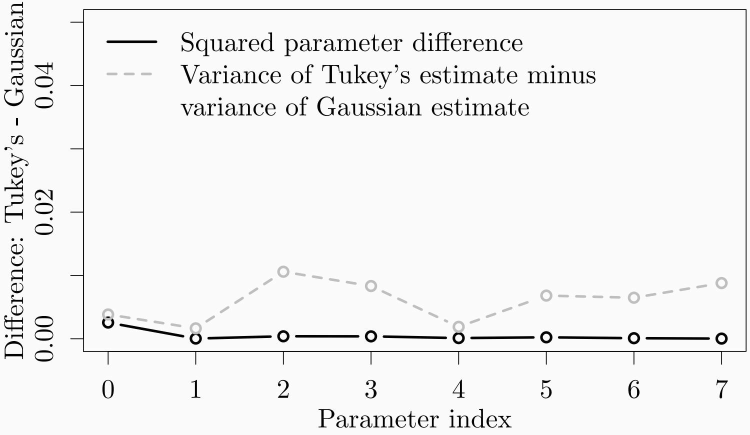
- Strong evidence for Gaussian $\mathcal{H}_1(y) = 272.9$ vs. Tukey's $\mathcal{H}_2(y) = 233.9$
- Rossell and Rubio (2018) also found evidence for Gaussian over (thicker) Laplace tails
- Similar $\hat{\beta}_j$ for both models, Gaussian has smaller Var($\hat{\beta}_j$) (bootstrap)

TGF-β data



TGF-β data - Bootstrap parameter variance estimates

- Black: Point estimate $\hat{\beta}_j$ under Tukey's - Gaussian loss
- Grey: variance under Tukey's - variance under normal (bootstrap)



Similar point estimates, but Gaussian more efficient

DLD data (Yuan et al., 2016)

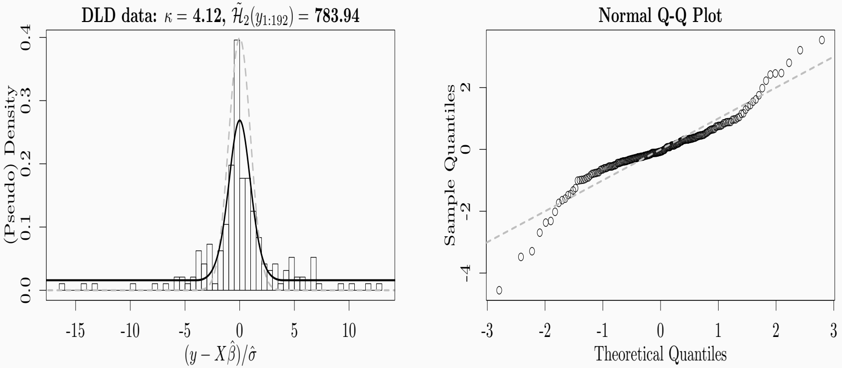
- RNA-sequencing gene expression data for $n = 192$ cancer patients
- DLD gene can perform several functions such as metabolism regulation
- For illustration, we select the 15 variables with the 5 highest loadings in each of the first 3 principal components

Results

- Strong evidence for Tukey's model ($\mathcal{H}_1(y) = 155.6$ vs $\mathcal{H}_2(y) = 783.9$)
- Rossell and Rubio (2018) selected Laplace over Gaussian tails
- $\hat{\beta}_j$'s from each model quite different, Var($\hat{\beta}_j$) lower for Tukey's

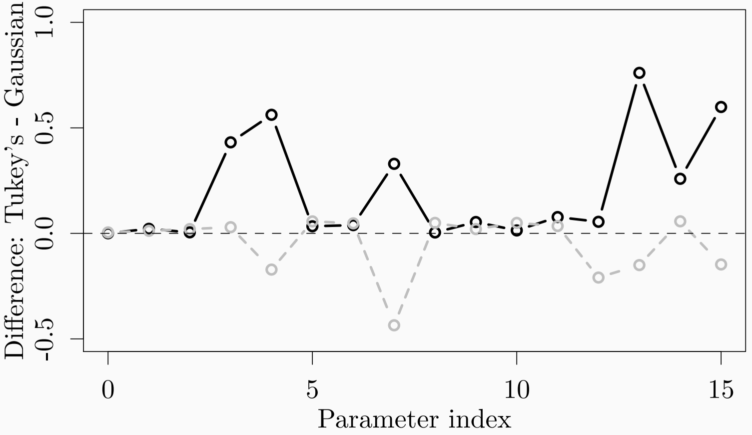
DLD data

Fitted Tukey's (black) vs Gaussian model (grey), and qq-normal plot



DLD data - Bootstrap parameter variance estimates

- Black: Point estimate $\hat{\beta}_j$ under Tukey's - Gaussian loss
- Grey: Variance under Tukey's - Variance under normal (bootstrap)



Kernel Density Estimation

Consider the KDE

$$\tilde{f}(x;\kappa) = \frac{1}{n\kappa}\sum_{i=1}^{n} K\left(\frac{x-y_i}{\kappa}\right)$$

where $\kappa > 0$ is the bandwidth

Tempting to define a likelihood for $y = (y_1, \ldots, y_n)$, and set prior on κ

$$f(y;\kappa) = \prod_{i=1}^{n} \tilde{f}(y_i;\kappa)$$

However, easy to see that $\int f(y;\kappa)dy = \infty$

Consequence: Bayesians don't do KDE

Proposal: consider the loss

$$\ell(y;\kappa,w) = -w\sum_{i=1}^{n} \log \tilde{f}(y_i;\kappa)$$

$w > 0$ is a tempering hyper-parameter, to be learnt from data

Mixture Model Experiments

Compare to

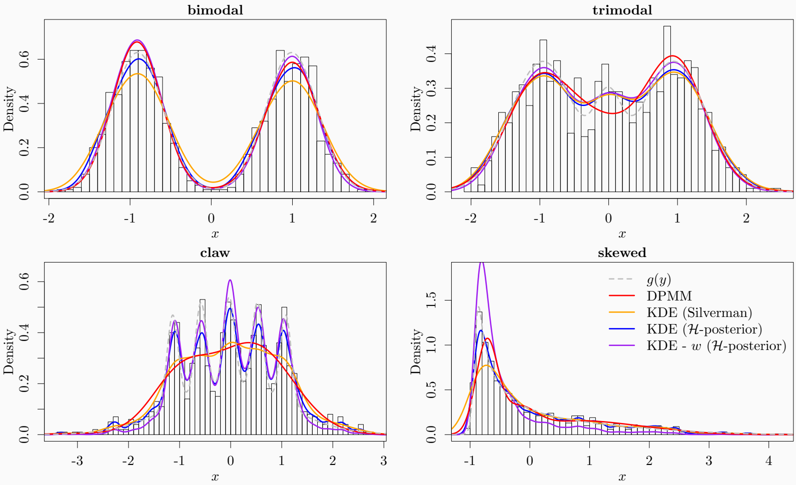
- R's '*density*'. Estimates the bandwidth via cross-validated MSE
- R package '*dirichletprocess*'. Uses Gaussian Dirichlet process mixture

Consider simulation settings from Marron and Wand (1992), all of the form

$$g(y) = \sum_{j=1}^{J} \pi_j N(y; \mu_j, \sigma_j)$$

Results

Similar estimates for bimodal/skewed. Tracks modes better in trimodal/claw



Not claiming that the $\mathcal{H}$-score leads to better density estimation. Just that it seems competitive, and opens a new avenue for Bayesian non-parametrics

Take-home messages

Viewing losses as defining (possibly improper) models enriches the probabilistic data analysis toolkit

- Interpretable via "relative probabilities"
- Decide between models vs losses in data-based manner
- Hyper-parameters affect the "model fit", and can hence be learnt

Future work

- Alternatives to Fisher divergence, particularly for multivariate/dependent data
- How to do model checking for improper models
- Applications: open-set classif., improper random effects etc.

Main reference

Jewson & Rossell. General Bayesian Loss Function Selection and the use of Improper Models. JRSS-B 2022 (in press). Also at arXiv:2106.01214

Funding

Jack: Juan de la Cierva (Spain)

David: Huawei Research grants, Ramón y Cajal 2015-18544 (Spain), Europa Excelencia EUR2020-112096 (Spain), Programa Estatal I+D+i PGC2018-101643-B-I00 (Spain)

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.

Calon, A., Espinet, E., Palomo-Ponce, S., Tauriello, D. V., Iglesias, M., Céspedes, M. V., . . . others (2012). Dependency of colorectal cancer on a TGF-β-driven program in stromal cells for metastasis initiation. *Cancer cell*, *22*(5), 571–584.

Dawid, A. P., Musio, M., & Ventura, L. (2016). Minimum scoring rule inference. *Scandinavian Journal of Statistics*, *43*(1), 123–138.

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.

Grünwald, P. (2012). The safe bayesian. In *International conference on algorithmic learning theory* (pp. 169–183).

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.

Lewis, J. R., MacEachern, S. N., & Lee, Y. (2021). Bayesian restricted likelihood methods: Conditioning on insufficient statistics in bayesian regression. *Bayesian Analysis*, *1*(1), 1–38.

Marron, J. S., & Wand, M. P. (1992). Exact mean integrated squared error. *The Annals of Statistics*, 712–736.

Matsuda, T., Uehara, M., & Hyvarinen, A. (2019). Information criteria for non-normalized models. *arXiv preprint arXiv:1905.05976*.

McAllester, D. A. (1999). Some PAC-Bayesian theorems. *Machine Learning*, *37*(3), 355–363.

Rossell, D., & Rubio, F. J. (2018). Tractable bayesian variable selection: beyond normality. *Journal of the American Statistical Association*, *113*(524), 1742–1758.

Yuan, T., Huang, X., Woodcock, M., Du, M., Dittmar, R., Wang, Y., . . . others (2016). Plasma extracellular rna profiles in healthy and cancer patients. *Scientific reports*, *6*(1), 1–11.