Bayes assisted goodness-of-fit for von Mises regression

Richard Lockhart + Contreras, Stephens, Sun

Simon Fraser University

ADISTA
Brussels, May 22, 2014

Outline

Apology

David Blackwell said

I've worked in so many areas — I'm sort of a dilettante. Basically, I'm not interested in doing research and I never have been. I'm interested in understanding, which is quite a different thing.

For when I don't finish

- Points of contact with talks of Garcia-Portugués, Jupp, Pewsey, Swan, Verdebout, at least. I am very grateful to have been invited here.
- We do frequentist model assessment via Bayes.
- We construct goodness-of-fit tests for directional regression models.
- They maximize a certain average power.
- Some regression models have complete sufficient statistic.
- For these models best test is conditional.
- Implementation via Markov Chain Monte Carlo.
- Methodology allows diagnosis after testing, in principle.

Some general principles

- Admissible nearly implies Bayes.
- Describe model departures via a stochastic process prior on a likelihood ratio.
- To test a goodness-of-fit null, pretend parameters are known, test fit, average results with respect to a posterior on the null.
- To test fit to an assumption about unobservable quantities: pretend they were observed and average results wrt a posterior.

Exponential family: von Mises distribution

Von Mises density relative to arc length on the unit circle:

$$f(y;\tau) = \frac{1}{2\pi I_o(\|\tau\|)} \exp\{\tau^T y\}$$

- y is a unit vector.
- $\tau = \kappa(\cos\theta_0, \sin\theta_0)^T \in \mathbb{R}^2$.
- *modal angle* θ_0.
- *concentration parameter* κ.

Add covariates

Responses y_i

Covariates for observation i form **matrix** c_i ($p \times 2$).

Model: for some $p \times 1$ parameter vector β we have:

$$\tau_i = c_i^T \beta.$$

Likelihood is (ignoring powers of 2π):

$$L(\beta) = \exp\{\beta' S - \sum \log I_o(\|\tau_i\|)\}$$

where

$$S = \sum_i c_i y_i$$

is a ($p \times 1$) complete sufficient statistic.

Location Finding

Guttorp and Lockhart (*JASA*, 1988) Bayesian aircraft finding.

- Detectors at locations $x_1, \ldots, x_n$ each in $\mathbb{R}^2$.
- Lost object at $v \in \mathbb{R}^2$.
- Take bearings y_i at each detector.
- Model y_i as von Mises unit vector with parameter

$$\tau_i = (v - x_i)\kappa = \begin{bmatrix} 1 & 0 & -x_{i1} \\ 0 & 1 & -x_{i2} \end{bmatrix} \begin{bmatrix} \kappa v_1 \\ \kappa v_2 \\ \kappa \end{bmatrix} \equiv c_i \beta$$

Location Finding

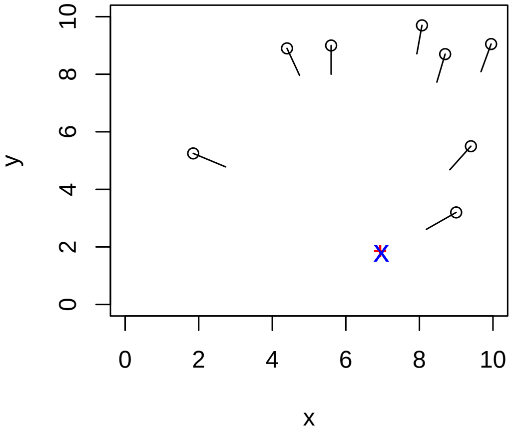
Guttorp and Lockhart (*JASA*, 1988) Bayesian aircraft finding.

- Detectors at locations $x_1, \dots, x_n$ each in $\mathbb{R}^2$.
- Lost object at $v \in \mathbb{R}^2$.
- Take bearings y_i at each detector.
- Model y_i as von Mises unit vector with parameter

$$\tau_i = (v - x_i)\kappa = \begin{bmatrix} 1 & 0 & -x_{i1} \\ 0 & 1 & -x_{i2} \end{bmatrix} \begin{bmatrix} \kappa v_1 \\ \kappa v_2 \\ \kappa \end{bmatrix} \equiv c_i \beta$$

- Warning: unrealistic model chosen to illustrate some ideas.
- Concentration is higher when object is further from the detector.

The Data



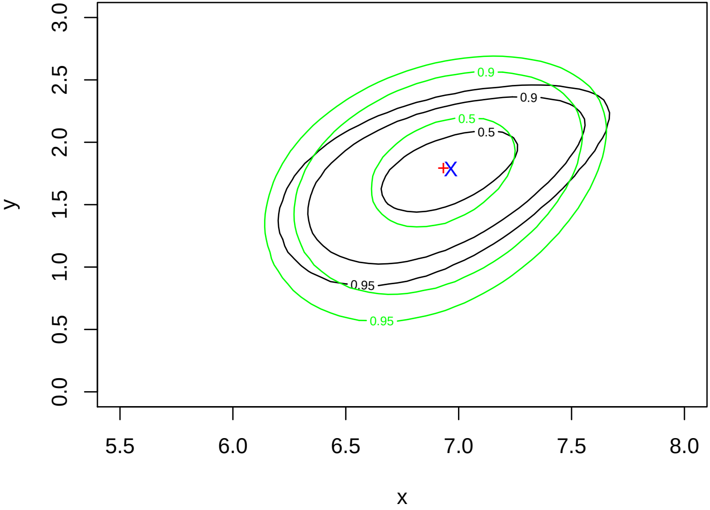
Parameters and statistics

- Parameter vector $\beta = (v, \kappa)$ has three components.
- Complete sufficient statistic has three components:

$$S = (\sum x_i^t y_i, \sum y_i).$$

- Peter and I wrote down obvious prior for a different model and worked with it.
- Next page has posteriors (for x) for old and new models.

Posteriors for location



Our focus — goodness-of-fit

Is the von Mises model any good? [I know we don't have enough data to answer the question.]

Ingredients of today's proposal:

- Null hypothesis: the model specified above is right.
- Alternative: von Mises assumption is wrong.
- Define priors on null and alternative hypothesis.
- Maximize average power subject to level α.
- Classical Neyman-Pearson approach.
- Prior makes alternative hypothesis simple.
- Existence of complete sufficient statistic permits solution to optimization problem.

Priors on the alternative?

- On null $y_i|\beta$ has density $f_i(y_i|\beta)$.
- On alternative y_i has density $g_i(y_i,\beta) = \ell_i(y_i,\beta)f(y_i,\beta)$.
- Describe prior on alternative in two parts:
- First pick parameter value β at random – density $\pi_1(\beta)$.
- Then model likelihood ratio ℓ as stochastic process.

Priors on the alternative — more detail

- Our choice: for stochastic process Z:

$$\ell_i(y_i, \beta) = C \exp(aZ(y_i, \beta, x_i)/\sqrt{n})$$

- Many choices for structure of Z: one convenient one is

$$Z(y, \beta, x) = Z^*(F(y, \beta, x))$$

- Take $Z^*(\cdot)$ to be a "stationary" Gaussian process on 'circle'.
- Factor of $n^{-1/2}$ gives contiguous alternatives (Le Cam).
- C is approximated by $\exp(-\int Z^2(y)dy/(2n))$.
- Today's favourite choice: Z has mean 0 and

$$\mathrm{Cov}(Z(s), Z(t)) = \frac{\rho(\cos(2\pi(t-s)) - \rho)}{1 + \rho^2 - 2\rho\cos(2\pi(t-s))}.$$

Neyman Pearson lemma

- Neyman Pearson: reject for large values of

$$\frac{\text{Marginal density on alternative}}{\text{Marginal density on Null}}$$

- Has form

$$\mathrm{E}\left[\exp\{T(\text{data},\text{pars})\}|\text{data}\right] \times \text{ratio of null marginals}$$

Neyman Pearson lemma

- Neyman Pearson: reject for large values of

$$\frac{\text{Marginal density on alternative}}{\text{Marginal density on Null}}$$

- Has form

$$\mathrm{E}\left[\exp\{T(\text{data}, \text{pars})\}|\text{data}\right] \times \text{ratio of null marginals}$$

- Need to find least favourable prior on null for denominator.

Neyman Pearson lemma

- Neyman Pearson: reject for large values of

$$\frac{\text{Marginal density on alternative}}{\text{Marginal density on Null}}$$

- Has form

$$\mathrm{E}\left[\exp\{T(\text{data},\text{pars})\}|\text{data}\right] \times \text{ratio of null marginals}$$

- Need to find least favourable prior on null for denominator.
- Can skip this if there is a complete sufficient statistic!

Conditional tests

If

$$P_{H_o,\beta}(T \geq t) \leq \alpha \leq P_{H_1}(T \geq t)$$

for all β then

$$P_{H_o}(T \geq t|S) \equiv \alpha$$

So do conditional test.

Conditional tests

If

$$P_{H_o,\beta}(T \geq t) \leq \alpha \leq P_{H_1}(T \geq t)$$

for all β then

$$P_{H_o}(T \geq t|S) \equiv \alpha$$

So do conditional test.

But how do we do it and how well does it work?

Implementation

How do we do compute *p*-value?

- Exact distribution is essentially impossible.
- Simulation.
  - Draw many samples from conditional dist of data given S.
  - Markov Chain Monte Carlo if that can't be done.
- Approximation.

MCMC

- Often can't easily generate samples $(y_1, \ldots, y_n)|S$.
- So find a Markov Chain whose stationary distribution is

  $$\mathcal{L}(y_1, \ldots, y_n|S)$$

  and which we can simulate.
- We have so far used Gibbs sampler.

Gibbs sampling for von Mises case -3

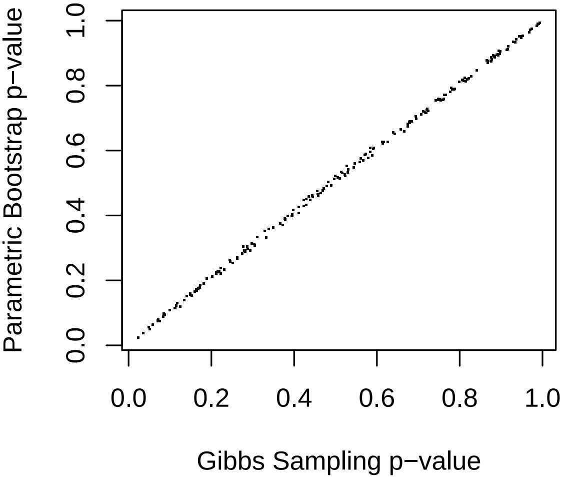
Skip this slide. Our method goes like this in iid sampling case:

- Start with original sample $\theta_1, \ldots, \theta_n$ and $S_n = S$.
- Compute the sufficient statistic S_3 for $\theta_1, \theta_2, \theta_3$.
- Compute conditional density of θ_3 given S_3: joint over marginal.
- Doesn't depend on von Mises parameter! So do uniform case!
- Joint – easy by change of variables. Write S_3 in polar co-ordinates R, Θ.
- Marginal: Angle Θ is uniform and independent of R.
- Marginal: Stephens (1962) uses elliptic integrals to get density of R.

Comparison with other methods

- Tried to show conditional tests better than parametric bootstrap.
- Generate many data sets. For each data set:
- Run MCMC to compute conditional p-value.
- Do parametric bootstrap: estimate parameters by ml; compute new test statistic; compare observed test statistic to simulations to get unconditional p-value.
- Plot two p values against each other.

P-value comparisons



The kicker

- Two methods produce nearly equal P-values.
- So nearly equal levels and powers.
- This is a theorem.
- Starting with work of Lars Holst (Ann Prob).
- Conditional law of data given S is singular wrt unconditional law.
- But: conditional dist of gof tests asymptotically same as unconditional.

Conclusions reiterated, rephrased, augmented, contradicted? 1

- Obviously no time to discuss actual data.
- Can maximize average power.
- Choice of “average” – (approximate) Gaussian process on alternative.

Conclusions reiterated, rephrased, augmented, contradicted? 1

- Obviously no time to discuss actual data.
- Can maximize average power.
- Choice of “average” – (approximate) Gaussian process on alternative.
- Best unbiased tests are conditional tests.

Conclusions reiterated, rephrased, augmented, contradicted? 1

- Obviously no time to discuss actual data.
- Can maximize average power.
- Choice of “average” – (approximate) Gaussian process on alternative.
- Best unbiased tests are conditional tests.
- Implementation via Markov Chain Monte Carlo.

Conclusions reiterated, rephrased, augmented, contradicted? 1

- Obviously no time to discuss actual data.
- Can maximize average power.
- Choice of “average” – (approximate) Gaussian process on alternative.
- Best unbiased tests are conditional tests.
- Implementation via Markov Chain Monte Carlo.
- The parametric bootstrap nearly implements conditional tests.

Thanks

Thanks!