David L. Dowe [© Jan. 2007]
(www.csse.monash.edu.au/∼dld ; dld At bruce.csse.monash.edu.au

MML and statistically consistent invariant (objective?) Bayesian probabilistic inference

Statistical invariance

Statistical consistency

– Fixed number of parameters
– Amount of data per parameter bounded above
  - Neyman-Scott problem

Statistical likelihood function
Inference: Maximum likelihood, etc.

Evidence-based medicine

- Statistical inference
- Machine learning
- Econometrics
- Inductive inference
- "Data mining"

Inference

One model (typically)

Prediction

Possibly more than one model

Models can be averaged

- non-weighted (equal weights), or
- weighted (different weights)

Easy problems

– Known likelihood function $f(D|H)$, $Prob(Data|Hypothesis)$, $f(\boldsymbol{x}|\boldsymbol{\theta})$

– Fixed number of parameters
Amount of data per parameter unbounded

– Little noise

Intermediate problems ...

Hard(er) **problems**

– (Unknown likelihood function)

– Much noise

– Amount of data per parameter bounded above - e.g.,
  - Neyman-Scott problem (with known likelihood function)

Desiderata (in inference)

Statistical invariance

– Circle: $\hat{A} = \pi\ \hat{r}^2$
– Cube: $\hat{l} \ = \ \hat{A}^{1/2} \ = \ \hat{V}^{1/3}$
– Cartesian/Polar: $(\hat{x}, \hat{y}) = (\hat{r}\cos(\hat{\theta}), \hat{r}\sin(\hat{\theta}))$

Statistical consistency

As we get more and more data, we converge more and more closely to the true underlying model
(But what if data-generating source is outside our model space?)

Efficiency

Not only are we statistically consistent, but as we get more and more data we converge as rapidly as is possible to any underlying model.

Some methods of inference

Maximum Likelihood: Given data D, choose (probabilistic) hypothesis H to maximise $f(D|H)$ and minimise $-\log f(D|H)$.

– Statistically invariant – but tends to over-fit, "finding" non-existent patterns in random noise

– Also, how do we choose between models of increasing complexity and increasingly good fit e.g., constant, linear, quadratic, cubic, ...?

– Also, maximum likelihood chooses the hypothesis to make the already observed data as likely as possible.

  But, shouldn't we choose H so as to maximise $Pr(H|D)$?

Bayesianism, prior prob's, $Pr(H|D)$
Prior probability, $Pr(H)$

$Pr(H).Pr(D|H) = Pr(H\&D) = Pr(D\&H) = Pr(D).Pr(H|D)$

So, $Pr(H|D) = \frac{Pr(H).Pr(D|H)}{Pr(D)} = \frac{1}{Pr(D)}(Pr(H).Pr(D|H))$

$$posterior(H|D) = \frac{prior(H)\ .\ likelihood(D|H)}{marginal(D)}$$

Probability vs probability *density*

What is your (friend's) height? weight? *Measurement accuracy* - used in MML in lower bound for some parameter estimates, but overlooked and ignored in classical approaches

Information Theory

Given data D already observed,

$max_H \; Pr(H|D) =$

$max_H \frac{1}{Pr(D)}(Pr(H).Pr(D|H)) =$

$max_H \; Pr(H).Pr(D|H) =$

$min_H \; -\log Pr(H) \; -\log Pr(D|H)$

Can do this if everything is a probability and not a density, whereupon $l_i = -\log_2 p_i$ is the binary code-length of an event of probability p_i

$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{21}$
$\frac{1}{4}$	$\frac{1}{4}$	$\frac{2}{21}$
$\frac{1}{4}$	$\frac{1}{4}$	$\frac{3}{21}$
$\frac{1}{8}$	$\frac{1}{4}$	$\frac{6}{21}$
$\frac{1}{16}$		$\frac{4}{21}$
$\frac{1}{16}$		$\frac{5}{21}$

Bayesian **Maximum A Posteriori** (*MAP*) maximises prior *density* multiplied by likelihood
This is not statistically invariant.
It also suffers the inconsistency and other problems of Max Likelihood.

Minimum Message Length (**MML**) is statistically invariant and has general statistical consistency properties (which Maximum Likelihood and Akaike's Information Criterion (AIC) don't have).

– MML is also far more efficient than Maximum Likelihood and AIC
– MML is always defined, whereas for some problems AIC is either undefined or poor

Turing Machine

$f : States \times Symbols \rightarrow \{L, R\} \cup Symbols.$

With binary alphabet,
$f : States \times \{0, 1\} \rightarrow \{L, R\} \cup \{0, 1\}.$

Any known computer program can be represented by a Turing Machine.

Universal Turing Machines (UTMs) are like a compiler and can be made to emulate *any* Turing Machine (TM).

Recalling from information theory that an event of probability p_i can be encoded by a binary code-word of length $l_i = \log_2 p_i$, and recalling from MML that choosing H to maximise $Pr(H|D)$ is equivalent to choosing H to minimise the length of a two-part message,

$$-\log Pr(H) \quad -\log Pr(D|H),$$

```
 -------------------------------------------
|    H1   |            Data given H1        |
 -------------------------------------------

 ---------------------------------------------
|    H2        |      Data given H2           |
 ---------------------------------------------
```

we can see the relationship between MML, (probabilistic) Turing machines and (two-part) Kolmogorov complexity.

Kolmogorov complexity

The *Kolmogorov complexity* of a string, s, relative to some (Universal) Turing machine, U, is the length, $|l|$, of the shortest input l to U such that

$U(l) = s \quad$ and then U halts.

MML is Bayesian, and the choice of UTM is Bayesian.

But does this appeal to UTMs and Kolmogorov complexity give us a (fairly?) objective(?) Bayesianism?

In practice, use *approximations* to MML, typically quantising (rounding off) in parameter space:

Approximations to (Strict) MML
For *discrete* variables, relatively easy.

For *continuous* variables (note measurement accuracy):
MMLD [or I_{1D}] ({1999,} 2002, ...)
$min_R - \log(\int_R h(\boldsymbol{\theta})\ d\theta) - \frac{\int_R h(\boldsymbol{\theta}).\log f(\boldsymbol{x}|\boldsymbol{\theta})\ d\theta}{\int_R h(\boldsymbol{\theta})\ d\theta}$

Wallace-Freeman (J RoyStatSoc 1987)
$-\log(h(\boldsymbol{\theta}).\frac{1}{\sqrt{\kappa_D^D\ Fisher(\boldsymbol{\theta})}}) \quad - \log f(\boldsymbol{x}|\boldsymbol{\theta}) + \frac{D}{2}$

Example (slightly hybrid): Univariate Polynomial Regression (x known)
$y = (\Sigma_{i=0}^{d} a_i\ x^i) + N(0, \sigma^2)$
1^{st} part of message (hypothesis, H):
$\hat{d};\ \hat{a_0}, ..., \hat{a_d}, \hat{\sigma^2}$
2^{nd} part of message: $Data|H$.

Neyman-Scott problem (1948)
We measure N people's heights J times each (say $J = 2$) & then infer

- the heights $\mu_1, ..., \mu_N$ of each of the N people,
- the accuracy (σ) of the measuring instrument.

We have JN measurements from which we need to estimate $N + 1$ parameters. $JN/(N + 1) \leq J$, so the amount of data per parameter is bounded above (by J).

$\hat{\sigma}^2_{MaximumLikelihood} \rightarrow \frac{J-1}{J}\sigma^2$,
and so for fixed J as $N \rightarrow \infty$ Maximum Likelihood is statistically inconsistent - under-estimating σ and "finding" patterns that aren't there.

Variants on Neyman-Scott problem

What makes Neyman-Scott difficult is that the amount of data per parameter is bounded above.
This is awful for Maximum Likelihood and Akaike's Information Criterion (AIC).

Other examples include

- latent factor analysis
- fully-parameterised mixture modelling

By acknowledging **uncertainty** (or quantising) when doing parameter estimation, MML is statistically consistent on all of these problems.

MML is about *inference*, seeking the *truth*.

– It gives a statistically invariant - and statistically consistent - Bayesian method of point estimation.
– It gives general consistency results where classical non-Bayesian approaches are known to break down.
– It is also efficient, working well on all range of real inference problems.

Conjecture (1998, ...) that only MML and very closely-related Bayesian methods are in general both statistically consistent and invariant.
Back-up Conjecture: If there are any such non-Bayesian methods, they will be far less efficient than MML.

Some of MML's many "friends"

Scoring probabilistic predictions

MML and Efficient Markets Hypothesis: markets *not* provably efficient

MML, Kolmogorov complexity and measures of "intelligence"

MML and Econometric Time Series

MML, Entropy and Time's Arrow

MML and Linguistics - inferring "dead" languages

MML, cosmological arguments and "Intelligent Design" (I.D.)

MML in medicine, psych' & bio':

Amer. J. Psychiatry:

Kissane D.W., S. Bloch, D.L. Dowe, R.D. Snyder, P. Onghena, D.P. McKenzie and C.S. Wallace (1996a). The Melbourne Family Grief Study, I: Perceptions of family functioning in bereavement. *American Journal of Psychiatry*, 153, 650-658.

Kissane D.W., S. Bloch, P. Onghena, D.P. McKenzie, R.D. Snyder, D.L. Dowe (1996b). The Melbourne Family Grief Study, II: Psychosocial morbidity and grief in bereaved families. *American Journal of Psychiatry*, 153, 659-666.

Pilowsky, I., Levine, S., & Boulton, D.M. (1969). The classification of depression by numerical taxonomy. British Journal of Psychiatry, 115, 937-945.

Prior, R. Eisenmajer, S. Leekam, L. Wing, J. Gould, B. Ong and D. L. Dowe (1998). Are there subgroups within the autistic spectrum? A cluster analysis of a group of children with autistic spectrum disorders. J. Child Psychol. Psychiat. Vol. 39, No. 6, pp893-902

Clarke, D.M., G.C. Smith, D.L. Dowe and D.P. McKenzie (2003). An empirically-derived taxonomy of common distress syndromes in the medically ill. J. Psychosomatic Research 54 (2003) pp323-330.

Edgoose, T., L. Allison and D. L. Dowe (1998). An MML Classification of Protein Structure that knows about Angles and Sequences, pp585-596, Proc. 3rd Pacific Symposium on Biocomputing (PSB-98), Hawaii, U.S.A., January 1998.

etc., ..., etc. etc., ..., etc.

Reading (on general MML):

– Wallace, C.S. and D.L. Dowe (1999a). "Minimum Message Length and Kolmogorov Complexity", Computer Journal, Vol. 42, No. 4, pp270-283
[As of May 2005, this has been the *Computer Journal*'s most downloaded article.]
– Wallace, C.S. (2005) [posthumous], "Statistical and Inductive Inference by Minimum Message Length", Springer (Series: Information Science and Statistics), 2005, XVI, 432 pp., 22 illus., ISBN: 0-387-23795-X
– Dowe, D.L., S. Gardner and G.R. Oppy (2007+). "Bayes not Bust! Why Simplicity is no problem for Bayesians", accepted (Thu 29/6/2006) to - and forthcoming in - British Journal for the Philosophy of Science (BJPS).
– Dowe, D.L. and G. Oppy (2001). "Universal Bayesian inference?". Behavioral and Brain Sciences [special issue re R. Shepard], Vol 24, No. 4, Aug 2001, pp662-663.
– Comley, J. W. and D.L. Dowe (2005). "Minimum Message Length and Generalized Bayesian Networks with Asymmetric Languages", Chapter 11 (pp265-294) in P. Gru:nwald, I. J. Myung & M. Pitt (eds.), Advances in Minimum Description Length: Theory and Applications, MIT Press, April 2005, ISBN 0-262-07262-9. [Final camera-ready copy submitted October 2003.]
  - {See also Comley, J. W. and D.L. Dowe (June 2003). "General Bayesian Networks and Asymmetric Languages", Proc. 2nd Hawaii International Conference on Statistics and Related Fields, 5-8 June, 2003.}
– Wallace, C. S. and D. M. Boulton (1968), "An information measure for classification", Computer Journal, Vol. 11, No. 2, August 1968, pp185-194.