Penalized likelihood*

Peter J. Green†

University of Bristol, UK.

January 7, 1997

Penalized likelihood is discussed in two rather different settings: function estimation and model choice. The connection is that, in both cases, a penalty term is added to the log-likelihood of the data, before it is used in some standard inferential criterion, in order to circumvent difficulties related to dimensionality. In function estimation, the problem is that the object of inference is infinite-dimensional, while in model choice, it is a general difficulty with comparison between models of differing dimension.

Penalized maximum likelihood estimation in nonparametric regression and density estimation were reviewed by Silverman [20] in volume 6 of the *Encyclopaedia*; this is an update and extension of that article; there is additional material especially in respect of aspects of penalized likelihood in nonlinear regression of various kinds, in inverse problems, and in model choice.

*Invited entry for the Encyclopaedia of Statistical Sciences, update volume.

†Department of Mathematics, University of Bristol, Bristol BS8 1TW, UK.

Email: P.J.Green@bristol.ac.uk.

1 Penalized likelihood in function estimation

In statistical inference about infinite-dimensional objects such as functions of one or more continuous variables, the principle of maximum likelihood is usually inadequate as a basis for estimation. The function in question may be, for example, a probability density or hazard function, a regression relationship, or the intensity function of a point process, so problems of this kind occur very widely indeed. Attempting to maximise the likelihood in such contexts usually results in an infinite value for the likelihood, and/or to degeneracy or lack of identifiability for the unknown function itself. In parametric approaches to inference, the inherent impossibility of identifying an infinite-dimensional object from a finite amount of data is avoided, of course, by assuming that the function has a particular parametric form (a normal density, a quadratic regression curve, etc.), so that inference about only a finite number of parameters is involved. Very occasionally, there are other ways round the difficulties posed by dimension, as in the partial likelihood approach to Cox's proportional hazards model. Generally, however, it is necessary to confront the nonparametric nature of the problem explicitly, and penalized likelihood provides a natural and unifying approach to doing so.

The failure of maximum likelihood as a principle for estimation of functions really derives from an inadequacy in the formulation of the inferential problem. No matter how successfully the likelihood function captures the relationship between unknown function and data, it cannot fully represent all that is known about the function, for it is impossible to imagine a scenario in which no other information is available, however informally, from the context. Bayesian inference provides one approach to the explicit acknowledgement of this prior information, and its integration into the process of inference; penalized likelihood provides an alternative, non-Bayesian, paradigm with a similar aim. Formal connections between the two approaches will be discussed in Section 1.5, later in this article.

The scope for the application of penalized likelihood is greatest in nonparametric and semiparametric regression, interpreting the term very broadly, and such applications will be emphasised here. A brief discussion of application to density estimation will also be given.

The emphasis in this article is on methodology, not theory; for careful and illuminating accounts of the asymptotic theory of penalized likelihood estimators, we refer the reader to Cox and O'Sullivan [3], and Gu and Qiu [10].

1.1 Nonparametric regression

The most basic situation both conveys the essential ideas and provides the key ingredient in the solution of more complicated problems. Suppose that responses $\{Y_i : i = 1, 2, \ldots, n\}$ are available, corresponding to recorded values (which we call design points) $t_1 \leq t_2 \ldots \leq t_n$ of a univariate explanatory variable t. As usual in regression we suppose that Y but not t is subject to error, and here we consider the model

$$Y_i = g(t_i) + \varepsilon_i \qquad \text{for } i = 1, 2, \ldots, n \tag{1}$$

where the errors ε_i have prescribed statistical properties (let us say, independently normally distributed with zero mean and equal variance σ^2, for definiteness), and the function g is to be estimated.

Under this model, and assuming the design points are distinct, it is easy to see that the likelihood is equally maximized by any function g that interpolates the data points (t_i, Y_i). (If there are tied design points, the optimising g's interpolate the mean response value at each distinct design point). Maximum likelihood has therefore failed to generate useful methodology – it has not produced a unique estimate, and in any case, typically, all such interpolating functions g are implausibly *rough* as explanations for the true dependence of Y on t, that is, they provide descriptions of the functional dependence that are too complicated, varying too rapidly.

The penalized likelihood solution to this difficulty replaces the log-likelihood for the problem

$$L(g) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - g(t_i))^2$$

by a new objective function

$$L(g) - \tfrac{1}{2}\lambda J(g) \tag{2}$$

where $J(g)$ is a non-negative real-valued functional measuring the "roughness" of g in a manner appropriate to the analyst's understanding of the problem. Small values of $J(g)$ correspond to a function g that is smooth (or simple): large values to one that is rough (or complicated); examples of such g can be seen in Figure 1. Thus, when this penalized

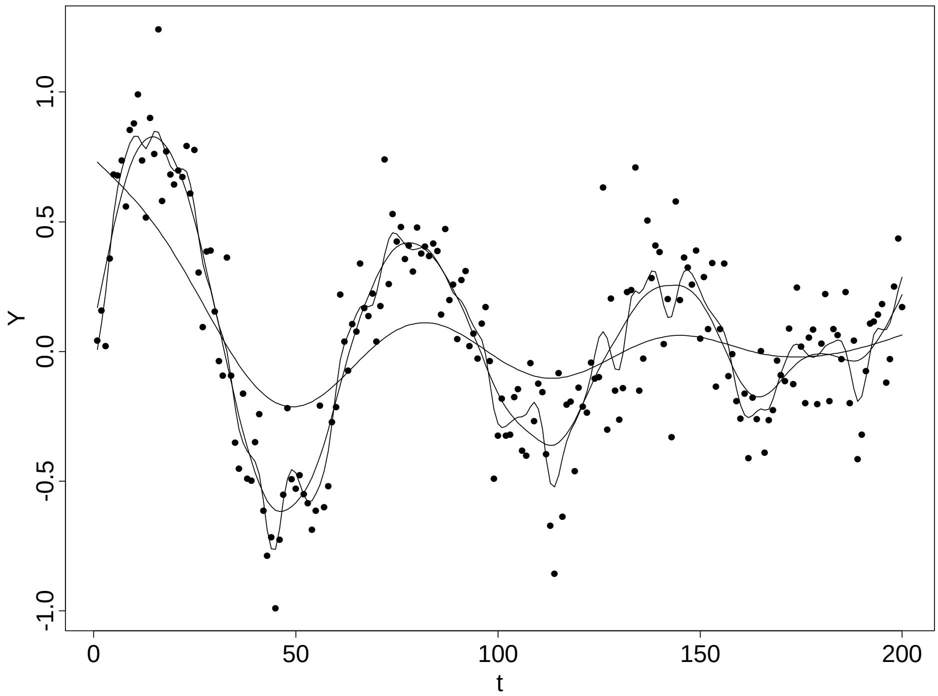

Figure 1: Penalized likelihood estimates for an artificial data set, with 3 different choices for the smoothing parameter λ, differing by factors of 100.

log-likelihood is maximized over g there will be a trade-off between fidelity to the data (large values of $L(g)$) and smoothness of the function (small values of $J(g)$). The smoothing parameter λ controls the relative importance of the two criteria (the $\frac{1}{2}$ factor is just for algebraic convenience).

Maximum likelihood estimation of a parametric version of the problem corresponds formally to the choice of $J(g)$ that gives the value zero whenever g is in the corresponding parameterized class of functions and infinity otherwise, so that we are effectively maximising $L(g)$ under a constraint. More generally, with penalized likelihood we relax this "hard" constraint by choosing $J(g)$ that is positive but finite for a suitable class of functions. The case of hard constraints is recovered formally by letting $\lambda \to \infty$.

A common choice of roughness functional $J(g)$ for nonparametric regression on one variable is the integrated squared second derivative

$$J(g) = \int_{-\infty}^{+\infty} g''(t)^2 dt \tag{3}$$

which has an appealing interpretation as being a global measure of curvature for g as well as physical motivation in terms of the deformation energy of a thin rod. It takes the value zero if and only if g is a first degree polynomial, so letting $\lambda \to \infty$ as described above reduces to simple linear regression of Y on t. Finally, its quadratic nature, combined with the quadratic log-likelihood resulting from the normally distributed errors, leads to appealing mathematical properties that in turn mean convenience in computing. Optimisation of a quadratic criterion is evidently going to yield linear estimates. More explicitly, it can be shown that the solution is the unique natural cubic spline interpolating points $\{(t_i, a_i), i = 1, \ldots, n\}$ where the vectors with components $\{a_i\}$ and $\{Y_i\}$ are related by the linear equations

$$(I + \lambda\sigma^2 K)\boldsymbol{a} = \boldsymbol{Y}. \tag{4}$$

Here K is a positive semi-definite matrix depending only on $\{t_i\}$ which possesses representations in terms of banded matrices that can be exploited for economical computation; this was discovered by Reinsch [16]; see also Green and Silverman [9] (§2.3.3). (A cubic spline is a function composed of pieces that are cubic polynomials, joined so that the result is smooth enough to have a continuous second derivative throughout; it is "natural" if the first and last pieces are linear functions. It is a standard result that there is a unique such curve interpolating any set of points $\{(t_i, a_i), i = 1, \ldots, n\}$ where the $\{t_i\}$ are distinct.)

In Figure 1, penalized likelihood estimates of g based on an artificial data set are displayed for various values of λ; they provide explanations for the data that on the basis of visual assessment might be considered too smooth, too rough, and just about right.

1.2 More general regression problems

Of course, many alternative methods are available for the problem of estimating a smooth regression curve of one variable: kernel methods, orthogonal series, local polynomials, etc. Among these the penalized likelihood solution we have described has little to distinguish it, except for its convenient characterisation through its optimisation of an explicit criterion, rather than being defined solely through the operations used to compute it.

However, the real strength of the penalized likelihood approach is seen when we come to address more general regression problems: the generalisations are trivial to define, straightforward to derive, and usually routine to compute. Without giving full details, which can all be found in Green and Silverman [9], the following list of generalisations gives a flavour of this process.

1.2.1 Unequally weighted data

If the errors ε_i have unequal variances σ^2/w_i, or there is some other reason to weight the responses unequally according to $\{w_i\}$, the only change to the estimation procedure described above is that (4) becomes

$$(W + \lambda\sigma^2 K)\boldsymbol{a} = W\boldsymbol{Y} \tag{5}$$

where W is the diagonal matrix with entries $\{w_i\}$. The computational details are hardly altered.

1.2.2 Other roughness functionals

Depending on context, it may be appropriate to replace the second derivative in defining $J(g)$ by some other order of derivative, say the mth. This corresponds to a different assumption about what is meant by "smooth": the resulting estimator is a polynomial spline of order $(2m-1)$, and is characterised in a very similar way to (1).

1.2.3 Semi-parametric regression

If the regression model

$$Y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + g(t_i) + \varepsilon_i$$

is appropriate, corresponding to parametric dependence on the explanatory variables $\boldsymbol{x}_i$, and nonparametric dependence on t_i, then the solution for g is still a natural cubic spline. The maximum penalized likelihood estimators for $\boldsymbol{\beta}$ and g can be obtained by alternately fitting a linear model and a cubic smoothing spline, and iterating to convergence, a process known as "backfitting" (see, for example, Hastie and Tibshirani [12]).

1.2.4 Additive models

Regression models in which two or more explanatory variables are treated nonparametrically, such as

$$Y_i = g(t_i) + h(u_i) + \varepsilon_i$$

can be handled by penalising the log-likelihood with two functionals, so that the function to be maximized is, say

$$L(g,h) - \tfrac{1}{2}\lambda_1 J_1(g) - \tfrac{1}{2}\lambda_2 J_2(h)$$

Again, the estimates can be computed by backfitting, in this case by alternately updating estimates of g and h using spline smoothing. The monograph by Hastie and Tibshirani [12] covers many further aspects of this style of modelling in both linear and nonlinear models.

1.2.5 Generalized linear models

In the class of generalized linear models (Nelder and Wedderburn, [14]), the normal/linear assumptions underlying linear models are relaxed. The responses $\{Y_i\}$ are supposed to be drawn from a one-parameter exponential family of distributions, with density

$$p(y_i;\theta_i,\phi) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i,\phi)\right)$$

where ϕ is an overall dispersion parameter, $b(\cdot)$ and $c(\cdot)$ are specified functions characterising the distribution. The natural parameter θ_i is a function of a parameter vector $\boldsymbol{\beta}$ and covariates $\boldsymbol{x}_i$ specific to the ith case, such that

$$G(\mu_i) = \boldsymbol{x}_i^T\boldsymbol{\beta} \quad \text{where} \quad \mu_i = E(Y_i) = b'(\theta_i), \tag{6}$$

where $G(\cdot)$ is a specified *link function.* This class covers many standard techniques such as logit and probit analysis, and log-linear models. The principles of modelling and estimation in a nonparametric or semiparametric context apply still, with few changes. For example, if (6) is replaced by

$$G(\mu_i) = \boldsymbol{x}_i^T\boldsymbol{\beta} + g(t_i)$$

where $(\boldsymbol{x}_i, t_i)$ are covariates, then g and $\boldsymbol{\beta}$ can be estimated by backfitting, as in section 1.2.3 above, within an outer loop that corresponds to Fisher's scoring algorithm. These computations use working weights and responses, in place of W and $\boldsymbol{Y}$ in (5), given in terms of the information matrix and score vector.

1.2.6 Quantile regression

One approach to the estimation of the whole distribution of a response Y as a function of a one-dimensional covariate t is the LMS method of Cole [2]. This provides flexibility and a nonparametric approach by assuming that Y is normally distributed, *after a Box-Cox power transformation.* The power, mean and spread all vary smoothly with t. Thus we specify

$$\frac{\{Y/\mu(t)\}^{\lambda(t)} - 1}{\lambda(t)\sigma(t)} \sim N(0,1)$$

(defined by continuity at $\lambda(t) = 0$), independently for each observation pair (t,Y). There are various methods for estimating the three smooth functions $\lambda(t)$, $\mu(t)$ and $\sigma(t)$. Using penalized likelihood means maximising

$$L(\lambda, \mu, \sigma) - \tfrac{1}{2}\alpha_\lambda J(\lambda) - \tfrac{1}{2}\alpha_\mu J(\mu) - \tfrac{1}{2}\alpha_\sigma J(\sigma)$$

where L is the log-likelihood derived from the normal assumption and the power transformation, and the roughness functional J would usually be taken to be that in (3). This has the advantage of yielding simultaneously smooth estimates of every quantile of Y given t that do not intersect, avoiding arbitrary grouping of the data, involving choice of only three smoothing parameters, and being relatively cheap to compute.

1.3 Inverse problems: regularisation and maximum entropy

In many areas of science and engineering, often involving aspects of signal or image processing, ill-posed inverse problems of the following kind occur. The description of the object in question is a high-dimensional vector $\boldsymbol{x}$, and this has to estimated from a vector of observations $\boldsymbol{Y}$ that is both noisy and indirect, so

$$\boldsymbol{Y} = K\boldsymbol{x} + \boldsymbol{\varepsilon} \tag{7}$$

Here $\boldsymbol{\varepsilon}$ represents noise, and K is some fixed known matrix, usually representing a discretized version of some mathematical transform. We can view this a modified form of equation (1), by taking the t_i there to be just the index i, $g(t_i)$ to be x_i, discretized from the start, and allowing now indirect estimation through the operator K, in place of the identity.

In practice, such problems are often ill-posed, in that the least squares solution to (7), if it is exists at all, is very sensitive to small perturbations in $\boldsymbol{Y}$: mathematically, K has few singular values differing substantially from 0.

A standard approach to solving the inverse problem that gets round this difficulty is least-squares regularisation: estimate $\boldsymbol{x}$ by minimising

$$||\boldsymbol{Y} - K\boldsymbol{x}||^2 + \alpha||\boldsymbol{x}||^2 \tag{8}$$

where $\|\cdot\|$ denotes the L_2 norm, and α is a smoothing parameter. There is an obvious analogue with our penalized likelihood objective function (2) for nonparametric regression, and it is possible to motivate regularisation from this perspective.

There is a huge amount of work on regularisation in the engineering literature. There is also a great deal of work on maximum entropy estimation, a method that is often discussed quite independently, but it is illuminating to discuss the two methods together.

Very often in applications of inverse problems, the components of the unknown $\boldsymbol{x}$ represent intensities or other qualities that are necessarily non-negative. In the simplest version of the method of maximum entropy (ME) for such problems (see, for example, Gull and Daniell, [11]), we replace the L_2-norm penalty in (8) by an entropy measure, obtaining

$$\|\boldsymbol{Y} - K\boldsymbol{x}\|^2 + 2\alpha \sum_i x_i \log x_i$$

as our objective function, to be minimized over $\boldsymbol{x}$. This is now not quadratic in $\boldsymbol{x}$, so that the solution is not linear in $\boldsymbol{Y}$, and must be obtained by numerical methods. Donoho, Johnstone, Hoch and Stern [5] provide an excellent analysis of the ME method from a statistical perspective, characterising the types of problems for which it can be expected to yield good performance, including signal-to-noise enhancement and superresolution.

There are also important variants of the inverse problem (7) that are nonlinear. For example, in emission tomography, $\boldsymbol{x}$ represents a vector of concentrations of a radioisotope, the components corresponding to voxels in a human body. The observations $\boldsymbol{Y}$ are counts of photons recorded in detectors placed around the patient (an exact description would depend on the particular equipment being used), and the objective is to estimate $\boldsymbol{x}$ from $\boldsymbol{Y}$. From basic physical principles of attenuation and superposition, there is a realistic probabilistic model stating that the components of $\boldsymbol{Y}$ are independent with Poisson distributions whose means are the corresponding components of $K\boldsymbol{x}$ for some suitable matrix K, which is ill-posed. Large Poisson regression problems of this kind, where $\boldsymbol{x}$ is to be estimated by maximum likelihood under non-negativity constraints, turn out to be a natural domain of application for the EM algorithm (Dempster, Laird and Rubin, [4]). For derivation of the approach for positron emission tomography, see Shepp and Vardi [18].

Because of the ill-posed character of K, maximum likelihood reconstruction from emission tomography data are not very satisfactory: high variance in the estimators is manifested in a rough, noisy-looking reconstruction. There has been interest in obtaining reconstructions that are smoother spatially, by either including an explicit smoothing step in an iterative reconstruction method (e.g. Silverman, Jones, Wilson and Nychka [22]) or by using penalized likelihood. See Green [7] for more background on single-photon emission tomography, showing penalized likelihood reconstructions, and Green [8] for a practical modification to the EM method that adapts it to maximum penalized likelihood estimation for quite a general range of roughness penalties. Those used in Green [7] are inspired by the Gibbs distributions of statistical physics, an example of the resulting penalized log likelihood being

$$\sum_i (Y_i \log(K\boldsymbol{x})_i - (K\boldsymbol{x})_i) - \tfrac{1}{2}\lambda \sum_{j\sim k} \log\cosh\left(\frac{x_j - x_k}{\delta}\right)$$

where the second summation is over physically adjacent pairs of voxels j, k, and δ is a scale constant.

1.4 Density estimation

Nonparametric density estimation is somewhat different in character from nonparametric regression, but some of the same considerations apply. The basic setup is that we wish to estimate a probability density function (pdf) g, given n independent and identically distributed observations $Y_1, Y_2, \ldots, Y_n$. The log-likelihood is of course just

$$L(g) = \sum_{i=1}^{n} \log g(Y_i),$$

and if we maximise this subject to the constraints of a parametric form for g, then sensible estimates generally arise. However, without such constraints, there are difficulties with maximum likelihood. For, consider the step function estimate of g given by

$$\widehat{g_m}(y) = \frac{m}{n} \sum_{i=1}^{n} I[|y - Y_i| < 1/(2m)]$$

where $I[\cdot]$ denotes the indicator function; this is evidently a proper pdf, but for all m sufficiently large that no pair of observations are closer than $1/(2m)$, we have $L(\widehat{g_m}) = n\log(m/n)$, which goes to infinity with m. Thus the maximum likelihood "estimator" degenerates into a set of spikes at the data points.

This is the analogue for density estimation of the "interpolation" problem with unconstrained maximum likelihood estimation of a regression function. There are many routes to density estimation avoiding this problem (see **Density estimation** article by H. G. Muller, in this volume, and the comprehensive but accessible book by Silverman [21]), including kernel methods, spectral methods, smoothed histograms, and orthogonal series estimates. But in the spirit of the appealing methodology obtained by penalized likelihood in regression, it is natural to consider such an approach for the present problem.

This apparently originates with Good and Gaskins [6], who replaced the log-likelihood by

$$L(g) - \tfrac{1}{2}\lambda J(g) = \sum_{i=1}^{n} \log g(Y_i) - \tfrac{1}{2}\lambda J(g)$$

for some roughness penalty $J(g)$ and smoothing parameter λ. They proposed

$$J(g) = \int g'^2/g = 4\int (\sqrt{g})'^2, \tag{9}$$

while Silverman [19] considers

$$J(g) = \int (\log g)'''^2 \tag{10}$$

and the integrated squared second derivative penalty from (3) might also be used.

As with nonparametric regression, establishing theoretical properties of MPLE's in the density estimation poses mathematical difficulties because of the implicit definition of the estimates. In fact, here the difficulties are more formidable because of the lack of linearity (and there is no explicit representation comparable with (4)), and the constraint that the estimate be a pdf. In contrast to nonparametric regression, here there are not even convenient practical methods of computation. Available methods provide approximations: Good and Gaskins [6] use orthogonal series methods, and Tapia and Thompson [23] discretize the problem, including the roughness penalty. It is therefore not surprising that the penalized likelihood paradigm has made much less impact on practical methodology in density estimation than it has in nonparametric regression.

1.5 The Bayesian perspective on penalized likelihood estimation

If is sometimes helpful to give the basic notion of penalized likelihood estimation a Bayesian interpretation.

Suppose we write the likelihood of the data $\boldsymbol{Y}$ given an unknown g as

$$p(\boldsymbol{Y}|g) \propto e^{L(g)}$$

and place a prior distribution on g of the form

$$p(g) \propto e^{-\frac{1}{2}\lambda J(g)} \tag{11}$$

than by Bayes' theorem

$$p(g|\boldsymbol{Y}) \propto e^{L(g)-\frac{1}{2}\lambda J(g)},$$

so that the maximum penalized likelihood estimator of g, maximising (2), is identified as the mode of its posterior distribution, or *maximum a posteriori* estimator.

This connection should not be surprising: the underlying notion and justification of penalized likelihood is very close to that of specifying prior belief or knowledge about g, albeit in a non-probabilistic form. Beyond this qualitative argument, the extent to which the connection is useful is limited by two factors:

(a) whether (11) is plausible in probabilistic terms as an expression of prior belief, for roughness functionals $J(g)$ of interest, and

(b) whether the *maximum a posteriori* criterion is appealing in the context.

Regarding (a), there are certainly difficulties. It is commonly the case the $J(g)$ is invariant to the addition of a constant, or other transformations of g, so that $p(g)$ is an improper distribution. More precisely, is it *partially improper*; for example, conditioning on the values of g at a few distinct points yields a proper distribution. But further, because of the infinite dimensions involved, there are apparent paradoxes. For example, in the case of cubic smoothing splines, and elsewhere, it turns out that the prior and posterior distributions of g are entirely concentrated *outside* the space of smooth functions over which the optimum (the cubic smoothing spline) is sought.

Regarding (b), there are general grounds for suspicion about using the mode as a summary of the distribution of a high-dimensional object. These issues are all explored further in Green and Silverman [9] (§3.8.3).

2 Penalized likelihood in model choice

The classical approach to model comparison operates in a nested model situation, and uses the generalized likelihood ratio test. If we have two models M_1 and M_2, with M_2 nested within M_1, then the likelihood ratio test rejects the simpler model M_2 if

$$-2 \log \Lambda$$

exceeds an appropriate percentile of the $\chi^2_{p_1 - p_2}$ distribution, where $\Lambda = L_1/L_2$, L_i denotes the maximized likelihood under model M_i, and p_i is the dimension of that model.

It has long been recognised that, even when the simpler model M_2 is true, the more complex model is favoured if the amount of data is large (Lindley's paradox; see [13]) Various approaches have been explored to counteract this tendency, and these amount to penalizing the likelihood ratio criterion, or, what is usually equivalent, comparing models on the basis of penalized log-likelihoods.

Akaike's criterion [1] uses

$$-2 \log \Lambda + 2(p_1 - p_2),$$

the Bayesian criterion of Schwarz [17] is

$$-2 \log \Lambda + (p_1 - p_2) \log n,$$

when the likelihoods are each based on n independent observations, while Mallows' C_p criterion amounts to using

$$-2\log\Lambda + (p_1 - p_2).$$

All of these criteria for model comparison give greater weight to the simpler model than does the unmodified likelihood ratio statistic, thus tilting the choice in favour of parsimony. Each of these also finds applications to the comparison of models M_1 and M_2 that are not nested.

Of the three criteria, that of Schwarz evidently involves the most radical perturbation to the likelihood ratio test. The mathematical basis for it is an asymptotic expansion of the Bayes factor in favour of M_1 against M_2. As a result, model choice based on this sets the two competing models on an equal footing, in contrast to the classical test, which focuses on detectability of departures from the simpler model, and hence has an interpretation and performance highly dependent on sample size. See O'Hagan ([15], chapter 7) for a thorough discussion of these issues.

References

[1] Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory*, 267–281. Akademia Kiado, Budapest. (The original appearance of Akaike's information criterion.)

[2] Cole, T.J. (1988) Fitting smoothed centile curves to reference data (with discussion). *Journal of the Royal Statistical Society, A*, **151**, 385–418. (Introduces the LMS method for quantile regression, using ad-hoc methods of curve estimation.)

[3] Cox, D. D. and O'Sullivan, F. (1990) Asymptotic analysis of penalized likelihood and related estimators. *Annals of Statistics*, **18**, 1676–1695.

[4] Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion).*Journal of the Royal Statistical*

Society, B, **39**, 1–37. (Seminal paper introducing the EM algorithm for indirect estimation.)

[5] Donoho, D. L., Johnstone, I. M., Hoch, J. C. and Stern, A. S. (1992) Maximum entropy and the nearly black object (with discussion). *Journal of the Royal Statistical Society, B*, **54**, 41–81. (Analysis of why and when maximum entropy performs well as a method of regularisation.)

[6] Good, I. J. and Gaskins, R. A. (1971) Nonparametric roughness penalties for probability densities. *Biometrika*, **58**, 255–277.

[7] Green, P. J. (1990) Bayesian reconstructions from emission tomography data using a modified EM algorithm. *IEEE Transactions on Medical Imaging,* **9**, 84–93. (Statistical modelling perspective on SPECT data, and a practical EM-like algorithm for penalized likelihood/Bayesian estimation in a large Poisson regression model.)

[8] Green, P. J. (1990) On use of the EM algorithm for penalized likelihood estimation. *Journal of the Royal Statistical Society, B*, **52**, 443–452. (Derives the OSL (one-step-late) algorithm for penalized likelihood estimation in incomplete-data problems, and discusses convergence rates.)

[9] Green, P. J. and Silverman, B. W. (1994) *Nonparametric regression and generalized linear models: a roughness penalty approach.* Chapman and Hall, London. (The penalized likelihood approach to nonparametric and semiparametric regression; mathematical derivations are based on simple linear algebra and calculus; stresses flexibility of approach.)

[10] Gu, C. and Qiu, C. (1993) Smoothing spline density estimation: theory. *Annals of Statistics*, **21**, 217–234.

[11] Gull, S. F. and Daniell, G. J. (1978) Image reconstruction from incomplete and noisy

data. *Nature*, **272**, 686–690. (Early article promoting maximum entropy approach to function estimation.)

[12] Hastie, T. and Tibshirani, R. J. (1990) *Generalized additive models.*. Chapman and Hall, London. (Monograph providing comprehensive coverage of nonparametric additive regression modelling, both linear and nonlinear.)

[13] Lindley, D. V. (1957) A statistical paradox. *Biometrika*, **44**, 187–192.

[14] Nelder, J. A. and Wedderburn, R. W. M. (1972) Generalized linear models. *Journal of the Royal Statistical Society, A*, **135**, 370–384. (The original paper identifying the class of generalized linear models, with algorithms and inference procedures.)

[15] O'Hagan, A. (1994) *Kendall's advanced theory of statistics, volume 2B: Bayesian inference.* Edward Arnold, London. (Comprehensive coverage of Bayesian inference.)

[16] Reinsch, C. (1967) Smoothing by spline functions. *Numerische Mathematik*, **10**, 177-183. (Early research paper on smoothing splines, including efficient algorithm.)

[17] Schwarz, G. (1978) Estimating the dimension of a model. *Annals of Statistics*, **6**, 464–464. (The Bayesian analogue of Akaike's information criterion.)

[18] Shepp. L. A. and Vardi, Y. (1982) Maximum likelihood reconstructions in positron emission tomography. *IEEE Transactions on Medical Imaging,* **1**, 113-122. (Statistical modelling perspective on PET data, and an EM algorithm for maximum likelihood estimation in a large Poisson regression model.)

[19] Silverman, B. W. (1982) On the estimation of a probability density function by the maximum penalized likelihood method. *Annals of Statistics*, **10**, 795–810. (On the theory of MPLE in density estimation and nonparametric regression, with many references.)

[20] Silverman, B. W. (1985) Penalized maximum likelihood estimation. *Encyclopaedia of Statistical Sciences*, **6**, 664–667. (Maximum penalized likelihood in density estimation and nonparametric regression.)

[21] Silverman, B. W. (1986) *Density estimation for statistics and data analysis.* Chapman and Hall, London. (Comprehensive guide to density estimation, by penalized likelihood and other methods.)

[22] Silverman, B. W., Jones, M. C., Wilson, J. D. and Nychka, D. W. (1990) A smoothed EM approach to indirect estimation problems, with particular reference to stereology and emission tomography. *Journal of the Royal Statistical Society, B*, **52**, 271–324. (Deals with a wide range of ill-posed indirect estimation problems, using a smoothed version of the EM algorithm.)

[23] Tapia, R. A. and Thompson, J. R. (1978) *Nonparametric probability density estimation.* Johns Hopkins University Press, Baltimore. (Emphasises the penalized likelihood approach.)

List of ESS articles to cite as cross-references:

Akaike's criterion,

Bayesian inference,

Density estimation,

Generalized linear models,

Maximum entropy,

Nonparametric regression,

Regression variables, selection of,

Spline functions.