Contribution to discussion of paper by
Spiegelhalter, Best, Carlin and van der Linde
RSS Ordinary meeting, 14 March 2002

Peter Green (University of Bristol)

I have two rather simple comments on this interesting, important and long-awaited paper.

The first concerns using basic distribution theory to give a surprising new perspective on p_D in the normal case, perhaps identifying a missed opportunity in exposition.

Consider first a decomposition of data as focus plus noise:

$$Y = X + Z$$

where X and Z are independent n-vectors, normally distributed with fixed means and variances, and $\mathrm{var}(Z)$ is non-singular. The deviance is

$$D(X) = (Y - X)^T \{\mathrm{var}(Z)\}^{-1}(Y - X)$$

and so

$$p_D = E[D(X)|Y] - D(E[X|Y]) = \mathrm{tr}\{\mathrm{var}(Z)^{-1}\mathrm{var}(Z|Y)\}, \tag{1}$$

using the standard expression for the expectation of a quadratic form. Several results in the paper have this form, possibly in disguise. However, $\mathrm{var}(Z|Y) = \mathrm{var}(Z) - \mathrm{cov}(Z,Y)\mathrm{var}(Y)^{-1}\mathrm{cov}(Y,Z) = \mathrm{var}(Z) - \mathrm{var}(Z)\mathrm{var}(Y)^{-1}\mathrm{var}(Z) = \mathrm{var}(Z)\mathrm{var}(Y)^{-1}\{\mathrm{var}(Y) - \mathrm{var}(Z)\}$, yielding the much more easily interpretable

$$p_D = \mathrm{tr}\{\mathrm{var}(Y)^{-1}\mathrm{var}(X)\}. \tag{2}$$

This allows a very clean derivation of examples in Sections 2.5 and 4.1–4.3. For example, in the Lindley and Smith model we have $\mathrm{var}(Z) = C_1$, $\mathrm{var}(X) = A_1 C_2 A_1^T$, and so

$$p_D = \mathrm{tr}\{(A_1 C_2 A_1^T + C_1)^{-1} A_1 C_2 A_1^T\} = \mathrm{tr}\{A_1^T C_1^{-1} A_1 (A_1^T C_1^{-1} A_1 + C_2^{-1})^{-1}\},$$

as in (21) of the paper.

Turning now to hierarchical models, consider a decomposition into k independent terms

$$Y = Z_1 + Z_2 + \cdots + Z_k,$$

where all Z_i are normal, and $\mathrm{var}(Z_k)$ is nonsingular. These represent all the various terms of the model: fixed effects with priors, random effects with different structures, errors at various levels; again all means and variances are fixed. Then for any level $\ell = 1, 2, \ldots, k-1$, we may take the sum of the first ℓ terms as the focus and the rest as noise.

Version (1) of p_D above is then not very promising:

$$p_D(\ell) = \mathrm{tr}\left\{ \mathrm{var}\left(\sum_{i=\ell+1}^{k} Z_i \right)^{-1} \mathrm{var}\left(\sum_{i=\ell+1}^{k} Z_i \,\middle|\, Y \right) \right\},$$

but (2) gives the more compelling

$$p_D(\ell) = \mathrm{tr}\left\{ \mathrm{var}(Y)^{-1} \mathrm{var}\left(\sum_{i=1}^{\ell} Z_i \right) \right\}. \tag{3}$$

Thus p_D has generated a decomposition of the overall degrees of freedom $n = \sum_\ell \mathrm{tr}\{\mathrm{var}(Y)^{-1}\mathrm{var}(Z_\ell)\}$ into non-negative terms attributable to the levels $\ell = 1, 2, \ldots, k$, just as in frequentist nested model

ANOVA. (One must take care with improper priors in using (3), and terms should be treated as limits as precisions go to 0.) Of course, (2) & (3) fail to hold with unknown variances or with non-normal models, but the observations above do provide further motivation for accepting p_D as a measure of complexity, and suggest exploring more thoroughly its role in hierarchical models.

My second point notes that the paper has no examples with discrete 'parameters'. Conditional distributions in hierarchical models with purely categorical variables can be computed using probability propagation methods (Lauritzen and Spiegelhalter (1988)), avoiding MCMC, so that p_D is again a cheap local computation. Presumably marginal posterior modes would be used for $\overline{\theta}$. Certainly this is a context where p_D can be negative. Can connections be drawn with existing model criticism criteria in probabilistic expert systems?

Additional reference

Lauritzen, S. L. and Spiegelhalter, D. J. (1988). Local computations with probabilities on graphical structures and their application to expert systems (with discussion). *Journal of the Royal Statistical Society* B, **50**, 157–224.