Counting degrees of freedom in hierarchical and other richly-parameterized models

By JAMES S. HODGES[1] and DANIEL J. SARGENT[2]

[1] *Division of Biostatistics, School of Public Health, University of Minnesota*

[2] *Division of Biostatistics, Mayo Clinic*

Correspondence author: James S. Hodges
Division of Biostatistics,
University of Minnesota,
2221 University Ave SE, Suite 200,
Minneapolis, Minnesota 55414 U.S.A.
telephone: (612) 626-0995
email: hodges@gopher.ccbr.umn.edu

February 12, 1998

Draft: Please do not circulate. Comments are most welcome.

Counting degrees of freedom in hierarchical and other richly-parameterized models

ABSTRACT

Drawing on linear model theory, we rigorously extend the notion of degrees of freedom to richly-parameterized models, including linear hierarchical and random-effect models, some smoothers, and combinations of these. The number of degrees of freedom can be and often is much smaller than the number of parameters. Our notion of degrees of freedom is compatible with similar ideas long associated with smoothers, but is applicable to new classes of models and can be interpreted using the projection theory of linear models. One use of this notion, which we examine through an example, is to control the size of a richly- parameterized model by setting prior distributions for its variances.

Key Words: degrees of freedom, hierarchical model, prior distribution, random-effect model, smoothing

1 Introduction

A common goal in statistical analyses is to find a model for a dataset which describes and explains key features of the data and is as simple as possible. In recent years, more and more elaborate models have been proposed for the increasingly complicated datasets made possible by computing advances. Often these models are hierarchical, that is, a simple statistical model (e.g., a linear regression) serves as the base for a sequence of models specified in a hierarchy. It is now fairly easy to fit such models but it is not clear how to judge their complexity.

For example, consider the balanced one-way random-effects model. Suppose we have a total of Nn observations $y_{ij}, i = 1, \ldots, N, j = 1, \ldots, n$. The one-way random-effects model assumes first, that for each i the observations y_{ij} have a common mean θ_i, and second, that the θ_i are draws from a distribution with mean μ. Commonly, the model is

$$y_{ij}|\theta_i, \sigma^2 \sim N(\theta_i, \sigma^2), i = 1, \ldots, N, j = 1, \ldots, n \quad (1)$$

$$\theta_i|\mu, \tau^2 \sim N(\mu, \tau^2), i = 1, \ldots, N. \quad (2)$$

A Bayesian analysis adds prior distributions for μ, σ^2, and τ^2.

If we ask how complex or large this model is, the answer depends on which viewpoint we adopt. In one approach, often associated with the label "variance component modeling", the θ_i are integrated out, leaving the model with three parameters — μ, σ^2, and τ^2. In another approach, often associated with the term "hierarchical [or multilevel] modeling", the θ_i remain in the model, so there are $N + 3$ parameters: N θ_i's, μ, σ^2, and τ^2. However, it would seem that the answer must depend on σ^2 and τ^2. For finite, positive σ^2, if $\tau^2 = \infty$, there are plainly $N + 1$ parameters: N θ_i's and σ^2. But if $\tau^2 = 0$, there are plainly two parameters, μ and σ^2. It seems awkward to suggest that the model's complexity changes discontinuously when τ^2 reaches either zero or infinity and is

constant otherwise, but no continuous measure of complexity is immediately obvious. The purpose of this paper is to formulate a notion of complexity for models like this one, and to measure it.

The complexity of single-level models is usually measured by *degrees of freedom* (DF). For linear models, DF are a property of the vector space into which the y_{ij} are projected to obtain fitted values $\hat{y}_{ij}$. In this paper we extend this linear-model theory to hierarchical and other richly-parameterized models. In Section 2 we review the relevant theory for the usual linear model, then show how hierarchical and other richly-parameterized models can be expressed formally as linear models. Section 3 proposes a method for counting DF for models that can be expressed in the form given in Section 2. Section 4 then gives examples. Section 5 relates our approach to similar notions of DF developed for smoothers (e.g., Hastie and Tibshirani 1990) and for very general modelling methods (Ye 1998). Section 6 illustrates how our approach may be helpful in understanding the implications of prior distributions for variances in richly-parameterized models, and examines a simple problem in detail. Section 7 concludes and briefly mentions areas for future research.

2 Background

2.1 Degrees of freedom in ordinary linear models

Consider the usual linear model

$$y = A\beta + \epsilon \tag{3}$$

where y is $d \times 1$, A is $d \times p$, β is $p \times 1$, ϵ is $d \times 1$, and $\text{cov}(\epsilon) = \sigma^2 I_d$, I_d being the identity matrix of dimension d. The DF of this model, ρ, is generally taken to be the dimension of the column space of A: if A is of full column rank, the model has p degrees of freedom. Regardless of A's rank, $\rho = tr[A(A'A)^- A']$ where A^- is the generalized inverse of X (see, e.g., Searle, Casella, and McCulloch 1992, p. 451). Another way to express this, which foreshadows our development for

richly-parameterized models, involves a reparameterization. Assuming A is full-rank, let $A = U\Delta V'$ be A's singular value decomposition, where U is $d \times p$, Δ and V are $p \times p$, Δ is diagonal, and U and V have orthonormal columns. If we define $\xi = \Delta V'\beta$, then model (3) becomes

$$y = U\xi + \epsilon \tag{4}$$

and

$$\rho = tr[U(U'U)^{-1}U'] = tr[UU'] = tr[U'U] = \sum_{k=1}^{p} U^{(k)\prime}U^{(k)} = p, \tag{5}$$

where the $U^{(k)}$ are the columns of U. The singular value decomposition has no special role here except to prove the existence of an orthonormal basis for A's column space; any other orthonormal basis for A's column space yields the same result.

2.2 Expressing richly-parameterized models as linear models

To count the DF in a richly-parameterized linear model, we use the technique of Hodges (1998) to express such models formally as linear models. Consider again the balanced one-way random-effects model of Section 1, which can be represented by two equations:

$$y_{ij} = \theta_i + \epsilon_{ij} \tag{6}$$

$$\theta_i = \mu + \delta_i, \tag{7}$$

where $\epsilon_{ij} \sim N(0, \sigma^2)$ and $\delta_i \sim N(0, \tau^2)$. Rewrite (7) as

$$0 = -\theta_i + \mu + \delta_i. \tag{8}$$

Combining equations (7), and (8), we obtain the linear model

$$\left[\begin{array}{c} y \\ \hline 0_N \end{array}\right] = \left[\begin{array}{ccc|c} 1_n & \cdots & 0 & \\ \vdots & \ddots & \vdots & 0_{N\times n} \\ 0 & \cdots & 1_n & \\ \hline \multicolumn{3}{c|}{-I_N} & 1_N \end{array}\right] \left[\begin{array}{c} \theta_1 \\ \vdots \\ \theta_N \\ \hline \mu \end{array}\right] + \left[\begin{array}{c} \epsilon \\ \hline \delta \end{array}\right] \tag{9}$$

where $y = \{y_{ij}\}$, 1_m and $0_{m\times r}$ are a column vector of 1's and a matrix of zeros of the specified dimensions respectively, $\epsilon = \{\epsilon_{ij}\}$, and $\delta = \{\delta_i\}$. This is a linear model in the formal sense that the left-hand side is known and the right hand side consists of known linear combinations of unknown parameters with a heteroskedastic vector of additive errors.

Hodges (1998) shows that Bayesian inferences based on (9) and prior distributions for μ, σ^2, and τ^2 are identical to inferences drawn from the typical formulation. In effect, (9) is simply an accounting identity (Whittaker 1998). Many richly-parameterized models can be reexpressed in a similar way, yielding the same Bayesian analysis as the usual formulation, specifically every hierarchical linear model in Bryk and Raudenbush (1992), every linear model with additive errors in West and Harrison (1989), and every variance component model in Chapters 1-9 of Searle et al (1992). When re-expressed, all these models fit into a general form like (9):

$$\left[\begin{array}{c} y \\ \hline 0_{c\times 1} \end{array}\right] = \left[\begin{array}{c|c} X_1 & 0_{d\times p} \\ \hline Z_1 & Z_2 \end{array}\right] \left[\begin{array}{c} \Theta_1 \\ \hline \Theta_2 \end{array}\right] + \left[\begin{array}{c} \epsilon \\ \hline \delta \end{array}\right] \tag{10}$$

where y is $d \times 1$, X_1 is $d \times q$, Z_1 is $c \times q$, Z_2 is $c \times p$, Θ_1 is $q \times 1$, Θ_2 is $p \times 1$, ϵ is $d \times 1$, and δ is $c \times 1$. The symbols d and p are used by analogy to the ordinary linear model (3). In more compact notation, (10) can be written

$$Y = X\Theta + E, \tag{11}$$

where X and Y are known, Θ is unknown, and E is the error term. We define rows of X and Y in (11) corresponding to X_1 in (10) to be "data cases". Data cases are the terms in the joint posterior into which the data y enter directly. We call rows of X and Y corresponding to Z_1 "constraint cases". These rows place restrictions — stochastic constraints — on possible values of Θ_1. One can add prior information about Θ_2 by adding further artificial cases, as in Theil (1971); we call these "prior cases". (Hodges 1998 gives more details.)

The covariance matrix Γ of the error term E has a block-diagonal structure with blocks Γ_1, Γ_2, and Γ_3 corresponding to the data, constraint, and prior cases, respectively. In general, Γ_1 and Γ_2 will contain unknown parameters, while the prior cases are distinguished from the constraint cases because Γ_3 is completely specified. For the remainder of this paper, we set $\Gamma_3^{-1} = 0$, that is, we use a flat prior for the highest-level parameters in Θ_2. This is equivalent to simply omitting the prior cases, which we henceforth do. This flat prior creates no pathologies except in computing Bayes factors or posterior probabilities of models of the form (11).

3 Counting degrees of freedom in richly-parameterized models

Suppose we have a richly-parameterized model in the form (10), with $\mathrm{cov}(E) = \Gamma$, and assume X is of full rank. For now, assume Γ is known. Pre-multiply (11) by $\Gamma^{-\frac{1}{2}}$ to create a homoskedastic problem, giving the model

$$\mathsf{Y} = \left[\begin{array}{c} \mathsf{y} \\ \hline 0 \end{array}\right] = \mathsf{X}\Theta + \mathsf{E} = \left[\begin{array}{c|c} \mathsf{X}_1 & 0 \\ \hline \mathsf{Z}_1 & \mathsf{Z}_2 \end{array}\right] \Theta + \mathsf{E} \tag{12}$$

where the sans-serif font indicates premultiplication by $\Gamma^{-\frac{1}{2}}$ and $\mathrm{cov}(\mathsf{E}) = I_{d+c}$. Next, compute the singular value decomposition of X, $\mathsf{X} = U\Delta V'$, and define $\Xi = \Delta V'\Theta$, where U is $(d+c) \times (p+q)$,

Δ and V are $(p+q) \times (p+q)$, and Δ is diagonal. The model can now be expressed as

$$\left[\begin{array}{c} \mathsf{y} \\ \hline 0 \end{array}\right] = \mathsf{Y} = U\Xi + \mathsf{E} = \left[\begin{array}{c} U_1 \\ \hline U_2 \end{array}\right] \Xi + \mathsf{E}, \tag{13}$$

where U_1, corresponding to the data cases, is $d \times (p+q)$, and U_2, corresponding to the constraint cases, is $c \times (p+q)$.

The key to counting DF in this model is to consider the dimension of the subspace onto which the observation vector y is projected, yielding the fitted values $\hat{\mathsf{y}}$. Usual linear model theory gives

$$\hat{\mathsf{y}} = \mathsf{X}_1\hat{\Theta} = \mathsf{X}_1(\mathsf{X}'\mathsf{X})^{-1}\mathsf{X}'\mathsf{Y} = \mathsf{X}_1(\mathsf{X}'\mathsf{X})^{-1}\mathsf{X}_1'\mathsf{y} \tag{14}$$

$$= U_1\hat{\Xi} = U_1(U'U)^{-1}U'\mathsf{Y} = U_1U_1'\mathsf{y}. \tag{15}$$

(In a frequentist interpretation, $\hat{\Xi}$ is the generalized least squares estimate of Ξ for the value of Γ that we have assumed is known. In a Bayesian interpretation, $\hat{\Xi}$ is the posterior mean of Ξ conditional on Γ.) In the usual linear model, we computed DF by expression (5), because $U(U'U)^{-1}U'$ projects the observation vector onto the column space of U and thus of A. In the richly-parameterized linear model (12), y is projected onto the column space of U_1 using $\hat{\mathsf{y}} = U_1U_1'\mathsf{y}$. It is, therefore, consistent with the usual linear model theory to define DF in richly-parameterized models to be

$$\rho = tr[U_1U_1'] = tr[U_1'U_1] = \sum_{k=1}^{p} U_1^{(k)\prime}U_1^{(k)}, \tag{16}$$

so the projection onto the kth column of U_1 can be said to have $U_1^{(k)\prime}U_1^{(k)}$ degrees of freedom. U_1U_1' is generally not an orthogonal projection.

By construction, the columns of U satisfy $U^{(k)\prime}U^{(k)} = 1$, $k = 1, 2, ..., r$. The columns of U_1,

however, satisfy $U_1^{(k)\prime}U_1^{(k)} \leq 1$ and except for trivial cases, $0 < U_1^{(k)\prime}U_1^{(k)} < 1$ for at least one k. Thus, y is projected onto a space whose basis vectors are, loosely speaking, less than full size. We can make this more precise in two ways. Suppose $U_1^{(k)\prime}U_1^{(k)} < U_1^{(l)\prime}U_1^{(l)}$. Suppose further that $\hat{\Xi}_k = \hat{\Xi}_l$; then $\hat{\mathsf{y}}$'s projection along $U_1^{(k)}$, $U_1^{(k)}\hat{\Xi}_k$, has smaller Euclidean norm than $\hat{\mathsf{y}}$'s projection along $U_1^{(l)}$. Heuristically, $U_1^{(k)}$'s "vote" (i.e., $\hat{\Xi}_k$) counts less than $U_1^{(l)}$'s "vote". But $U_1^{(k)}$ will also tend to *receive* a smaller vote in the fitted model than $U_1^{(l)}$, because $max(\hat{\Xi}_k) = max(U_1^{(k)\prime}\mathsf{y}) = (\mathsf{y}'\mathsf{y})^{\frac{1}{2}}(U_1^{(k)\prime}U_1^{(k)})^{\frac{1}{2}} < max(\hat{\Xi}_l)$.

As with the ordinary linear model, we have used the singular value decomposition of X to demonstrate existence, but for any $(p+q)$-dimensional orthogonal matrix P, $U^* = UP$ is easily shown to produce the same count of DF as U. Another basis for the column space of X simply produces a different allocation of the DF among directions on which y is projected.

In this discussion, we have assumed $\Gamma = \mathrm{Cov}(E)$ is known; in practice, elements of it are unknown. From a Bayesian perspective, we have conditioned on Γ; removing this conditioning implies that the DF in a model and dataset is a function of unknown parameters and hence has a posterior distribution. (This definition of DF meshes tidily with Bayesian computations using Markov chain Monte Carlo [MCMC]). In other words, in a Bayesian fit, the data adaptively determine the complexity of the fitted model. We discuss this further in Section 6, where we consider setting a prior distribution on ρ as a way to induce a prior on the unknown elements of Γ. From a frequentist or smoothing perspective, choosing Γ is equivalent to choosing smoothing or penalty constants, and we can do so by estimating Γ, by cross-validation, or by other means.

4 Examples

4.1 One-way random effects model

Consider the balanced one-way random-effects model from Section 1. Using the form of (10), with the design matrix from (9), the error covariance matrix Γ is $(Nn+N)\times(Nn+N)$ and diagonal, the first Nn diagonal elements being $\sigma^2 = \frac{1}{h_e}$ and the last N elements $\tau^2 = \frac{1}{h}$. Premultiply X by $\Gamma^{-\frac{1}{2}}$ to obtain

$$\mathsf{X} = \left[\begin{array}{ccc|c} h_e^{\frac{1}{2}}1_n & \cdots & 0 & \\ \vdots & \ddots & \vdots & 0_{Nn} \\ 0 & \cdots & h_e^{\frac{1}{2}}1_n & \\ \hline & -h^{\frac{1}{2}}I_N & & h^{\frac{1}{2}}1_N \end{array}\right]. \tag{17}$$

Due to X's simple form, it is more transparent to count DF by the formula (suggested by (15))

$$\rho = tr(\mathsf{X}_1(\mathsf{X}'\mathsf{X})^{-1}\mathsf{X}_1) = \frac{nN + \frac{\sigma^2}{\tau^2}}{n + \frac{\sigma^2}{\tau^2}}. \tag{18}$$

Equation (18) shows how each piece of the model affects ρ. Most important, ρ is determined by the ratio $\frac{\sigma^2}{\tau^2}$, not by the individual variances. Note that $\rho \geq 1$ and $\rho \to 1$ as $\frac{\sigma^2}{\tau^2}$ approaches infinity — that is, when within-group variation (σ^2) is extremely large compared to between-group variation (τ^2), the fitted θ_i are shrunk to μ. Also, $\rho \leq N$, with $\rho = N$ in the limit as $\frac{\sigma^2}{\tau^2}$ approaches zero or as n approaches infinity, that is, as the variation of each group's average becomes small compared to between-group variation.

When we think of (say) two degrees of freedom in the ordinary linear model (3), we think of a projection space in which the fitted values can vary arbitrarily in the plane specified by the two basis vectors. In the random-effects model, if we set $\frac{\sigma^2}{\tau^2}$ so $\rho = 2$, we have a projection space with N dimensions, but the fitted values can no longer vary arbitrarily in that N-dimensional space.

Instead, they are constrained (by the constraint cases) to lie in a "small" irregular ball centered on a point on the vector corresponding to $\theta_i \equiv \mu$, where "small" is specified precisely by $\rho = 2$ and the assumption of normal errors. The center of the ball can move arbitrarily along the vector corresponding to $\theta_i \equiv \mu$.

4.2 Simple smoother

If we have T observations that are ordered (in time, say) and equally spaced, the simple Kalman smoother (Kalman 1960) assumes that pairwise differences of adjacent observations have mean zero. Specifically, the model is

$$y_t = \theta_t + \epsilon_t, \epsilon_t \sim N(0, \sigma^2), t = 1, \ldots, T \tag{19}$$

$$\theta_t = \theta_{t-1} + \delta_t, \delta_t \sim N(0, \tau^2), t = 2, \ldots, T. \tag{20}$$

Re-expressed as in Section 2.2, this gives the linear model

$$\left[\begin{array}{c} y_1 \\ \vdots \\ y_T \\ \hline 0_{T-1} \end{array}\right] = \left[\begin{array}{cccccc} & & I_T & & & \\ \hline 1 & -1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & -1 & \cdots & 0 & 0 \\ & & & \vdots & & \\ 0 & 0 & 0 & \cdots & 1 & -1 \end{array}\right] \left[\begin{array}{c} \theta_1 \\ \vdots \\ \theta_T \end{array}\right] + \left[\begin{array}{c} \epsilon \\ \hline \delta \end{array}\right] \tag{21}$$

Unlike the random-effects model in Section 4.1, this model has no closed-form expression for ρ. However, the basis matrix U obtained from the singular value decomposition has two interesting properties. The first property is a pattern in the columns of U_1. One column of U_1 is constant across $t = 1, 2, ..., T$ and has 1 DF for all σ^2 and τ^2. The k^{th} largest DF, $k = 2, 3, ..., T$, belongs

to a column that is approximately a $(k-1)^{th}$-order polynomial in T, where the approximation improves as T grows. The DF of the columns decline at a slower-than-exponential rate in the order of these polynomials. The total DF ρ declines as $\frac{\sigma^2}{\tau^2}$ increases, that is, as the permissible change in θ_t between t's is reduced for fixed σ^2. For example, for $T = 10$ and $\frac{\sigma^2}{\tau^2}$ taking the values 1, 2, 4, 8, and 16, ρ takes the values 4.9, 3.8, 2.9, 2.3, and 1.8, respectively.

The second interesting property is that for all σ^2 and τ^2, there is an orthonormal basis matrix for the column space of X whose first d rows are $U_1\Psi$, where U_1 is given by the aforementioned singular-value decomposition and Ψ is a diagonal matrix. Thus, the basis for y's projection space is, in a specific sense, invariant to changes in σ^2 and τ^2.

4.3 Error distributions that are scale mixtures of normals

The error terms for the data, constraint, and prior cases need not be normally distributed. If, instead, they are scale mixtures of normals, e.g., with a t or double-exponential distribution, then the error distribution for each case can be written as normal conditional on a mixing variate that has the appropriate distribution. The foregoing theory can then be applied conditional on the mixing variates and any scale factors. It is not necessary for all cases to have the same kind of scale-mixed normal distribution. As with normal errors, this lends itself readily to Bayesian computations using MCMC.

5 Relationship with other approaches

5.1 Degrees of freedom in the smoother literature

The notion of DF has a long history for smoothers and related methods like generalized additive

models. The notion of DF in Section 3 was developed for a class of models that includes some smoothers. As it turns out, our approach can be used to extend the familiar theory of DF for smoothers (e.g., Hastie and Tibshirani [henceforth HT] 1990) to richly-parameterized linear models that can be expressed in the form (10).

For a linear smoother operator S — where $\hat{y} = Sy$ is the smooth of y — HT give three definitions of DF: $trace(S)$, $trace(2S - SS')$ and $trace(SS')$ (HT 1990, Section 3.5). In Section 3, we defined our measure of DF as the trace of the projection (smoothing) matrix $U_1 U_1'$, which corresponds to HT's first definition. The main goal of this section is to derive the other two measures of DF within our framework.

Before we do, it is reasonable to ask: which definition of DF is preferable? We agree with HT (section 3.5, p. 53) that each definition is useful for a different purpose. The considerations that motivate $trace(2S - SS')$ are appropriate for (non-Bayesian) tests of significance, while $trace(U_1 U_1')$ (i.e. $trace(S)$) characterizes the posterior behavior of the parameter vector Θ and the fitted values arising from it. For Bayesian purposes, such as setting prior distributions on unknown elements of Γ (see Section 6), we prefer the latter.

As a first step in deriving $trace(2S - SS')$ and $trace(SS')$ in our framework, note that HT obtained these definitions from frequentist derivations where the true smooth function is fixed but unknown and hypothetical repeated draws are made of the error vector. In our constraint-case formulation (10), this corresponds to fixing Θ_2, fixing the constraint-case error vector δ at the value corresponding to the unknown true Θ_1, and then deriving properties over repeated draws of the data-case error vector ϵ. In HT's framework, a smoother is given and draws are made from the error vector; the analog, in our framework, is to condition on a value of Γ and draw the data-case error vector ϵ from its true distribution, which has a covariance matrix that, in general, differs from Γ_1. To emulate HT, we need to fix Γ_1 and make repeated draws from the true distribution of ϵ.

Thus, we define the true covariance matrix of ϵ to be γ_1.

HT obtain the definition $\rho = trace(2S - SS')$ as a property of the residual sum of squares — in our formulation, the data-case residual sum of squares. To follow their derivation in our framework, fix a value of Γ, transform the problem to the form (12), and make the canonical reparameterization to arrive at (13). Using the subscript Γ for emphasis and defining $H_\Gamma = U_\Gamma U_\Gamma'$, the residual vector from the maximum likelihood or least-squares fit is

$$\hat{\mathsf{E}}_\Gamma = (I - H_\Gamma)\mathsf{Y} = (I - H_\Gamma)(U_\Gamma \Xi + \mathsf{E}) = (I - H_\Gamma)\Gamma^{-\frac{1}{2}} \left[\begin{array}{c} \epsilon \\ \hline \delta \end{array} \right]. \tag{22}$$

Using the notation $U_\Gamma' = [U_{1\Gamma}' | U_{2\Gamma}']$, the data-case residuals are

$$[I - U_{1\Gamma}U_{1\Gamma}' | U_{1\Gamma}U_{2\Gamma}']\Gamma^{-\frac{1}{2}} \left[\begin{array}{c} \epsilon \\ \hline \delta \end{array} \right]. \tag{23}$$

Let the data-case residual sum of squares be denoted by $DCRSS$. Following HT, the expectation of $DCRSS$ — over repeated draws of ϵ with δ held fixed — is:

$$\mathcal{E}(DCRSS) = bias + trace(\gamma_1^{\frac{1}{2}}\Gamma_1^{-\frac{1}{2}}(I - U_{1\Gamma}'U_{1\Gamma})^2\Gamma_1^{-\frac{1}{2}}\gamma_1^{\frac{1}{2}}) \tag{24}$$

where $bias = \delta'\Gamma_2^{-\frac{1}{2}}U_{2\Gamma}U_{1\Gamma}'U_{1\Gamma}U_{2\Gamma}'\Gamma_2^{-\frac{1}{2}}\delta$, and $\mathcal{E}$ indicates expectation. If $\Gamma_1 = \hat{\sigma}^2 I_d$ and $\gamma_1 = \sigma^2 I_d$, d being the number of data cases, then

$$\mathcal{E}(DCRSS) = bias + \frac{\sigma^2}{\hat{\sigma}^2}(d - trace(2U_{1\Gamma}'U_{1\Gamma} - (U_{1\Gamma}'U_{1\Gamma})^2)). \tag{25}$$

Following HT (1990, section 3.5), this last expression motivates the definition $\rho = trace(2U_1'U_1 -$

$(U_1'U_1)^2)$ as the number of DF in this model.

The definition $\rho = trace(SS')$ arises from a similar exercise with the data-case regression sum of squares ($DCRegSS$). Specifically,

$$\mathcal{E}(DCRegSS) = bias + trace(\gamma_1^{\frac{1}{2}}\Gamma_1^{-\frac{1}{2}}(U_{1\Gamma}'U_{1\Gamma})^2\Gamma_1^{-\frac{1}{2}}\gamma_1^{\frac{1}{2}}), \tag{26}$$

where the term $bias$ is the same as in (24). Again, if $\Gamma_1 = \hat{\sigma}^2 I_d$ and $\gamma_1 = \sigma^2 I_d$, then

$$\mathcal{E}(DCRegSS) = bias + \frac{\sigma^2}{\hat{\sigma}^2}trace((U_{1\Gamma}'U_{1\Gamma})^2), \tag{27}$$

motivating the definition $\rho = trace((U_{1\Gamma}'U_{1\Gamma})^2)$.

The foregoing extends the three definitions of DF from linear smoothers to all richly-parameterized linear models fitting into our framework. HT (1990, Section 6.8) and others (e.g., Gray 1992) have also extended the definition $\rho = trace(2S - SS')$ to generalized additive models and various nonlinear smoothers. We have, as yet, made no such extension.

5.2 Ye's generalized degrees of freedom

Consider a real d-vector $Y = (y_1, ..., y_d)'$ having a multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance $\sigma^2 I_d$, and let i index the elements of Y. Suppose $\mathcal{M}$ is a modeling procedure mapping Y to $\hat{\boldsymbol{\mu}} \in \Re^d$. $\mathcal{M}$ can be a simple procedure like the mean or a regression, or a complicated procedure like a tree-fitter that prunes by cross-validation. Ye (1998) defines the generalized degrees of freedom of $\mathcal{M}$ by

$$GDF = \sum_{i=1}^{n} \frac{\partial \mathcal{E}_{\boldsymbol{\mu}}[\hat{\mu}_i(Y)]}{\partial \mu_i} = \sum_{i=1}^{n} \mathcal{E}_{\boldsymbol{\mu}}[\hat{\mu}_i(Y)(y_i - \mu_i)]/\sigma^2 \tag{28}$$

where the notation $\mathcal{E}_{\boldsymbol{\mu}}$ indicates that the expectation is over the distribution of Y, conditional on $\boldsymbol{\mu}$. Ye (1998) shows that GDF is the sum over data cases of the average sensitivity of $\hat{\mu}_i(Y)$ to a small change in y_i, and thus measures the flexibility of the modeling procedure $\mathcal{M}$.

Conditional on Γ, our measure of DF equals GDF. Also, for the fitted values $X_1\mathcal{E}(\Theta_1|Y)$, if we define DF $= \mathcal{E}(\rho|Y)$, then our measure of DF again equals GDF. GDF is completely general in the sense that $\mathcal{M}$ can be literally anything. However, GDF can be written in closed form only for simple $\mathcal{M}$ and can usually be computed only by simulation. In this sense, our definition of DF is more explicit, for the cases in which both are defined.

6 DF and prior distributions for variances

One is rarely in a position to set prior distributions for the variances in a richly-parameterized model and be confident of their implications. Sometimes relevant previous data are available, but in the more common absence of *bona fide* prior data, it is hard to think even qualitatively about the consequences of a given prior. The usual expedient — flat or otherwise improper priors — causes problems; see O'Hagan (1976, 1985), Hoeschele (1989), Hobert and Casella (1996), and the latter part of this section.

The foregoing notion of DF offers an alternative approach to specifying such priors. A richly-parameterized model's DF is a function of Γ; if you specify desired features of a prior on DF, these constrain or fully specify a prior on the unknown elements of Γ. Suppose, for example, you fix $\mathcal{E}(\rho) = 3$. If the prior standard deviation of ρ is small, then $\mathcal{E}(\rho|data)$ will be near 3; in the random-effects model, this means you have inserted strong prior information about the ratio $\frac{\sigma^2}{\tau^2}$. (In this sense, someone using a smoother in, say, S+ and fixing the DF is implicitly using a point prior on $\frac{\sigma^2}{\tau^2}$ in a particular richly-parameterized model.) If, on the other hand, the prior standard

deviation of ρ is large, then $\mathcal{E}(\rho|data)$ will be shrunk toward 3 a bit, but it will be primarily influenced by the data (and the model specification) — that is, you have inserted weak prior information about the ratio $\frac{\sigma^2}{\tau^2}$.

In the remainder of this section, we consider in detail the one-way random-effects model with conjugate priors for the precisions, to illustrate how DF can be helpful in showing the implications of particular priors in richly-parameterized models. Before we do, though, we can derive two facts about priors and posteriors on ρ without any further assumptions. First, ρ takes values in a closed, bounded set, the lower bound being no less than zero and the upper bound no greater than $p+q$. Thus, if ρ has a proper distribution, it has moments of all orders. The second, related fact is that ρ's variance is bounded above by a function of its expectation. If $\rho \in [l, u]$, then $var(\rho) \leq (\mathcal{E}(\rho) - l)(u - \mathcal{E}(\rho))$, with equality occurring when ρ's distribution places all its mass on l and u.

Consider, then, the one-way random-effects model, defining the precisions $h_e = \frac{1}{\sigma^2}$ and $h = \frac{1}{\tau^2}$. As noted, $\rho \equiv \rho(h_e, h) = \frac{Nn+r}{n+r}$, where $r = \frac{h}{h_e} = \frac{\sigma^2}{\tau^2}$. If we assume the usual conjugate priors $h_e \sim$ Gamma(α_d, λ_d), having mean α_d/λ_d (d indicating data cases) and $h \sim$ Gamma(α_c, λ_c), having mean α_c/λ_c (c indicating constraint cases), then straightforward distribution theory gives ρ's prior density as:

$$f(\rho) = K \frac{(N-\rho)^{\alpha_c - 1}(\rho - 1)^{\alpha_d - 1}}{[\lambda_d(\rho - 1) + n\lambda_c(N - \rho)]^{\alpha_c + \alpha_d}} \tag{29}$$

where

$$K = \frac{\Gamma(\alpha_d + \alpha_c)\lambda_d^{\alpha_d}\lambda_c^{\alpha_c} n^{\alpha_c}(N-1)}{\Gamma(\alpha_d)\Gamma(\alpha_c)} \tag{30}$$

and $\rho \in [1, N]$; this is a generalized beta density. If we set $\lambda_d = n\lambda_c$, (29) is a beta distribution; we can obtain a flat prior on ρ by also setting $\alpha_d = \alpha_c = 1$. (Purists note! The prior distribution then depends on the within-group sample size n.) For $\lambda_d \neq n\lambda_c$, ρ's moments appear to be intractable.

But r's moments are readily obtained as:

$$\mathcal{E}(r) = \frac{\alpha_c \lambda_d}{\lambda_c(\alpha_d - 1)}, \qquad \alpha_d > 1 \tag{31}$$

$$var(r) = \frac{\lambda_d^2 \alpha_c(\alpha_c + \alpha_d - 1)}{\lambda_c^2(\alpha_d - 1)^2(\alpha_d - 2)}, \tag{32}$$

$$= \frac{\alpha_c + \alpha_d - 1}{\alpha_c(\alpha_d - 2)} \mathcal{E}(r)^2, \quad \alpha_d > 2. \tag{33}$$

We can now see some possibly undesirable features of the conjugate priors for h_e and h. To make $var(\rho)$ small, we must make $var(r)$ small, but for fixed $\mathcal{E}(r)$, we can only make $var(r)$ small by making both α_c and α_d large, that is, by inserting strong prior information about both h and h_e. On the other hand, to achieve the maximal $var(\rho)$ for fixed $\mathcal{E}(\rho) \in (1, N)$, the prior on ρ must place all its mass on 1 and N. But when α_c and α_d are both positive, $f(\rho)$ gives positive probability to $(1, N)$. Thus, the conjugate priors for h and h_e provide near-maximal prior variance for ρ only to the extent that their scale parameters approach zero (i.e.,they are nearly improper).

These properties arise because we assumed h and h_e independent *a priori*. An alternative prior specification has $h_e \sim \text{Gamma}(\alpha_d, \lambda_d)$, and $h = r h_e$ with r distributed independently of h_e. Then ρ's prior depends only on the prior for r, which can take any desired form. Analytically this is less convenient than conjugate priors, but it creates no difficulty in, for example, the MCMC computing package BUGS (Spiegelhalter et al 1995, 1997).

It is straightforward to derive the posterior distribution of ρ under the conjugate priors for h and h_e and a normal or flat prior for μ. Let $\rho = 1 + (N-1)s$, where $s = \frac{n h_e}{n h_e + h} = \frac{n}{n+r}$ is the complement of the usual shrinkage factor. Using a flat prior for μ, s has posterior density

$$f(s|data) \propto s^{Nn/2 + \alpha_d - 3/2}(1-s)^{N/2 + \alpha_d - 1} R^{-(Nn/2 + \alpha_d + \alpha_c)} \tag{34}$$

where $R = N(n-1)s + (N-1)Fs(1-s) + (2\lambda_d s + 2n\lambda_c(1-s))/RMS$, F being the usual F-statistic for the between-group effect and RMS being the usual regression mean square (*cf.* O'Hagan 1985). The last two terms in R represent the effect of the priors for h and h_e. Popular improper priors for h or h_e correspond to setting λ_c or λ_d to zero, respectively, for various values of α_c and α_d. For example, a flat prior on h_e corresponds to $\alpha_d = 1$ and $\lambda_d = 0$. Setting $\lambda_d = 0$ creates a problem in this posterior only if $n = 1$, in which case $f(s|data)$ contains the factor $(1-s)^{-\alpha_c - 1}$ and is thus improper for $\alpha_d \geq 0$. Similarly, a flat prior on h corresponds to $\alpha_c = 1$ and $\lambda_c = 0$, but if $\lambda_c = 0$, $f(s|data)$ contains the factor $s^{-\alpha_c - 3/2}$ and is improper for all n and N for $\alpha_c \geq 0$. A proper prior for μ does not correct this problem. In other words, such an improper prior for h creates a profound instability in the resulting model fit. This is consistent with the emerging folk wisdom that one should avoid improper priors on the constraint-case precisions, while improper priors on the data-case precisions are less problematic.

7 Conclusion

We have developed a measure of model complexity that generalizes the notion of degrees of freedom used for linear smoothers to a large class of richly-parameterized linear models. Our definition also provides an approach to specifying prior distributions for the variances of those models. Future papers will develop such priors, focusing on their implications for the resulting fitted values.

Acknowledgments

JSH was supported in part by the Minnesota Oral Health Clinical Research Center, NIH/NIDR P30-DE09737), while DJS was supported in part by National Institute of Allergy and Infectious

Diseases (NIAID) Contract NO1-AI05073 and NIAID Graduate Training in Biostatistics Grant # 1-T32-A107432-01A1.

References

Bryk AS, Raudenbush SW (1992). *Hierarchical Linear Models: Applications and Data Analysis Methods.* Newbury Park, California: SAGE Publications.

Gray RJ (1992). Flexible methods for analyzing survival data using splines, with application to breast cancer prognosis. *Journal of the American Statistical Association*, **87**, 942–951

Hastie TJ, Tibshirani RJ (1990). *Generalized Additive Models.* New York: Chapman and Hall.

Hobert JP, Casella G (1996) The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. *Journal of the American Statistical Association*, **91**, 1461–1473

Hodges JS (1998). Some algebra and geometry for hierarchical models, applied to diagnostics (with discussion). To appear, *Journal of the Royal Statistical Society, Series B.*

Hoeschele I (1989) A note on local maxima in maximum likelihood, restricted maximum likelihood, and Bayesian estimation of variance components. *Journal of Statistical Computing and Simulation*, **33**,149–160.

Kalman RE (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering* **82**, 34–45.

O'Hagan A (1976). On posterior joint and marginal modes. *Biometrika*, **63**: 329–333.

O'Hagan A (1985). Shoulders in hierarchical models. *Bayesian Statistics 2*, eds. JM Bernardo, MH DeGroot, DV Lindley, AFM Smith, Amsterdam: North-Holland, 697-710.

Searle SR, Casella G, McCulloch CE (1992). *Variance Components.* New York: Wiley.

Spiegelhalter DJ, Thomas A, Best N, Gilks WR (1995). BUGS: Bayesian Inference Using Gibbs Sampling, version 0.5. Technical report, Medical Research Council Biostatistics Unit, Institute of Public Health, Cambridge University.

Spiegelhalter DJ, Thomas A, Best N, Gilks WR (1997). BUGS 0.6: Bayesian Inference Using Gibbs Sampling (Addendum to manual). Technical report, Medical Research Council Biostatistics Unit, Institute of Public Health, Cambridge University.

Theil H (1971). *Principles of Econometrics.* New York: Wiley

West M, Harrison W (1989). *Bayesian Forecasting and Dynamic Models.* New York: Springer.

Whittaker J (1998). Discussion of Hodges (1998). *Journal of the Royal Statistical Society, Series B*, forthcoming.

Ye J, (1998) On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association*, forthcoming.