Partitioning Degrees of Freedom in Hierarchical and Other Richly Parameterized Models

Yue Cui

Department of Biometrics
Abbott Labs
Abbott Park, IL 60064

James S. Hodges, Xiaoxiao Kong, and Bradley P. Carlin

Division of Biostatistics
University of Minnesota
Minneapolis, MN 55414
(*hodges@ccbr.umn.edu*)

Hodges and Sargent (2001) have developed a measure of a hierarchical model's complexity, degrees of freedom (DF), that is consistent with definitions for scatterplot smoothers, is interpretable in terms of simple models, and enables control of a fit's complexity by means of a prior distribution on complexity. But although DF describes the complexity of the whole fitted model, in general it remains unclear how to allocate DF to individual effects. Here we present a new definition of DF for arbitrary normal-error linear hierarchical models, consistent with that of Hodges and Sargent, that naturally partitions the n observations into DF for individual effects and for error. The new conception of an effect's DF is the ratio of the effect's modeled variance matrix to the total variance matrix. This provides a way to describe the sizes of different parts of a model (e.g., spatial clustering vs. heterogeneity), to place DF-based priors on smoothing parameters, and to describe how a smoothed effect competes with other effects. It also avoids difficulties with the most common definition of DF for residuals. We conclude by comparing DF with the effective number of parameters, p_D, of Spiegelhalter et al. (2002). Technical appendices and a data set are available online as supplemental materials.

KEY WORDS: Degrees of freedom; Hierarchical model; Model complexity; Prior distribution.

1. INTRODUCTION AND MOTIVATION

Hodges and Sargent (2001) developed a measure of a hierarchical model's complexity, degrees of freedom (DF), that is consistent with the definition for scatterplot smoothers and interpretable in terms of simple models. DF describes complexity of the whole fitted model but in general it is unclear how to allocate DF to individual parts of a model. Here we present an example that introduces the problem of measuring the complexity of a model's components.

Example 1 (Periodontal measurements: clustering and heterogeneity). Periodontal attachment loss (AL)—the extent of a tooth's root (in millimeters) that is no longer attached to surrounding bone by periodontal ligament—is used to diagnose and monitor periodontal disease. We present analyses of AL measurements for 12 research subjects, for each of whom AL was measured on a quadrant of 7 teeth. On each tooth, 3 sites (distal, direct, and mesial) on both the buccal (tongue) side and lingual (cheek) side were measured, giving $7 \times 3 \times 2 = 42$ sites per subject. Figure 1 is a schematic of one subject's measurements. A site's AL measurement is the sum of true loss and measurement error, which is substantial. Spatial correlation in true AL arises because if a person has poor hygiene in an area of the mouth, sites in that area may be more prone to loss, and because bacterial infection (the ultimate cause of periodontal damage) is transmitted among the sites on a given tooth. Nonspatial heterogeneity arises from local features of the dentition, for example, tooth malalignments that predispose a site to gum recession or features that make measurement difficult and affect all examiners similarly.

Gilthorpe et al. (2003) used nonspatial random effects to merge the "linear" and "burst" theories of progressive attachment loss, while Reich and Hodges (2008a) used spatial random effects to improve detection of disease progression. The preceding paragraph suggests a model with both spatial clustering and nonspatial heterogeneity in true AL, however. We consider a model commonly used by epidemiologists (Besag, York, and Mollie 1991). For simplicity, we model one subject and one measurement at each site; Section 3.1 presents an analysis of all 12 subjects. The data set is available online with the supplemental materials.

Let $\mathbf{y} \in R^N$ be observations on N sites on a spatial lattice with G islands (unconnected groups of sites); $N = 42$ and $G = 1$ in our example. Model $\mathbf{y}$ as multivariate normal with mean $\boldsymbol{\delta} + \boldsymbol{\xi}$ and error covariance $\sigma^2\mathbf{I}_N$. Nonspatial heterogeneity is captured by $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_N)'$, modeled as $\mathrm{N}(\mathbf{0}, \tau_h^2\mathbf{I}_N)$. Spatial clustering is captured by $\boldsymbol{\delta}$ as follows. Neighbor relations among sites are summarized in an $N \times N$ matrix $\mathbf{Q}$, with nondiagonal entries $q_{ij} = -1$ if site i and j are neighbors and 0 otherwise, while diagonal entry q_{ii} is the number of site i's neighbors. Figure 1 shows the neighbor pairs that we used; other choices are possible (e.g., Reich, Hodges, and Carlin 2007). Then $\boldsymbol{\delta}$ is modeled with an intrinsic conditional autoregressive model, $\mathrm{CAR}(\mathbf{Q}, \tau_c^2)$. Conditional on τ_c^2, $\boldsymbol{\delta}$ has (improper) joint density

$$f(\boldsymbol{\delta}|\tau_c^2) \propto (\tau_c^2)^{-(N-G)/2} \exp\left(-\frac{1}{2\tau_c^2}\boldsymbol{\delta}'\mathbf{Q}\boldsymbol{\delta}\right).$$

This density is always improper because $\mathbf{1}_N$ is an eigenvector of $\mathbf{Q}$ with eigenvalue 0; the impropriety implicitly allows $\boldsymbol{\delta}$ a nonzero intercept, which we make explicit later. Also note that

DOI 10.1198/TECH.2009.08161

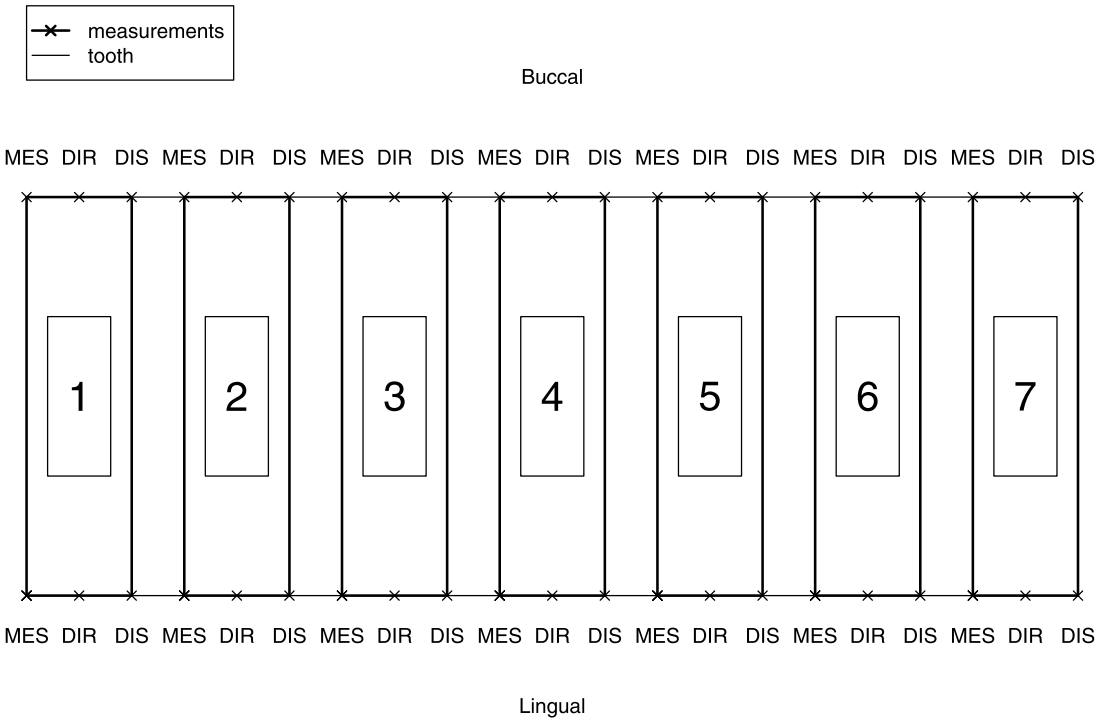

Figure 1. Example 1, neighbor relations in periodontal data. "X" indicates a measurement site, lines indicate neighbor relations, teeth are numbered from the central incisor (1) to the second molar (7); MES: mesial site, DIR: direct site, DIS: distal site.

$\boldsymbol{\delta}'\mathbf{Q}\boldsymbol{\delta} = \sum_{j\sim i}(\delta_i - \delta_j)^2$, where $j \sim i$ means that j is i's neighbor; thus $\mathbf{Q}$'s specification has the effect of shrinking neighbors toward each other.

The model's components for clustering ($\boldsymbol{\delta}$) and heterogeneity ($\boldsymbol{\xi}$) are controlled by variances τ_c^2 and τ_h^2. As τ_h^2 is increased, ξ_i's magnitude increases but without a spatial pattern. The effect of increasing τ_c^2 is determined by the spatial neighbor pairs. To see how, consider a simplified model without $\boldsymbol{\xi}$, so that $\mathbf{y} = \boldsymbol{\delta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\delta} \sim \mathrm{CAR}(\mathbf{Q}, \tau_c^2)$ as before, and $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$, and assume that $G = 1$. Given $r = \sigma^2/\tau_c^2$, the posterior mean and best linear unbiased predictor (BLUP) of $\boldsymbol{\delta}$ are $\hat{\boldsymbol{\delta}} = (\mathbf{I}_N + r\mathbf{Q})^{-1}\mathbf{y}$. Let $\mathbf{Q}$ have spectral decomposition $\mathbf{Q} = \mathbf{V}\mathbf{D}\mathbf{V}'$, where $\mathbf{D} = \mathrm{diag}(d_1, \ldots, d_{N-1}, 0)$, $d_1 \geq \cdots \geq d_{N-1} > 0$, and the orthonormal $\mathbf{V}$ partitions as $\mathbf{V} = (\mathbf{V}_1 | \frac{1}{\sqrt{N}}\mathbf{1}_N)$ where $\mathbf{V}_1$ is $N \times (N-1)$ (Hodges, Carlin, and Fan 2003). The single zero eigenvalue and conforming partition of $\mathbf{V}$ arise because there are $G = 1$ islands. Then $\hat{\boldsymbol{\delta}} = \mathbf{V}\hat{\boldsymbol{\phi}}$, where $\hat{\boldsymbol{\phi}} = (\mathbf{I}_N + r\mathbf{D})^{-1}\mathbf{V}'\mathbf{y}$. Because $d_N = 0$, $\hat{\phi}_N = \frac{1}{\sqrt{N}}\bar{y}$ regardless of r, whereas for $i < N$, $\hat{\phi}_i = (\mathbf{V}_1'\mathbf{y})_i/(1 + rd_i)$, which is shrunk toward 0 by $rd_i > 0$. For the lattice in Figure 1, the three smallest eigenvalues, d_{N-1}, d_{N-2}, and d_{N-3}, are smaller than the largest, d_1, by factors of 248, 62, and 28, respectively; the corresponding columns of $\mathbf{V}_1$ describe roughly linear, quadratic, and cubic trends along the long axis of Figure 1. Columns of $\mathbf{V}_1$ corresponding to increasingly larger d_i describe increasingly higher-frequency variation in the spatial structure, whose coefficients $\hat{\phi}_i$ are shrunk toward 0 to increasingly greater degrees for given $r = \sigma^2/\tau_c^2$. Thus for small τ_c^2, the fitted values $\hat{\boldsymbol{\delta}}$ mostly reflect a damped large-scale structure (Reich and Hodges 2008b), and as τ_c^2 increases, damping is reduced in all $\hat{\boldsymbol{\phi}}$, and the fit becomes increasingly wiggly.

After this model is fitted, any of several measures can be used to describe the whole model's complexity, but the distinct contributions of the heterogeneity and clustering components are not well understood. In this article we propose a partition of a readily interpretable measure of complexity (Hodges and Sargent 2001; henceforth H&S). H&S used a model in which observed data $\mathbf{y} \in R^d$ are assumed to be multinormal with mean $\mathbf{X}_1\boldsymbol{\theta}$ and covariance matrix $\boldsymbol{\Gamma}_0$, with $\mathbf{X}_1$ known and $\boldsymbol{\theta}$ constrained by a model or prior distribution, $\mathbf{Z}\boldsymbol{\theta} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Gamma}_1)$, with $\mathbf{Z}$ known. $\boldsymbol{\Gamma}_0$ and $\boldsymbol{\Gamma}_1$ are usually functions of a few unknowns. H&S reformulated this as a linear model,

$$\mathbf{y} = \begin{pmatrix} \mathbf{y} \\ \mathbf{0} \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{Z} \end{pmatrix} \boldsymbol{\theta} + \begin{pmatrix} \boldsymbol{\epsilon} \\ \boldsymbol{\delta} \end{pmatrix} = \mathbf{X}\boldsymbol{\theta} + \mathbf{e}, \tag{1}$$

where $\mathbf{e} = (\boldsymbol{\epsilon}, \boldsymbol{\delta})' \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Gamma})$ for $\boldsymbol{\Gamma} = \mathrm{diag}(\boldsymbol{\Gamma}_0, \boldsymbol{\Gamma}_1)$, so $\boldsymbol{\epsilon}$ and $\boldsymbol{\delta}$ are independent a priori conditional on unknowns in $\boldsymbol{\Gamma}$. (This formulation is equivalent to that of H&S, but we are about to discard it, so we will not belabor the equivalence.) Define $\mathbf{X} = (\mathbf{X}_1'\mathbf{Z}')'$; then, given $\boldsymbol{\Gamma}$, $\mathbf{X}_1\boldsymbol{\theta}$ has posterior mean or BLUP $\mathbf{X}_1(\mathbf{X}'\boldsymbol{\Gamma}^{-1}\mathbf{X})^{-1}\mathbf{X}_1'\boldsymbol{\Gamma}_0^{-1}\mathbf{y}$. H&S used linear model theory to rationalize defining the complexity or DF of this model fit given $\boldsymbol{\Gamma}$ as

$$\rho = \mathrm{tr}[\mathbf{X}_1(\mathbf{X}'\boldsymbol{\Gamma}^{-1}\mathbf{X})^{-1}\mathbf{X}_1'\boldsymbol{\Gamma}_0^{-1}],$$

and showed ρ is the same as other measures of complexity when both are defined. Lu, Hodges, and Carlin (2007) extended this definition to generalized linear hierarchical models.

H&S's DF can be used for several purposes. Model complexity is used in model-comparison criteria, although (for reasons discussed later) we do not consider such criteria here. The original aims of H&S were to describe the complexity of a fitted hierarchical model and to control that complexity by putting

a prior distribution on it, inducing a prior on unknowns in $\boldsymbol{\Gamma}$. (In Sec. 2.6 we present an accessible introduction to this idea.) Several authors have specified priors this way, for example, Paciorek and Schervish (2006), Hodges et al. (2007), and Lu, Hodges, and Carlin (2007), who extended uniform shrinkage priors (Daniels 1999; Natarajan and Kass 2000).

H&S's approach has some limitations that restrict its use. Except for special cases, such as balanced single-error term ANOVA (Hodges et al. 2007), it is unclear how to allocate DF to individual effects. We use examples to illustrate this limitation and later show the advantage of removing it.

Example 1 (Revisited). The model for $\boldsymbol{\delta}$ can be reexpressed as $\mathbf{V}_1'\boldsymbol{\delta} \sim \mathrm{N}(\mathbf{0}, \tau_c^2\mathbf{D}_1^{-1})$, where $\mathbf{D}_1 = \mathrm{diag}(d_1, \ldots, d_{N-1})$. The models for $\boldsymbol{\delta}$ and $\boldsymbol{\xi}$ are then combined as in (1),

$$\begin{pmatrix}\mathbf{y}\\ \mathbf{0}\\ \mathbf{0}\end{pmatrix} = \begin{pmatrix}\mathbf{I} & \mathbf{I}\\ -\mathbf{V}_1' & \mathbf{0}\\ \mathbf{0} & -\mathbf{I}\end{pmatrix}\begin{pmatrix}\boldsymbol{\delta}\\ \boldsymbol{\xi}\end{pmatrix} + \begin{pmatrix}\boldsymbol{\epsilon}\\ \boldsymbol{\phi}_c\\ \boldsymbol{\phi}_h\end{pmatrix}$$

for $\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I})$, $\boldsymbol{\phi}_c \sim \mathrm{N}(\mathbf{0}, \tau_c^2\mathbf{D}_1^{-1})$, and $\boldsymbol{\phi}_h \sim \mathrm{N}(\mathbf{0}, \tau_h^2\mathbf{I}_N)$. H&S's DF is then (Lu, Hodges, and Carlin 2007)

$$\rho = \mathrm{tr}\left(\begin{pmatrix}\mathbf{I}_N & \mathbf{I}_N\\ \mathbf{I}_N & \mathbf{I}_N\end{pmatrix}\begin{pmatrix}\mathbf{I}_N + \frac{\sigma^2}{\tau_c^2}\mathbf{Q} & \mathbf{I}_N\\ \mathbf{I}_N & (1+\frac{\sigma^2}{\tau_h^2})\mathbf{I}_N\end{pmatrix}^{-1}\right).$$

We have assumed that $\mathbf{Q}$ describes a connected map, so $G = 1$. After some labor, ρ becomes

$$\rho = 1 + \sum_{i=1}^{N-1}\frac{(\tau_c^2/\sigma^2)/d_i + (\tau_h^2/\sigma^2)}{(\tau_c^2/\sigma^2)/d_i + (\tau_h^2/\sigma^2) + 1}.$$

While this suggests how total DF in the fit might be attributed to $\boldsymbol{\delta}$ and $\boldsymbol{\xi}$, any such partition must be rigorously justified. We do this in Section 2.5.

Example 2 (Global mean surface temperature: dynamic linear growth model). Summary measures of the earth's surface temperature are used to describe global climate. Consider the series y_t, $t = 0, \ldots, 124$, of global average temperature deviations (units 0.01 degrees Celsius) from 1881 to 2005. Figure 2 plots y_t, available at *http://data.giss.nasa.gov/gistemp/tabledata/GLB.Ts.txt*; Section 3.2 explains the rest of the plot. We smooth y_t using the linear growth model, which captures variation using time-varying mean and trend and independent noise (West and Harrison 1997, chap. 2). This model has equations for observation error, variation in local mean, and variation in trend,

$$\begin{aligned} y_t &= \mu_t + n_t, \qquad t = 0, \ldots, T,\\ \mu_t &= \mu_{t-1} + \beta_{t-1} + w_{1,t}, \qquad t = 1, \ldots, T,\\ \beta_t &= \beta_{t-1} + w_{2,t}, \qquad t = 1, \ldots, T-1. \end{aligned}$$

Let $\boldsymbol{\mu}, \boldsymbol{\beta}, \mathbf{n}, \mathbf{w}_1$, and $\mathbf{w}_2$ be the vectors of $\mu_t, \beta_t, n_t, w_{1t}$, and w_{2t}, respectively. Assume that $\mathbf{n}$, $\mathbf{w}_1$, and $\mathbf{w}_2$ are mutually independent and assume $\mathbf{n} \sim \mathrm{N}(\mathbf{0}, \sigma_n^2\mathbf{I}_{T+1})$, $\mathbf{w}_1 \sim \mathrm{N}(\mathbf{0}, \sigma_1^2\mathbf{I}_T)$, and $\mathbf{w}_2 \sim \mathrm{N}(\mathbf{0}, \sigma_2^2\mathbf{I}_{T-1})$, so σ_n^2, σ_1^2, and σ_2^2 describe the size of observational error n_t and the smoothness of level μ_t and trend β_t. The equation for β_t is the simplest CAR model, and the model for updating μ_t is similar; thus the variances σ_1^2 and σ_2^2 play roles like those of the variances in Example 1.

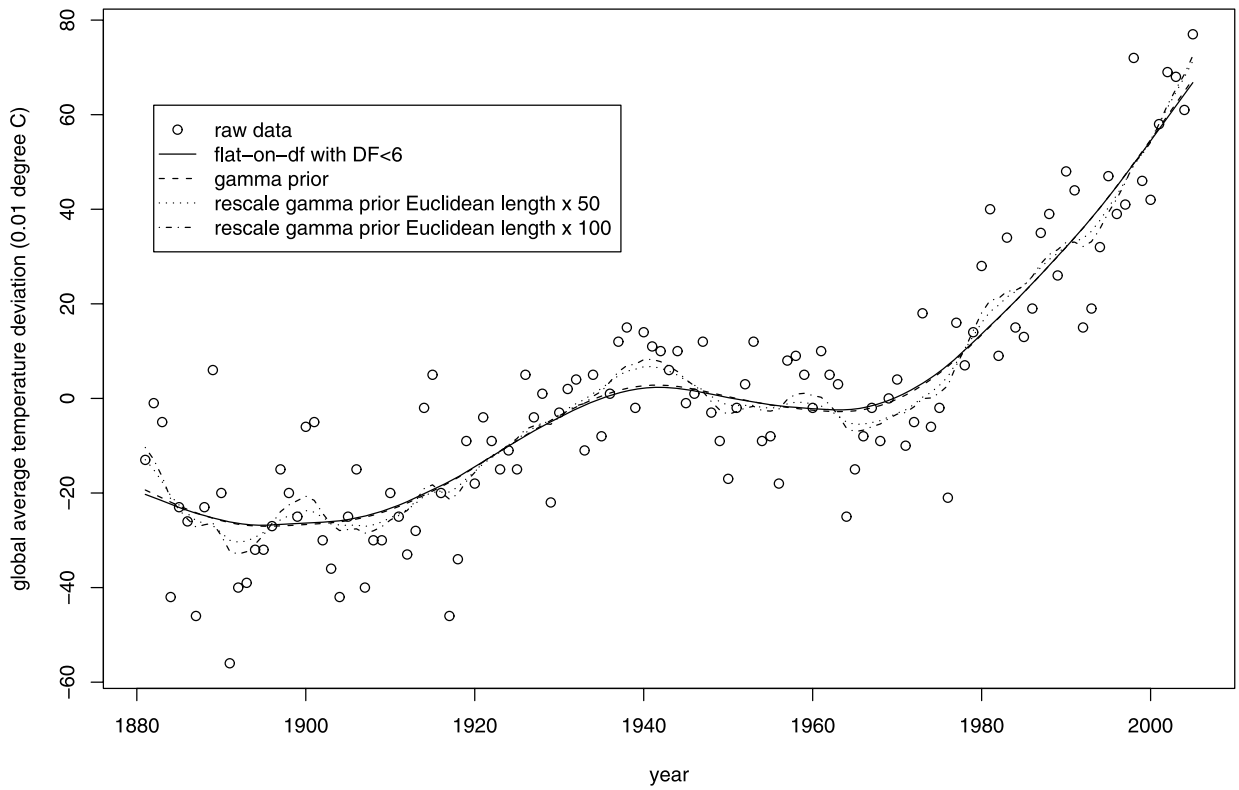

Figure 2. Example 2, global mean surface data, data and fitted smooths for gamma priors.

Following H&S, combine the three equations and reformulate them as a linear model:

$$\begin{pmatrix} \mathbf{y}_{(T+1)\times 1} \\ \mathbf{0}_{T\times 1} \\ \mathbf{0}_{(T-1)\times 1} \end{pmatrix} = \begin{pmatrix} \mathbf{I}_{T+1} & \mathbf{0}_{(T+1)\times T} \\ \mathbf{Z}_1 & -\mathbf{I}_T \\ \mathbf{0}_{(T-1)\times (T+1)} & \mathbf{Z}_2 \end{pmatrix} \begin{pmatrix} \boldsymbol{\mu} \\ \boldsymbol{\beta} \end{pmatrix} + \begin{pmatrix} \mathbf{n} \\ \mathbf{w}_1 \\ \mathbf{w}_2 \end{pmatrix},$$

where $\mathbf{Z}_1 = [\mathbf{0}_{T\times 1}; \mathbf{I}_T] - [\mathbf{I}_T; \mathbf{0}_{T\times 1}]$, $\mathbf{Z}_2 = [\mathbf{0}_{(T-1)\times 1}; \mathbf{I}_{T-1}] - [\mathbf{I}_{T-1}; \mathbf{0}_{(T-1)\times 1}]$. Then

$$\rho = \operatorname{tr}\left[\begin{pmatrix} \mathbf{I}_{T+1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \times \begin{pmatrix} \mathbf{I}_{T+1} + \frac{\sigma_n^2}{\sigma_1^2}\mathbf{Z}_1'\mathbf{Z}_1 & -\frac{\sigma_n^2}{\sigma_1^2}\mathbf{Z}_1' \\ -\frac{\sigma_n^2}{\sigma_1^2}\mathbf{Z}_1 & \frac{\sigma_n^2}{\sigma_1^2}\mathbf{I}_T + \frac{\sigma_n^2}{\sigma_2^2}\mathbf{Z}_2'\mathbf{Z}_2 \end{pmatrix}^{-1} \right].$$

This is difficult to simplify and thus offers little intuition about how to attribute DF separately to the local mean and the local linear trend (but again, see Sec. 2.5).

Usually, as in these examples, hierarchical models specify multiple sources of variation in the data. The first example has three: spatial clustering, local heterogeneity, and error. In the second example, variation comes from perturbations in the trend and local mean and from observation error. In these examples, it would be desirable to attribute DF to these meaningful effects. Such a partition of DF has at least two uses. First, it is an interpretable measure of each effect's share of the fit's complexity. Effects compete to explain variation when their design matrices are not orthogonal. In the absence of smoothing, this produces well-known collinearity effects; smoothing complicates the situation. Smoothed effects compete with unsmoothed effects (e.g., Reich, Hodges, and Zadnik 2006) and with one another, which influences the severity of collinearity effects and the degree to which smoothed effects are made smooth. Allocating a fit's DF to effects provides a view of this competition. For the model in Example 1, Best et al. (1999) used $\mathrm{SD}(\delta_i)/(\mathrm{SD}(\xi_i) + \mathrm{SD}(\delta_i))$ as a measure of clustering's contribution to model fit, where "SD" refers to the standard deviation of the δ_i and ξ_i themselves, not to τ_e or τ_s. In a Markov chain Monte Carlo (MCMC), this quantity's posterior distribution is estimated by computing the above ratio at each cycle of the MCMC using the current draws of the δ_i and ξ_i. This ratio is odd and hard to interpret; it would be more natural if variances were used instead of standard deviations, as in the intra-class correlation. Even then, however, it would depend on the scaling of the design matrices for the δ_i and ξ_i. [In defense of Best et al. (1999), it appears they merely intended to describe an aspect of their simulation results, and that this measure has taken on a life of its own.] In contrast, DF is independent of scaling (see Sec. 2.4) and directly describes complexity in the fitted values arising from the two parts of the model, with separate DF for clustering (DF_c) and heterogeneity (DF_h). The change from prior to posterior in the distribution of $\mathrm{DF}_c/(\mathrm{DF}_c + \mathrm{DF}_h)$ demonstrates the data's evidence regarding the proportion of variation explained by clustering. Section 3.1 explores this.

Second, a partition of DF can be useful for specifying a prior distribution on variance parameters. For the linear growth model, West and Harrison (1997, chap. 2) and Prado, West, and Krystal (2001) assumed fixed σ_1^2/σ_n^2 and σ_2^2/σ_n^2 to control complexity in the fit. It would be handy to avoid fixing these ratios, but because they are hard to conceptualize, specifying priors for them is difficult. A prior on interpretable DF avoids this difficulty.

In this article we propose a new definition of DF, consistent with that of H&S, which attributes a fit's total DF to individual effects. This definition involves a novel conception of an effect's DF as, loosely speaking, the effect's fraction of the model's total variance. The new definition applies to any model with a mixed-effect form and normal distributions for random effects and errors, for example, penalized splines in the mixed-model representation (Ruppert, Wand, and Carroll 2003) and at least some Gaussian process models. Section 2 defines the class of models, then gives and rationalizes the new definition. Section 3 applies this definition to the examples and explores its uses. The new definition is consistent with that given by Ruppert, Wand, and Carroll (2003, sec. 8.3) except for DF for residuals (Ruppert, Wand, and Carroll 2003, sec. 3.14). Ruppert, Wand, and Carroll (2003) also gave an approximation to an effect's DF, noting that "for all examples we have observed ... there is practically no difference between the approximate and exact degrees-of-freedom values" (p. 176). Section 3 gives an example in which the approximation differs substantially from the exact DF. Section 4 considers DF for residuals, in particular why the most common formula (Ruppert, Wand, and Carroll 2003, sec. 3.14), when added to the DF in the fitted model, gives a total less than the sample size, an oddity avoided by the new definition. Section 5 concludes. Technical results are provided in the Appendix, which is available online with the supplemental materials.

Model comparison criteria generally penalize a measure of model fit with a measure of model complexity. For example, Vaida and Blanchard (2005) derived a penalty based on ρ for a conditional Akaike information criterion for linear mixed models. In defining the deviance information criterion (DIC), Spiegelhalter et al. (2002) defined their penalty, the effective number of parameters p_D, in terms of a measure of model fit. However, in the light of Plummer (2008), it is not clear that the "right" penalty for using the same data to fit and evaluate a model is a simple function of any measure of model complexity, and thus describing model complexity and comparing models are distinct problems. We defer further consideration of model comparison, DIC, and p_D to Section 5.

2. THE NEW DEFINITION OF DF

2.1 Model Specification and Notation

Consider a linear model, which we call "the standard model," written as

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\theta}_1 + \mathbf{X}_2\boldsymbol{\theta}_2 + \boldsymbol{\epsilon}, \tag{2}$$

where $\mathbf{y} \in R^n$ contains the observations, $\mathbf{X}_1 \in R^{n\times p}$ is the design matrix for fixed effects (FE) that are not smoothed, and $\mathbf{X}_2 \in R^{n\times L}$ is the design matrix for smoothed effects, including random effects, effects representing spatial clustering, and so on. Vectors $\boldsymbol{\theta}_1 \in R^p$ and $\boldsymbol{\theta}_2 \in R^L$ are mean-structure parameters. Partition $\mathbf{X}_2$'s columns as $\mathbf{X}_2 = [\mathbf{X}_{21}, \ldots, \mathbf{X}_{2l}]$, $l \le L$, and conformably partition $\boldsymbol{\theta}_2 = (\boldsymbol{\theta}_{21}', \ldots, \boldsymbol{\theta}_{2l}')'$. Model the jth cluster of smoothed parameters $\boldsymbol{\theta}_{2j}$ as $\boldsymbol{\theta}_{2j}|\boldsymbol{\Gamma}_{2j} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Gamma}_{2j})$, with the

$\boldsymbol{\theta}_{2j}|\boldsymbol{\Gamma}_{2j}$ mutually independent; define $\boldsymbol{\Gamma}_2$ as the block diagonal matrix with the $\boldsymbol{\Gamma}_{2j}$ being the blocks on the diagonal. The normally distributed error $\boldsymbol{\epsilon}$ has mean $\mathbf{0}$ and covariance $\boldsymbol{\Gamma}_0$. $\boldsymbol{\Gamma}_0$ and $\boldsymbol{\Gamma}_{2j}$ are nonnegative definite and not necessarily proportional to the identity matrix or even diagonal. As the examples indicate, this is a large class of models. It includes Gaussian process models that can be written as $\mathbf{y} = \mathbf{X}_1\boldsymbol{\theta}_1 + \mathbf{S} + \boldsymbol{\epsilon}$, where $\mathbf{X}_1\boldsymbol{\theta}_1$ represents regressors, $\mathbf{X}_2 = \mathbf{I}_n$, $\boldsymbol{\theta}_2 = \mathbf{S} \in R^n$ with $\boldsymbol{\Gamma}_2 = \operatorname{cov}(\mathbf{S})$ nondiagonal and usually a function of unknowns, and $\operatorname{cov}(\boldsymbol{\epsilon})$ is diagonal.

Further notation includes $R(\mathbf{X})$ and $N(\mathbf{X})$ for the column and null spaces of the matrix $\mathbf{X}$ and $P_{\mathbf{X}}$ for the orthogonal projection onto $R(\mathbf{X})$. Appendix (a) gives necessary facts about the Moore–Penrose generalized inverse, denoted by $\mathbf{X}^+$.

2.2 Motivation for the New Definition

First, consider a linear model that is simpler than the standard model (2),

$$\mathbf{y} = \mathbf{X}_2\boldsymbol{\theta} + \boldsymbol{\epsilon}, \tag{3}$$

where $\boldsymbol{\theta}$ is smoothed using the model or prior $\boldsymbol{\theta} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Gamma}_2)$, and $\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Gamma}_0)$, for positive definite $\boldsymbol{\Gamma}_0$, $\boldsymbol{\Gamma}_2$. H&S wrote the model on $\boldsymbol{\theta}$ as "constraint cases," and combined it with (3) to give

$$\mathbf{y} = \begin{pmatrix} \mathbf{y} \\ \mathbf{0} \end{pmatrix} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}^* = \begin{pmatrix} \mathbf{X}_2 \\ \mathbf{I} \end{pmatrix} \boldsymbol{\theta} + \begin{pmatrix} \boldsymbol{\epsilon} \\ \boldsymbol{\xi} \end{pmatrix},$$

where $\boldsymbol{\epsilon}^*$ is multinormal with mean $\mathbf{0}$ and block diagonal covariance $\boldsymbol{\Gamma} = \operatorname{diag}(\boldsymbol{\Gamma}_0, \boldsymbol{\Gamma}_2)$. H&S's DF is

$$\begin{aligned} \rho &= \operatorname{tr}[\mathbf{X}_2(\mathbf{X}'\boldsymbol{\Gamma}^{-1}\mathbf{X})^{-1}\mathbf{X}_2'\boldsymbol{\Gamma}_0^{-1}] \\ &= \operatorname{tr}[\mathbf{X}_2(\mathbf{X}_2'\boldsymbol{\Gamma}_0^{-1}\mathbf{X}_2 + \boldsymbol{\Gamma}_2^{-1})^{-1}\mathbf{X}_2'\boldsymbol{\Gamma}_0^{-1}]. \end{aligned} \tag{4}$$

Rewrite the matrix inverse in (4) as (Schott 1997, thm. 1.7)

$$(\mathbf{X}_2'\boldsymbol{\Gamma}_0^{-1}\mathbf{X}_2 + \boldsymbol{\Gamma}_2^{-1})^{-1} = \boldsymbol{\Gamma}_2 - \boldsymbol{\Gamma}_2\mathbf{X}_2'(\mathbf{X}_2\boldsymbol{\Gamma}_2\mathbf{X}_2' + \boldsymbol{\Gamma}_0)^{-1}\mathbf{X}_2\boldsymbol{\Gamma}_2. \tag{5}$$

The left side of (5) is familiar from Bayesian linear models as the conditional posterior covariance of $\boldsymbol{\theta}|\mathbf{y}, \boldsymbol{\Gamma}$, while the right side is familiar as the conditional covariance of $\boldsymbol{\theta}$ given $\mathbf{y}$ from the joint multivariate normal distribution of $\boldsymbol{\theta}$ and $\mathbf{y}$. Use (5) to rewrite (4),

$$\begin{aligned} \rho &= \operatorname{tr}[\mathbf{X}_2\boldsymbol{\Gamma}_2\mathbf{X}_2'\boldsymbol{\Gamma}_0^{-1} \\ &\quad - \mathbf{X}_2\boldsymbol{\Gamma}_2\mathbf{X}_2'(\mathbf{X}_2\boldsymbol{\Gamma}_2\mathbf{X}_2' + \boldsymbol{\Gamma}_0)^{-1}\mathbf{X}_2\boldsymbol{\Gamma}_2\mathbf{X}_2'\boldsymbol{\Gamma}_0^{-1}] \\ &= \operatorname{tr}[\mathbf{X}_2\boldsymbol{\Gamma}_2\mathbf{X}_2'(\mathbf{X}_2\boldsymbol{\Gamma}_2\mathbf{X}_2' + \boldsymbol{\Gamma}_0)^{-1} \\ &\quad \times (\mathbf{X}_2\boldsymbol{\Gamma}_2\mathbf{X}_2' + \boldsymbol{\Gamma}_0 - \mathbf{X}_2\boldsymbol{\Gamma}_2\mathbf{X}_2')\boldsymbol{\Gamma}_0^{-1}] \\ &= \operatorname{tr}[\mathbf{X}_2\boldsymbol{\Gamma}_2\mathbf{X}_2'(\mathbf{X}_2\boldsymbol{\Gamma}_2\mathbf{X}_2' + \boldsymbol{\Gamma}_0)^{-1}], \end{aligned}$$

which has the form trace [ratio of {modeled variance matrix} to {total variance matrix}]. This suggests defining an effect-specific DF as that effect's contribution of variance to total variance. In the standard model (2), the mean-structure parameters $\boldsymbol{\theta}_{2j}$ are independent of each other conditional on their covariances $\boldsymbol{\Gamma}_{2j}$, and the variance in $\mathbf{y}$ arising from the effect represented by $\boldsymbol{\theta}_{2j}$ is $\mathbf{X}_{2j}\boldsymbol{\Gamma}_{2j}\mathbf{X}_{2j}'$. Along with variation from the error $\boldsymbol{\epsilon}$, the total modeled variance for $\mathbf{y}$ is $\sum_1^l \mathbf{X}_{2j}\boldsymbol{\Gamma}_{2j}\mathbf{X}_{2j}' + \boldsymbol{\Gamma}_0 = \mathbf{X}_2\boldsymbol{\Gamma}_2\mathbf{X}_2' + \boldsymbol{\Gamma}_0$. Thus, in this special case when all effects are smoothed ($\boldsymbol{\theta}_1$ is null) and all $\boldsymbol{\Gamma}_j$ are positive definite, the reformulated DF for $\mathbf{X}_{2j}$ is $\operatorname{tr}[\mathbf{X}_{2j}\boldsymbol{\Gamma}_{2j}\mathbf{X}_{2j}'(\mathbf{X}_2\boldsymbol{\Gamma}_2\mathbf{X}_2' + \boldsymbol{\Gamma}_0)^{-1}]$.

Green (2002) used a different approach to derive an equivalent decomposition of the effective number of parameters p_D when $\boldsymbol{\Gamma}_0$ and $\boldsymbol{\Gamma}_2$ are completely specified, in which case p_D is identical to H&S's DF. But Green's derivation does not appear to extend to cases in which $\boldsymbol{\Gamma}_0$ or $\boldsymbol{\Gamma}_2$ are functions of unknowns. The development above and below is reminiscent of that of Gelman and Pardoe (2006), who developed measures of explained variance and pooling for each of the effects in a model (which they call "levels"), although their intent and results were rather different.

2.3 Definition of DF

Section 2.2's reformulation of ρ for model (3), with all effects smoothed, suggests a way to rewrite H&S's DF as a sum of meaningful quantities. However, for the standard model (2), the new formulation must accommodate nonsmoothed effects $\mathbf{X}_1$. A nonsmoothed effect can be viewed as the limit of a smoothed effect, for which the prior covariance goes to infinity and imposes no constraint. The new definition uses $\lambda\boldsymbol{\Gamma}_1 \in R^{n\times n}$ as $\boldsymbol{\theta}_1$'s covariance matrix, where $\boldsymbol{\Gamma}_1$ is unspecified but positive definite and λ is a positive scalar. In the limit, as λ goes to $+\infty$, $\boldsymbol{\Gamma}_1$ disappears. For the standard model (2), then, define DF for nonsmoothed effects, denoted by $\mathrm{DF}(\mathbf{X}_1)$, as

$$\begin{aligned} \mathrm{DF}(\mathbf{X}_1) &= \lim_{\lambda\to+\infty} \operatorname{tr}[\mathbf{X}_1\lambda\boldsymbol{\Gamma}_1\mathbf{X}_1' \\ &\qquad \times (\mathbf{X}_1\lambda\boldsymbol{\Gamma}_1\mathbf{X}_1' + \mathbf{X}_2\boldsymbol{\Gamma}_2\mathbf{X}_2' + \boldsymbol{\Gamma}_0)^+] \\ &= \operatorname{tr}[P_{\mathbf{X}_1}] = \operatorname{rank}(\mathbf{X}_1). \end{aligned}$$

The proof is provided in Appendix (b), which, like all other Appendices given later, is available online as supplemental material. For the smoothed effect $\boldsymbol{\theta}_{2j}$, DF is defined analogously,

$$\begin{aligned} \mathrm{DF}(\mathbf{X}_{2j}) &= \lim_{\lambda\to+\infty} \operatorname{tr}[\mathbf{X}_{2j}\boldsymbol{\Gamma}_{2j}\mathbf{X}_{2j}'(\mathbf{X}_1\lambda\boldsymbol{\Gamma}_1\mathbf{X}_1' + \mathbf{X}_2\boldsymbol{\Gamma}_2\mathbf{X}_2' + \boldsymbol{\Gamma}_0)^+] \\ &= \operatorname{tr}\big\{\mathbf{X}_{2j}\boldsymbol{\Gamma}_{2j}\mathbf{X}_{2j}' \\ &\qquad \times [(\mathbf{I} - P_{\mathbf{X}_1})(\mathbf{X}_2\boldsymbol{\Gamma}_2\mathbf{X}_2' + \boldsymbol{\Gamma}_0)(\mathbf{I} - P_{\mathbf{X}_1})]^+\big\}. \end{aligned}$$

The proof is provided in Appendix (c). Similarly, for the error term $\boldsymbol{\epsilon}$,

$$\begin{aligned} \mathrm{DF}(\boldsymbol{\epsilon}) &= \lim_{\lambda\to+\infty} \operatorname{tr}[\mathbf{I}\boldsymbol{\Gamma}_0\mathbf{I}'(\mathbf{X}_1\lambda\boldsymbol{\Gamma}_1\mathbf{X}_1' + \mathbf{X}_2\boldsymbol{\Gamma}_2\mathbf{X}_2' + \boldsymbol{\Gamma}_0)^+] \\ &= \operatorname{tr}\big\{\boldsymbol{\Gamma}_0[(\mathbf{I} - P_{\mathbf{X}_1})(\mathbf{X}_2\boldsymbol{\Gamma}_2\mathbf{X}_2' + \boldsymbol{\Gamma}_0)(\mathbf{I} - P_{\mathbf{X}_1})]^+\big\}. \end{aligned}$$

This follows by the argument given in Appendix (c). In general, this differs from residual DF as defined by Ruppert, Wand, and Carroll (2003, sec. 3.14) or Hastie and Tibshirani (1990, sec. 3.5).

Based on this definition, the DF of a smoothed effect $\mathbf{X}_{2j}$ or error $\boldsymbol{\epsilon}$ is the fraction of variation contributed by $\mathbf{X}_{2j}$ or $\boldsymbol{\epsilon}$ out of variation not accounted for by the unsmoothed effects $\mathbf{X}_1$. Because computing $\mathrm{DF}(\mathbf{X}_{2j})$ can be laborious, Cui (2008, p. 79) derived the approximation $\mathrm{DF}_p(\mathbf{X}_{2j}) = \operatorname{tr}\{\mathbf{X}_{2j}\boldsymbol{\Gamma}_{2j}\mathbf{X}_{2j}'[(\mathbf{I} - P_{\mathbf{X}_1})(\mathbf{X}_{2j}\boldsymbol{\Gamma}_{2j}\mathbf{X}_{2j}' + \boldsymbol{\Gamma}_0)(\mathbf{I} - P_{\mathbf{X}_1})]^+\}$, which is equivalent to the approximation of Ruppert, Wand, and Carroll (2003, pp. 175–176) when both are defined. In Section 3.2 we briefly consider this approximation's accuracy.

2.4 Properties of the Definition

We argue for the new definition's validity by presenting four of its properties:

(DF.a) The new definition is consistent with H&S's DF; the sum of DF over all effects equals H&S's ρ for the whole model. Appendix (d) gives a proof.

(DF.b) The sum of DF in the model fit and error is fixed. When $\boldsymbol{\Gamma}_0$ is positive definite, $\mathrm{DF}(\mathbf{X}_1)+\sum_1^J \mathrm{DF}(\mathbf{X}_{2j})+\mathrm{DF}(\boldsymbol{\epsilon}) = \dim(\mathbf{y})$, the dimension of $\mathbf{y}$. Appendix (e) gives a proof. For the definition of DF in residuals given by Ruppert, Wand, and Carroll (2003) and Hastie and Tibshirani (1990), $\mathrm{DF}(\mathbf{X}_1) + \sum_1^J \mathrm{DF}(\mathbf{X}_{2j}) + \mathrm{DF}(\text{residuals}) < \dim(\mathbf{y})$, with the deficiency arising from DF(residuals).

(DF.c) The DF of an effect has a reasonable upper bound, $\mathrm{DF}(\mathbf{X}_{2j}) \le \mathrm{rank}((\mathbf{I}-P_{\mathbf{X}_1})\mathbf{X}_{2j}) = \mathrm{rank}([\mathbf{X}_1, \mathbf{X}_{2j}]) - \mathrm{rank}(\mathbf{X}_1) \le \mathrm{rank}(\mathbf{X}_{2j})$ for $j = 1, \ldots, J$. Appendix (f) gives a proof.

(DF.d) Scale-invariance property. DF, and particularly priors on DF, avoid problems arising from scaling of columns of the design matrix. (We discuss prior distribution on DF in Sec. 2.6.) Suppose that in the standard model (2), the covariance of each smoothed effect $\boldsymbol{\theta}_{2j}$ and the covariance of the error term $\boldsymbol{\epsilon}$ are both characterized by a single parameter, $\boldsymbol{\Gamma}_{2j} = \sigma_j^2 \boldsymbol{\Gamma}_{2j}^0$ and $\boldsymbol{\Gamma}_0 = \sigma_0^2 \boldsymbol{\Gamma}_0^0$, where the σ_j^2 are unknown scalars and $\boldsymbol{\Gamma}_{2j}^0$, $\boldsymbol{\Gamma}_0^0$ are known and positive definite. Let $\mathbf{B} \in R^{p\times p}$ be nonsingular, and let $\mathbf{H}_j$ be a matrix such that $\mathbf{H}_j \boldsymbol{\Gamma}_{2j}^0 \mathbf{H}_j' = \boldsymbol{\Gamma}_{2j}^0$, for example, $\boldsymbol{\Gamma}_{2j}^0$ is the identity and $\mathbf{H}_j$ is orthogonal. Then the posterior of $\mathbf{X}\boldsymbol{\theta}$ arising from independent priors on $(\mathrm{DF}(\mathbf{X}_{21}), \ldots, \mathrm{DF}(\mathbf{X}_{2l}))$ and σ_0^2 is the same when $\mathbf{X}_1$ is transformed to $\mathbf{X}_1^* = \mathbf{X}_1\mathbf{B}$, and $\mathbf{X}_2$ is transformed so that $\mathbf{X}_{2j}^* = t_j \mathbf{X}_{2j}\mathbf{H}_j$ for nonzero scalars t_j. Appendix (g) gives a proof.

2.5 The New Definition Applied to the Examples

Example 1 (Model with clustering and heterogeneity). Assume that the spatial map represented by $\mathbf{Q}$ is connected, so $G = 1$. (This is not necessary.) The CAR model for the spatial clustering effects $\boldsymbol{\delta}$ models $\mathbf{V}_1'\boldsymbol{\delta}$ as $\mathrm{N}(\mathbf{0}, \tau_c^2 \mathbf{D}_1^{-1})$, while $\mathbf{1}_N'\boldsymbol{\delta}$ is not smoothed; thus reexpress this model in the standard form (2) by taking $\mathbf{X}_1 = \mathbf{1}_N, \mathbf{X}_{21} = \mathbf{V}_1, \mathbf{X}_{22} = \mathbf{I}_N$, and $\boldsymbol{\theta}_1 = \mathbf{1}_N'\boldsymbol{\delta}/N, \boldsymbol{\theta}_{21} = \mathbf{V}_1'\boldsymbol{\delta}, \boldsymbol{\theta}_{22} = \boldsymbol{\xi}$, $\boldsymbol{\Gamma}_{21} = \tau_c^2 \mathbf{D}_1^{-1}, \boldsymbol{\Gamma}_{22} = \tau_h^2 \mathbf{I}_N$, and $\boldsymbol{\Gamma}_0 = \sigma^2 \mathbf{I}_N$. The constraint cases become $\boldsymbol{\theta}_{21} \sim \mathrm{N}(\mathbf{0}, \tau_c^2 \mathbf{D}_1^{-1})$ and $\boldsymbol{\theta}_{22} \sim \mathrm{N}(\mathbf{0}, \tau_h^2 \mathbf{I}_N)$. Appendix (h) derives these expressions for $\mathrm{DF}(\boldsymbol{\delta})$ and $\mathrm{DF}(\boldsymbol{\xi})$:

$$\begin{aligned}\mathrm{DF}(\boldsymbol{\delta}) &= \mathrm{rank}(\mathbf{X}_1) + \mathrm{tr}\big\{\mathbf{X}_{21}\boldsymbol{\Gamma}_{21}\mathbf{X}_{21}'[(\mathbf{I}_N - P_{\mathbf{X}_1}) \\ &\quad \times (\mathbf{X}_{21}\boldsymbol{\Gamma}_{21}\mathbf{X}_{21}' + \mathbf{X}_{22}\boldsymbol{\Gamma}_{22}\mathbf{X}_{22}' + \boldsymbol{\Gamma}_0)(\mathbf{I}_N - P_{\mathbf{X}_1})]^+\big\} \\ &= 1 + \sum_{i=1}^{N-1} \frac{(\tau_c^2/\sigma^2)/d_i}{(\tau_c^2/\sigma^2)/d_i + (\tau_h^2/\sigma^2) + 1}; \\ \mathrm{DF}(\boldsymbol{\xi}) &= \mathrm{tr}\big\{\mathbf{X}_{22}\boldsymbol{\Gamma}_{22}\mathbf{X}_{22}'[(\mathbf{I}_N - P_{\mathbf{X}_1}) \\ &\quad \times (\mathbf{X}_{21}\boldsymbol{\Gamma}_{21}\mathbf{X}_{21}' + \mathbf{X}_{22}\boldsymbol{\Gamma}_{22}\mathbf{X}_{22}' + \boldsymbol{\Gamma}_0)(\mathbf{I}_N - P_{\mathbf{X}_{\mathbf{1}_N}})]^+\big\} \\ &= \sum_{i=1}^{N-1} \frac{(\tau_h^2/\sigma^2)}{(\tau_c^2/\sigma^2)/d_i + (\tau_h^2/\sigma^2) + 1}.\end{aligned}$$

Thus $\mathrm{DF}(\boldsymbol{\delta}) + \mathrm{DF}(\boldsymbol{\xi})$ is H&S's DF. The DF of each component is a function of *both* variance ratios $(\tau_c^2/\sigma^2, \tau_h^2/\sigma^2)$, which sheds light on how the effects compete. If τ_c^2/σ^2 increases for fixed τ_h^2/σ^2, then $\mathrm{DF}(\boldsymbol{\delta})$ increases and $\mathrm{DF}(\boldsymbol{\xi})$ decreases. When there is more than one island ($G > 1$), it is easy to show that the DF for clustering and heterogeneity are the sums of the respective DF for each island.

Example 2 (Linear growth model). Write this model as a linear model in μ_0, the initial local mean; β_0, the initial local trend; $\mathbf{w}_1$, noise in the local mean; and $\mathbf{w}_2$, noise in the trend:

$$\begin{aligned}\begin{pmatrix} y_0 \\ y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_T \end{pmatrix} &= \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ \vdots & \vdots \\ 1 & T \end{pmatrix} \begin{pmatrix} \mu_0 \\ \beta_0 \end{pmatrix} \\ &\quad + \begin{pmatrix} 0 & 0 & \cdots & & & 0 \\ 1 & 0 & \cdots & & & 0 \\ 1 & 1 & 0 & \cdots & & 0 \\ 1 & 1 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & & & \ddots & \vdots \\ 1 & 1 & 1 & 1 & \cdots & 1 \end{pmatrix} \begin{pmatrix} w_{11} \\ w_{12} \\ \vdots \\ w_{1T} \end{pmatrix} \\ &\quad + \begin{pmatrix} 0 & 0 & \cdots & & & 0 \\ 0 & 0 & \cdots & & & 0 \\ 1 & 0 & 0 & \cdots & & 0 \\ 2 & 1 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & & \ddots & \vdots \\ T-1 & T-2 & T-3 & T-4 & \cdots & 1 \end{pmatrix} \\ &\quad \times \begin{pmatrix} w_{21} \\ w_{22} \\ \vdots \\ w_{2T-1} \end{pmatrix} + \begin{pmatrix} n_0 \\ n_1 \\ \vdots \\ n_T \end{pmatrix}.\end{aligned}$$

In terms of the standard model, the unsmoothed effects are $\boldsymbol{\theta}_1 = (\mu_0, \beta_0)'$, the smoothed effects are $\boldsymbol{\theta}_2 = (\mathbf{w}_1', \mathbf{w}_2')'$, the smoothing covariances are $\boldsymbol{\Gamma}_{21} = \sigma_1^2 \mathbf{I}_T$ and $\boldsymbol{\Gamma}_{22} = \sigma_2^2 \mathbf{I}_{T-1}$, and the error covariance is $\boldsymbol{\Gamma}_0 = \sigma_n^2 \mathbf{I}_{T+1}$. $\mathrm{DF}(\mathbf{w}_1), \mathrm{DF}(\mathbf{w}_2)$, and $\mathrm{DF}(\mathbf{n})$ follow straightforwardly, but the expressions do not simplify, and thus we omit them and illustrate their properties with some special cases.

When the local mean and trend do not vary ($\sigma_1^2 = \sigma_2^2 = 0$), the model reduces to a linear regression with intercept μ_0 and slope β_0, with 2 DF. Thus when $\sigma_1^2 > 0$ or $\sigma_2^2 > 0$, $\mathrm{DF}(\mathbf{w}_1)$ and $\mathrm{DF}(\mathbf{w}_2)$ describe complexity in the fit, beyond a linear trend, attributable to these two sources. Note that the matrix $(\mathbf{X}_1|\mathbf{X}_{21})$ is saturated, and $\mathbf{X}_{22} = \mathbf{X}_{21}\mathbf{B}$ for a specific $\mathbf{B}$. Thus if $\sigma_1^2 > 0$ and $\sigma_2^2 > 0$, then $\mathbf{w}_1$ and $\mathbf{w}_2$ compete with each other. Tables 1a and 1b show how they compete for various (σ_1^2, σ_2^2); noise variance σ_n^2 is fixed at 1, and T is fixed at 124. When $\sigma_2 = 0$ (Table 1a), $\mathrm{DF}(\mathbf{w}_1)$ increases as σ_1 grows, with analogous results when $\sigma_1 = 0$ and σ_2 grows. As for the clustering and heterogeneity model, for fixed $\sigma_1 > 0$ (Table 1b), $\mathrm{DF}(\mathbf{w}_1)$ is reduced as σ_2 grows, because a larger σ_2 allows $\mathbf{w}_2$ to compete more aggressively, and so $\mathrm{DF}(\mathbf{w}_2)$ grows. The analogous result occurs when σ_1 increases for fixed σ_2.

Table 1a. DF for $\mathbf{w}_1$ and $\mathbf{w}_2$ in the linear growth model in different scenarios ($\sigma_n = 1$): Assuming $\sigma_j = 0$ for $j = 1$ or 2

ϕ	DF($\mathbf{w}_1$) if $\sigma_1 = \phi, \sigma_2 = 0$	DF($\mathbf{w}_2$) if $\sigma_1 = 0, \sigma_2 = \phi$
0.01	0.1	3.4
0.1	4.8	13.1
1	54.3	47.4
10	120.6	116.4

Table 1b. DF for $\mathbf{w}_1$ and $\mathbf{w}_2$ in the linear growth model in different scenarios ($\sigma_n = 1$): Assuming various values of σ_2 for fixed σ_1

	$\sigma_1 = 0.1$		$\sigma_1 = 1$	
σ_2	DF($\mathbf{w}_1$)	DF($\mathbf{w}_2$)	DF($\mathbf{w}_1$)	DF($\mathbf{w}_2$)
0.01	3.3	2.8	54.1	0.2
0.1	1.3	12.8	49.5	5.2
1	0.4	47.2	27.6	38.1
10	0.02	116.4	2.2	114.4

2.6 Using Priors on DF to Induce Priors on Smoothing Parameters

The most familiar notion of DF is for linear models with one error term (e.g., Weisberg 1985), in which a model's DF is the fixed, known rank of its design matrix. As extended to scatterplot smoothers (e.g., Hastie and Tibshirani 1990), a fit's DF is not fixed and known before the model is fit, but rather is a function of tuning constants chosen to control the fit's smoothness. For DF as redefined here (e.g., in the model of Example 1), the vector of the effect-specific DF, $(\mathrm{DF}(\boldsymbol{\delta}), \mathrm{DF}(\boldsymbol{\xi}))$, is a one-to-one function of the vector of variance ratios $(\tau_c^2/\sigma^2, \tau_h^2/\sigma^2)$ and again is not fixed or known before the model is fit. Because of this one-to-one function, placing a prior distribution on $(\tau_c^2/\sigma^2, \tau_h^2/\sigma^2)$ induces a prior on $(\mathrm{DF}(\boldsymbol{\delta}), \mathrm{DF}(\boldsymbol{\xi}))$. But placing a prior on $(\mathrm{DF}(\boldsymbol{\delta}), \mathrm{DF}(\boldsymbol{\xi}))$ to induce a prior on the unknown variance ratios $(\tau_c^2/\sigma^2, \tau_h^2/\sigma^2)$ is equally legitimate. Such a prior has the advantage of being specified on the interpretable complexity measure instead of the usually more obscure variances or variance ratios. We can state this more precisely. In the standard model (2), suppose that each smoothed effect $\boldsymbol{\theta}_{2j}$ and the error term $\boldsymbol{\epsilon}$ have covariances characterized by one parameter, say $\boldsymbol{\Gamma}_{2j} = \sigma_j^2 \boldsymbol{\Gamma}_{2j}^0$ and $\boldsymbol{\Gamma}_0 = \sigma_0^2 \boldsymbol{\Gamma}_0^0$, where the σ_j^2 are unknown scalars and $\boldsymbol{\Gamma}_{2j}^0, \boldsymbol{\Gamma}_0^0$ are known and positive definite. Further assume that $\mathbf{X}_{2j} \not\subseteq R(\mathbf{X}_1)$. Then $\mathrm{DF}(\cdot)$ is a one-to-one mapping between $\mathbf{q} = (\mathrm{DF}(\mathbf{X}_{21}), \ldots, \mathrm{DF}(\mathbf{X}_{2l}))'$ on $\mathbf{q}$'s range and $\mathbf{s} = (\log(\sigma_1^2/\sigma_0^2), \ldots, \log(\sigma_l^2/\sigma_0^2)) \in R^l$, and thus a prior on $\mathbf{q}$ induces a unique prior on $\mathbf{s}$. Appendix (i) gives a proof and the Jacobian. By property (DF.d) in Section 2.4, under these assumptions, the posterior distribution of $\mathbf{X}_{2j}\boldsymbol{\theta}_{2j}$ using a prior on $\mathbf{q}$ is invariant under certain transformations of $\mathbf{X}_{2j}$. Sometimes it is desirable to put a prior on functions of $\mathbf{q}$, e.g., $\mathrm{DF}(\boldsymbol{\delta})/(\mathrm{DF}(\boldsymbol{\delta}) + \mathrm{DF}(\boldsymbol{\xi}))$ in Example 1 (discussed later). If an l-variate function of $\mathbf{q}$, $\mathbf{u} = (u_1(\mathbf{q}), \ldots, u_l(\mathbf{q}))$ is a one-to-one mapping, then a prior on $\mathbf{u}$ induces a unique prior on $\mathbf{s}$.

When $\boldsymbol{\Gamma}_{2j}$'s form is more complex than that assumed so far, a prior on DF partially specifies a prior on $\boldsymbol{\Gamma}_{2j}$, and a complete prior specification requires a prior on other functions of $\boldsymbol{\Gamma}_{2j}$. For example, if $\boldsymbol{\Gamma}_{2j} = \mathrm{diag}(\sigma_{j_1}^2, \sigma_{j_2}^2)$, then a uniform prior on $\mathrm{DF}(\mathbf{X}_{2j})$ induces a prior on a scalar function of $(\sigma_{j_1}^2/\sigma_0^2, \sigma_{j_2}^2/\sigma_0^2)$. A complete specification requires a prior on another function of $(\sigma_{j_1}^2/\sigma_0^2, \sigma_{j_2}^2/\sigma_0^2)$, for example, on $\sigma_{j_1}^2/\sigma_{j_2}^2$.

3. EXAMPLES: PRIORS ON DF, PARTITIONING DF, COMPETING EFFECTS

3.1 Periodontal Measurements: Clustering and Heterogeneity Model

Here we analyze AL data as described in Section 1 for 12 subjects without missing measurements taken from a larger data set of 410 subjects. The data set is available on line with the supplemental materials. Each site in each subject was measured once by each of two examiners; thus each subject provided 84 AL measurements on 42 sites. Zhao (1999) investigated differences between examiners using all 410 subjects and found tiny systematic differences between examiners. For each pair of examiners, the differences between the examiners' measurements showed small spatial correlation, and the standard deviation of the differences was similar for the different examiner pairs. Thus for our purposes here, we simply treat the two examinations on each patient as providing two iid measurements at each site.

Let y_{ijk} be the kth measurement of site j for subject i, $i = 1, \ldots, 12, j = 1, \ldots, 42, k = 1, 2$. Model y_{ijk} as $y_{ijk} = \mu + \alpha_i + \delta_{ij} + \xi_{ij} + \epsilon_{ijk}$, where μ is the grand mean and α_i is subject i's random effect, modeled as an independent $\mathrm{N}(0, \sigma_\alpha^2)$ draw. The vector $\boldsymbol{\delta}_i = (\delta_{i1}, \ldots, \delta_{i42})$ captures spatial clustering for subject i. We assume that these are independent across i and model each $\boldsymbol{\delta}_i$ as $\mathrm{CAR}(\mathbf{Q}, \sigma_{c,i}^2)$, with neighbor pairs as in Figure 1. We model heterogeneity effects, ξ_{ij} as $\mathrm{N}(0, \sigma_{h,i}^2)$, assumed to be iid within subject. The ϵ_{ijk} capture unsystematic measurement error and are modeled as independent $\mathrm{N}(0, \sigma_0^2)$, with σ_0^2 being common to all subjects.

Subject i has smoothing variances $\sigma_{c,i}^2$ and $\sigma_{h,i}^2$, and thus $\mathrm{DF}(\boldsymbol{\delta}_i)$ and $\mathrm{DF}(\boldsymbol{\xi}_i)$, allowing subject-specific inferences about the relative contribution to the fit of clustering and heterogeneity. An obvious prior for this purpose is a prior on $f_i = \mathrm{DF}(\boldsymbol{\delta}_i)/(\mathrm{DF}(\boldsymbol{\delta}_i) + \mathrm{DF}(\boldsymbol{\xi}_i))$, the fraction, for subject i, of the fitted complexity attributed to clustering. A flat prior on f_i demonstrates no preference between clustering and heterogeneity without constraining the total DF in subject i's fit, $\mathrm{DF}(\boldsymbol{\delta}_i) + \mathrm{DF}(\boldsymbol{\xi}_i)$, so it can be considered natural. We used independent flat priors on f_i, $\mathrm{DF}(\boldsymbol{\delta}_i) + \mathrm{DF}(\boldsymbol{\xi}_i)$, and $\mathrm{DF}(\boldsymbol{\alpha})$. Table 2 gives the posterior means of the f_i; the posterior medians are similar.

In Table 2, the overall DF for spatial clustering has a posterior mean of 82.00 out of 492, and subject-specific values are 1.66–16.75 out of 41. Overall DF for heterogeneity has a posterior mean of 158.26 out of 503; subject-specific values are 1.60–28.12 out of 42. The posterior mean of f_i ranges from 0.18 to 0.52, with a median of 0.36, compared with the prior mean of 0.50; thus for most subjects, the data indicate that heterogeneity should receive more fitted complexity than clustering. Compared with $\sigma_{c,i}/(\sigma_{c,i} + \sigma_{h,i})$ (Eberly and Carlin 2000), f_i has a direct interpretation and accounts for all other sources of variation; that is, because measurement error variance σ_0^2 is common

Table 2. Periodontal data: posterior mean DF for spatial clustering and site heterogeneity and posterior mean of f_i, the fraction of DF for spatial clustering

Source	DF($\boldsymbol{\delta}_i$)	DF($\boldsymbol{\xi}_i$)	$E(f_i\|\mathbf{y})$
Sum of subjects	82.00	158.26	0.34
Subject #11	6.32	28.12	0.18
Subject #3	7.74	27.24	0.22
Subject #9	1.88	4.15	0.31
Subject #2	1.66	3.44	0.33
Subject #5	10.10	20.14	0.33
Subject #6	10.34	19.06	0.35
Subject #12	11.53	20.29	0.36
Subject #8	5.52	8.08	0.41
Subject #4	16.75	18.07	0.48
Subject #1	1.84	1.89	0.49
Subject #10	1.76	1.60	0.52
Subject #7	6.57	6.18	0.52

NOTE: The maximum possible DF for $\boldsymbol{\delta}_i$ and $\boldsymbol{\xi}_i$ are 41 and 42 for individual subjects and 492 and 503 for all subjects combined.

to all subjects, f_i depends indirectly on $\sigma^2_{c,j}/\sigma^2_0$ and $\sigma^2_{h,j}/\sigma^2_0$ for $j \neq i$.

Are these results reasonable? Figure 3 shows the data and the posterior of f_i for subjects 11, 7, and 10, who have the smallest (subject 11) and the largest (subjects 7 and 10) $E(f_i|\mathbf{y})$. Based on Section 1's interpretation, when the DF are low for both clustering and heterogeneity, the data should have little large-scale trend or unpatterned noise respectively. When the DF are large for both, we should see large differences in the local level between different parts of the mouth, as well as sizeable unpatterned variation around that large-scale trend. When DF is large for heterogeneity but small for clustering, we should see little large-scale trend but much unpatterned variation. Subjects 11 and 7 have a similar posterior DF for clustering (about 6 DF), while subject 11 has a much larger DF for heterogeneity (about 28 DF). The data in Figure 3 indicate a similarly modest large-scale trend for these two subjects, but with much greater unpatterned variation for subject 11. Now compare subjects 7 and 10, who have the same $E(f_i|\mathbf{y})$, but subject 10 has a smaller posterior expected DF for both components by a factor of about 4. As expected, the data for subject 10 show much less trend and variation than the data for subject 7.

The random effects $\boldsymbol{\alpha}$, $\boldsymbol{\delta}$, and $\boldsymbol{\xi}$ compete with one another; the column spaces for $\boldsymbol{\alpha}$ and $\boldsymbol{\delta}$ are orthogonal to each other, but both are contained in the column space for $\boldsymbol{\xi}$. To give a taste of how this affects partitioning of DF, Table 3 shows results from simulated data like the periodontal data set analyzed earlier. For simplicity, all 12 simulated subjects have the same σ^2_c and σ^2_h in both the data and the model. We fixed σ^2_0 at 1 and sampled 100 artificial data sets from each of 6 sets of true variance ratios $(\sigma^2_\alpha/\sigma^2_0, \sigma^2_c/\sigma^2_0, \sigma^2_h/\sigma^2_0)$, drawing new α_i, $\boldsymbol{\delta}_i$, and $\boldsymbol{\xi}_i$ for each artificial data set. The analyses used a flat prior on $(\text{DF}(\boldsymbol{\alpha}), \text{DF}(\boldsymbol{\delta}), \text{DF}(\boldsymbol{\xi}))$; flat priors on $\text{DF}(\boldsymbol{\alpha}), \text{DF}(\boldsymbol{\delta}) + \text{DF}(\boldsymbol{\xi})$, and $\text{DF}(\boldsymbol{\delta})/(\text{DF}(\boldsymbol{\delta}) + \text{DF}(\boldsymbol{\xi}))$ give similar results. The DF allocated to subjects, DF($\boldsymbol{\alpha}$), are similar in all cases considered, and we do not discuss this further here.

In Table 3, consider the columns labeled "True DF"; these values are DF as a function of the true variances. The model with both clustering and heterogeneity [rows (1, 1, 1) and (0.25, 0.25, 0.25)] allocates less complexity to clustering than the model with only clustering [rows (1, 1, 0) and (0.25, 0.25, 0)]. The presence of heterogeneity reduces the true DF for clustering to a proportionally greater extent when the true variances are 1 (42% reduction) than when they are 0.25 (21%). The presence of clustering has a similar effect on the true DF for heterogeneity (29% and 19% reduction, respectively). In other words, the two effects compete but to a lesser degree when each is constrained more by a smaller true variance. Now consider the columns labeled "Est[imated] DF," which for each row in the table is the posterior median DF for each artificial data set, averaged over the 100 artificial data sets. When heterogeneity is truly absent [rows (1, 1, 0) and (0.25, 0.25, 0)], the analysis nonetheless allocates some DF to heterogeneity, (35 and 32 when the variances are 1 and 0.25, respectively), and the estimated DF for the competing clustering effect are below the true values by 23 and 10 DF respectively. That is, some of heterogeneity's DF are "stolen" from clustering, about 2/3 when the variances are 1 (and competition is more fierce) and about 1/3 when the variances are 0.25. An analogous but less pronounced effect occurs when clustering is truly absent. We discuss the implications of this finding in Section 5.

3.2 Global Mean Surface Temperature: Linear Growth Model

Using the linear growth model to smooth the global mean surface temperature series, we consider two types of priors on $(\sigma^2_n, \sigma^2_1, \sigma^2_2)$: independent Gamma(0.001, 0.001) on each $1/\sigma^2_j$ and DF-induced priors. Figure 2 plots the data and three fits (posterior means) arising from gamma priors. For one fit, the analysis used the design matrix's original scaling; for the other two fits, the columns of the design matrix were multiplied by 50 and 100. All smooths capture the increase from 1881 to 1940, the decrease from 1940 to 1970, and the increase after 1970. The fit with the original scaling smooths the most, with $\text{DF}(\mathbf{w}_1) + \text{DF}(\mathbf{w}_2)$ having posterior mean 5.6. The gamma prior's effect is not invariant to the scaling of $\mathbf{X}$'s columns, so the posteriors differ noticeably when the same gamma prior is used with rescaled design matrixes. When $\mathbf{X}$'s columns are multiplied by 100, the posterior means of $\text{DF}(\mathbf{w}_1)$ and $\text{DF}(\mathbf{w}_2)$ sum to 21.5. It was simply luck that in the "natural" scaling, the gamma prior gave a reasonable result. In contrast, priors specified on DF avoid the problem of scaling. Instead, they control the fit's smoothness directly by constraining DF in the fit. Figure 4 plots fits from flat priors on $(\text{DF}(\mathbf{w}_1), \text{DF}(\mathbf{w}_2))$ with five different constraints on total DF in the smoothed effects. The fit becomes smoother as the constraint on total DF is reduced. Figure 5 shows histograms of 10,000 draws from the posterior of $\text{DF}(\mathbf{w}_1)$ and $\text{DF}(\mathbf{w}_2)$ arising from flat priors on them that are constrained so that the total smoothed DF is <6. Both posteriors are skewed, but in different directions. For the local mean, $E(\text{DF}(\mathbf{w}_1)|\mathbf{y}) = 0.86$, while for the local slope, $E(\text{DF}(\mathbf{w}_2)|\mathbf{y}) = 4.58$.

This example also presents an instance in which the approximate DF (Sec. 2.3; Ruppert, Wand, and Carroll 2003, sec. 8.3) performs poorly. For the flat priors on $(\text{DF}(\mathbf{w}_1), \text{DF}(\mathbf{w}_2))$ with total DF constrained to be <6, the posterior mean of exact and

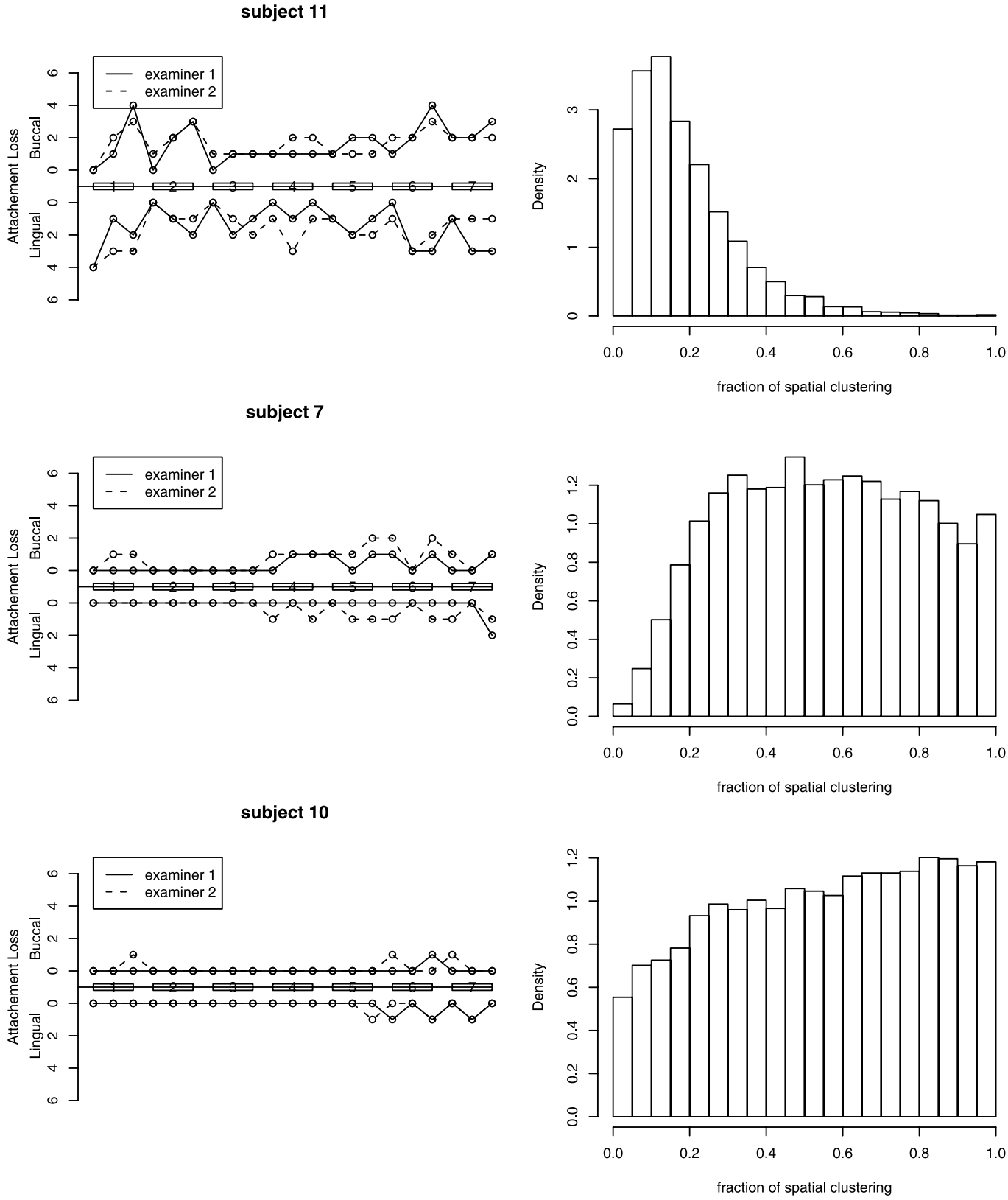

Figure 3. Example 1, AL data and histograms of draws from the posterior of f_i, the fraction of model complexity attributed to spatial clustering. The top two panels are for subject 11, the middle two panels are for subject 7, and the bottom two panels are for subject 10, who had $E(f_i|\mathbf{y}) = 0.18$, 0.52, and 0.52 respectively.

approximate DF for $\mathbf{w}_1$ are 0.86 and 1.87 respectively, differing by more than 1 DF, while the posterior mean of exact and approximate DF for $\mathbf{w}_2$ are closer, 4.58 and 4.76, respectively. When total DF is constrained to be <10, the posterior mean of exact and approximate DF are 2.97 and 5.13 for $\mathbf{w}_1$—differing by more than 2 DF—and 5.14 and 5.75 for $\mathbf{w}_2$.

Table 3. True DF and average posterior median DF for clustering and heterogeneity, for various true $(\sigma_\alpha^2/\sigma_0^2, \sigma_c^2/\sigma_0^2, \sigma_h^2/\sigma_0^2)$, with 100 simulated data sets per scenario

	True DF		Est DF		MC SE	
$(\frac{\sigma_\alpha^2}{\sigma_0^2}, \frac{\sigma_c^2}{\sigma_0^2}, \frac{\sigma_h^2}{\sigma_0^2})$	δ	ξ	δ	ξ	δ	ξ
(1, 1, 0)	245	0	222	35	20	19
(1, 1, 1)	143	233	143	233	24	29
(1, 0, 1)	0	328	7	323	3	15
(0.25, 0.25, 0)	121	0	111	32	15	17
(0.25, 0.25, 0.25)	95	133	97	132	18	33
(0.25, 0, 0.25)	0	164	7	160	3	28

NOTE: Est DF is the average of the 100 posterior medians, and MC SE is the Monte Carlo standard error of that estimate.

Other choices are possible for DF-based priors. For example, West and Harrison (1997) usually fix the signal-to-noise ratio for smoothing variances, which amounts to point priors on the corresponding DF. This could be relaxed by specifying priors on DF with the same centers as these point priors, but with positive variances.

The Gamma(0.001, 0.001) prior has been criticized on various grounds. Alternative priors on variances or standard deviations have been proposed; Gelman (2006) has provided an influential critique and some alternatives. These alternatives are all scale-dependent, like the gamma prior. We do not propose a flat prior on DF as a default, however; for this model and data set, a flat prior on $(\mathrm{DF}(\mathbf{w}_1), \mathrm{DF}(\mathbf{w}_2))$ without a constraint on total DF gives a grossly undersmoothed fit. It is still true that overparameterized models often need strong prior information.

4. DEGREES OF FREEDOM FOR RESIDUALS

The standard model (2), with $\boldsymbol{\Gamma}_0$ set to $\sigma_\epsilon^2\mathbf{I}$, can be reexpressed as

$$\begin{aligned}\mathbf{y} &= \mathbf{f} + \boldsymbol{\epsilon},\\ \boldsymbol{\epsilon} &\sim \mathrm{N}(\mathbf{0}, \sigma_\epsilon^2\mathbf{I}),\\ \mathbf{f} &\sim \mathrm{N}(\mathbf{X}_1\boldsymbol{\theta}_1, \mathbf{X}_2\boldsymbol{\Gamma}_2\mathbf{X}_2'),\end{aligned}$$

where $\mathbf{f}$ is a true but unknown function that we want to estimate, evaluated at the design values corresponding to the rows of $\mathbf{X}_1$. This notation is most familiar for penalized splines represented as mixed linear models (as in Ruppert, Wand, and Carroll 2003), where $\mathbf{f}$ is the true smooth function relating a predictor x and dependent variable $\mathbf{y}$. It appears that the approach of Ruppert, Wand, and Carroll (2003) requires that $\mathbf{X}_1$ have full rank; in this section we assume that $\mathbf{X}_1$ has full rank.

Treating σ_ϵ^2 and $\boldsymbol{\Gamma}_2$ as known, the fitted values from this model—the BLUPs or conditional posterior means—are $\hat{\mathbf{y}} = \mathbf{S}\mathbf{y}$ for

$$\mathbf{S} = [\,\mathbf{X}_1 \quad \mathbf{X}_2\,]\left[\begin{bmatrix}\mathbf{X}_1'\\ \mathbf{X}_2'\end{bmatrix}[\,\mathbf{X}_1 \quad \mathbf{X}_2\,] + \sigma_\epsilon^2\begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\Gamma}_2{}^{-1}\end{bmatrix}\right]^{-1}\begin{bmatrix}\mathbf{X}_1'\\ \mathbf{X}_2'\end{bmatrix}.$$

Here $\mathbf{S}$ is a linear smoother. The DF in the fitted values $\hat{\mathbf{y}}$ is $\mathrm{tr}(\mathbf{S})$ under all published definitions. The most common definition for residual DF in linear mixed models is derived as follows (Hastie and Tibshirani 1990, sec. 3.5; Ruppert, Wand, and Carroll 2003,

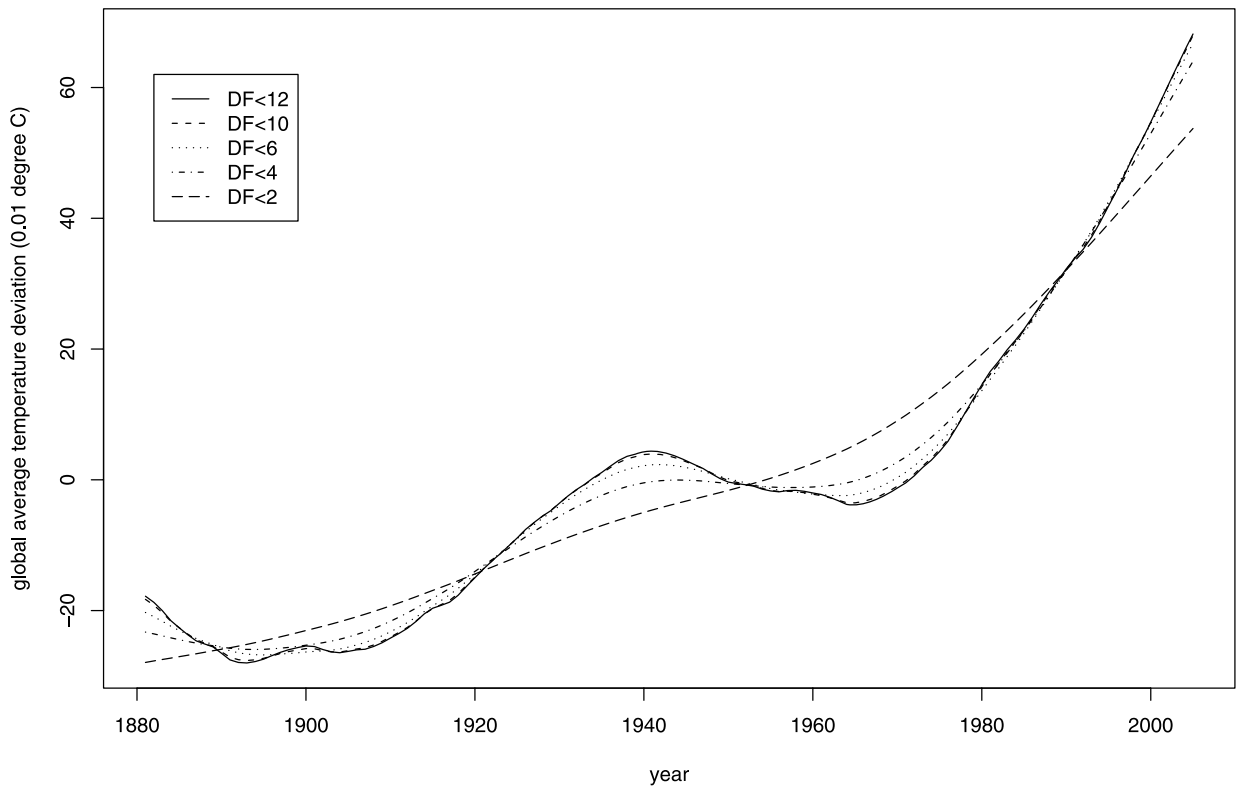

Figure 4. Example 2, global mean surface data, DF prior with different sum constraints on total smoothed DF.

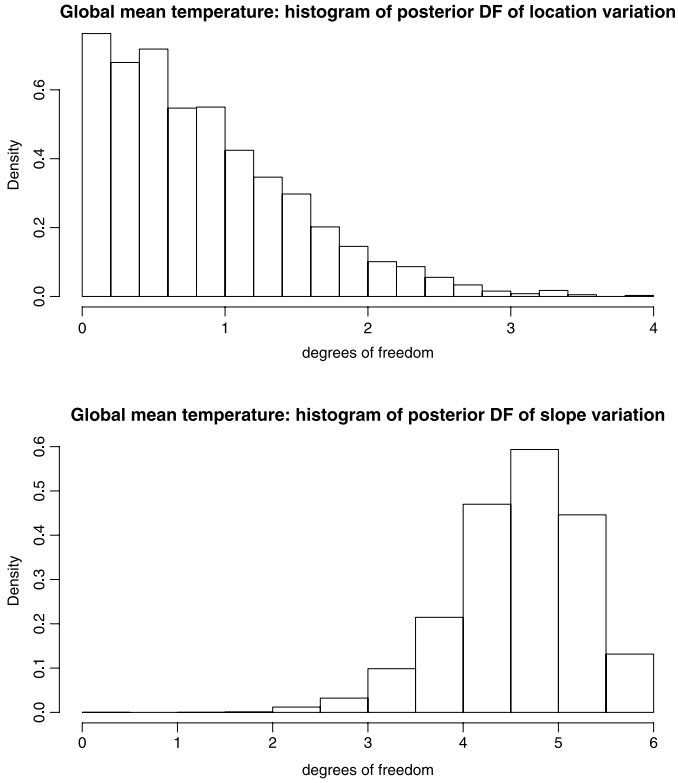

Figure 5. Example 2, global mean surface data, histograms of posterior DF for local mean (top) and local trend (bottom), for the flat prior on DF with sum of total smoothed DF constrained to <6.

sec. 3.14). Assuming that $\mathbf{f}$ is fixed, the mean squared error—the expectation of the residual sum of squares given $\boldsymbol{\Gamma}_2$, $\boldsymbol{\Gamma}_0$—is

$$\begin{aligned}
&\mathrm{MSE}(\mathbf{y}|\boldsymbol{\Gamma}_2, \boldsymbol{\Gamma}_0, \mathbf{f}) \\
&\quad = E_\epsilon[(\hat{\mathbf{y}} - \mathbf{y})'(\hat{\mathbf{y}} - \mathbf{y})|\boldsymbol{\Gamma}_2, \boldsymbol{\Gamma}_0, \mathbf{f}] \\
&\quad = E_\epsilon[\mathbf{y}'(\mathbf{S} - \mathbf{I})'(\mathbf{S} - \mathbf{I})\mathbf{y}|\boldsymbol{\Gamma}_2, \boldsymbol{\Gamma}_0, \mathbf{f}] \\
&\quad = \sigma_\epsilon^2 \operatorname{tr}((\mathbf{S} - \mathbf{I})'(\mathbf{S} - \mathbf{I})) + \mathbf{f}'(\mathbf{S} - \mathbf{I})'(\mathbf{S} - \mathbf{I})\mathbf{f} \\
&\quad = \sigma_\epsilon^2(n + \operatorname{tr}(\mathbf{S}'\mathbf{S}) - 2\operatorname{tr}(\mathbf{S})) + \mathbf{f}'(\mathbf{S} - \mathbf{I})'(\mathbf{S} - \mathbf{I})\mathbf{f}. \quad (6)
\end{aligned}$$

The term $\mathbf{f}'(\mathbf{S} - \mathbf{I})'(\mathbf{S} - \mathbf{I})\mathbf{f}$ arises from bias. "Assuming the bias term ... is negligible" (Ruppert, Wand, and Carroll 2003, p. 83), by analogy with ordinary linear models, the DF in the residuals is $n + \operatorname{tr}(\mathbf{S}'\mathbf{S}) - 2\operatorname{tr}(\mathbf{S})$, and {residual sum of squares}$/(n + \operatorname{tr}(\mathbf{S}'\mathbf{S}) - 2\operatorname{tr}(\mathbf{S}))$ is an unbiased estimate of σ_ϵ^2. Because DF in the fit is $\operatorname{tr}(\mathbf{S})$,

$$\begin{aligned}
&(\text{DF in fit}) + (\text{DF in residuals}) \\
&\quad = \operatorname{tr}(\mathbf{S}) + n + \operatorname{tr}(\mathbf{S}'\mathbf{S}) - 2\operatorname{tr}(\mathbf{S}) \\
&\quad = n + \operatorname{tr}(\mathbf{S}'\mathbf{S}) - \operatorname{tr}(\mathbf{S}) < n.
\end{aligned}$$

The inequality holds because $\mathbf{S}$ is symmetric with eigenvalues in $[0, 1]$ and at least one eigenvalue < 1, so $\operatorname{tr}(\mathbf{S}'\mathbf{S}) < \operatorname{tr}(\mathbf{S})$. This raises the question: where are the missing degrees of freedom?

Taking the mixed-effects model (2) literally, $\mathbf{f}$ is random, so we can remove the conditioning on $\mathbf{f}$ in (6). To do so, we need the claim, proved in Appendix (j), that if $\mathbf{f}$ is distributed as normal with mean $\mathbf{X}_1\boldsymbol{\theta}_1$ and covariance $\mathbf{X}_2\boldsymbol{\Gamma}_2\mathbf{X}_2'$, then

$$\begin{aligned}
&E_\mathbf{f}[\mathbf{f}'(\mathbf{S} - \mathbf{I})'(\mathbf{S} - \mathbf{I})\mathbf{f}|\boldsymbol{\Gamma}_2, \boldsymbol{\Gamma}_0] \\
&\quad = \operatorname{tr}\big((\mathbf{S} - \mathbf{I})'(\mathbf{X}_2\boldsymbol{\Gamma}_2\mathbf{X}_2' + \mathbf{X}_1\boldsymbol{\theta}_1(\mathbf{X}_1\boldsymbol{\theta}_1)')(\mathbf{S} - \mathbf{I})\big) \\
&\quad = \operatorname{tr}((\mathbf{S} - \mathbf{I})'(-\sigma_\epsilon^2\mathbf{S})).
\end{aligned}$$

We can now remove the conditioning on $\mathbf{f}$. By the foregoing claim,

$$\begin{aligned}
&\mathrm{MSE}(\mathbf{y}|\boldsymbol{\Gamma}_2, \boldsymbol{\Gamma}_0) \\
&\quad = E_\mathbf{f}[\mathrm{MSE}(\mathbf{y}|\boldsymbol{\Gamma}_2, \boldsymbol{\Gamma}_0, \mathbf{f})] \\
&\quad = \sigma_\epsilon^2(n + \operatorname{tr}(\mathbf{S}'\mathbf{S}) - 2\operatorname{tr}(\mathbf{S})) + E_\mathbf{f}[\mathbf{f}'(\mathbf{S} - \mathbf{I})'(\mathbf{S} - \mathbf{I})\mathbf{f}|\boldsymbol{\Gamma}_2, \boldsymbol{\Gamma}_0] \\
&\quad = \sigma_\epsilon^2(n + \operatorname{tr}(\mathbf{S}'\mathbf{S}) - 2\operatorname{tr}(\mathbf{S})) + \operatorname{tr}((\mathbf{S} - \mathbf{I})'(-\sigma_\epsilon^2\mathbf{S})) \\
&\quad = \sigma_\epsilon^2(n - \operatorname{tr}(\mathbf{S})).
\end{aligned}$$

By the same rationale used to define the usual residual DF, we can now define residual DF as $n - \operatorname{tr}(\mathbf{S})$, the same as the new definition of DF in Section 2.3. Therefore,

$$(\text{DF in fit}) + (\text{DF in residuals}) = \operatorname{tr}(\mathbf{S}) + n - \operatorname{tr}(\mathbf{S}) = n,$$

as defined in Section 2; the missing DF are in the bias term of mean squared error. The difference between the two definitions can be small or large. For the global mean surface temperature dataset, using the prior with total DF constrained to be <6 and fixing the variance ratios at their posterior medians, the residual DF is 115.7 under the widely used definition and 117.6 under the new definition. But for the AL data, with the variance ratios fixed at their posterior medians, the residual DF is 699.6 under the widely used definition and 764.4 under the new definition.

The foregoing highlights an awkward element of contemporary non-Bayesian theory for mixed-effects models, specifically those with random effects that do not meet the definition of, say, Scheffé (1959, p. 238), that a random effect's levels can "be regarded as a random sample from some population about which we wish to make our statistical inferences, rather than making them about the particular [levels] in the experiment." Penalized splines (as in Ruppert, Wand, and Carroll 2003, sec. 4.9) present the clearest example of this difficulty. The theory of penalized splines presumes that the smooth function $\mathbf{f}$ is fixed but unknown (e.g., Ruppert, Wand, and Carroll 2003, p. 58); the spline is fit using a random-effects model because this gives a minimization problem formally identical to the minimization problem in the original definition of penalized splines, and also allows us to use all the usual mixed linear model tools. For penalized splines, it seems meaningless outside a Bayesian framework to take an expectation over the distribution of $\mathbf{f}$, which is presumed to be fixed if unknown and is formally represented as a random effect for convenience only. Thus the mean squared error given $\boldsymbol{\Gamma}_0$, $\boldsymbol{\Gamma}_2$, and $\mathbf{f}$ is (6), a quantity that is fixed but unknown because $\mathbf{f}$ is unknown. Although in expectation over $\mathbf{f}$, total DF in the model and residuals add to the sample size n, for a given $\mathbf{f}$ this equality fails in general, and so "the books do not balance" as they do in conventional linear models.

The standard derivation of residual DF is intended to produce denominator DF for F-tests and is rationalized by the assumption that the squared length of the bias is small. For uses of this machinery described by Ruppert, Wand, and Carroll (2003) and

Hastie and Tibshirani (1990), in which the functions **f** are quite smooth, bias is likely small, and this presumption seems reasonable as it indeed was for the global mean surface temperature example. However, methods tend to outgrow their inventors' intentions, inevitably so for a scheme as inclusive as the one in Ruppert, Wand, and Carroll (2003), so this standard derivation will be extended to cases where bias is not necessarily small. Thus a theoretically more tidy treatment would seem desirable; our new definition of DF may provide such a treatment.

From a Bayesian standpoint, $\boldsymbol{\Gamma}_2$ can describe either variation in $\boldsymbol{\theta}_2$ for old-fashioned random effects or uncertainty about $\boldsymbol{\theta}_2$ for new-fangled random effects, and no problem arises because a probability distribution, once specified, obeys the same rules regardless of its rationale. But except when designing experiments, a Bayesian viewpoint does not consider expectations over varying **y**, so the usual derivation of residual DF is of no interest. Thus a Bayesian interpretation cannot justify removing the conditioning on **f** in the usual derivation of residual DF.

The new definition given in Section 2, while not explicitly Bayesian, does treat the covariance matrixes $\boldsymbol{\Gamma}_{2j}$ as Bayesians do, irrespective of their rationale. In the new definition, an effect's DF describes that effect's proportion of total variance, partitioning the sample size n naturally (property DF.b) while treating all sources of variation in the same manner, that is, without singling out error variation for distinctive treatment as is done in the usual derivation of residual DF.

5. DISCUSSION

We have presented a new conception of DF for models using random effects and normal probability distributions, which was anticipated in part by Green (2002). Although the resulting definition of DF arises from linear model theory, it defines a scale for complexity in the space of covariance structures and can be loosely interpreted as the fraction of variation attributed to individual effects. The new definition rationalizes partitioning the total DF in the data set into pieces for effects in the model and for error. The latter avoids difficulties in the most common definition of residual DF (Section 4). Conceiving an effect's DF as the fraction of variance attributed to the effect suggests a way of defining DF for nonnormal random-effects models, possibly avoiding the linear and normal approximations explicit in the approach of Lu, Hodges, and Carlin (2007) and implicit in the approach of Ruppert, Wand, and Carroll (2003, p. 212).

Our examples illustrate two uses of the new definition. The first use is to induce prior distributions on smoothing variances. In some cases, such variances are meaningful quantities, but priors on them are often nonintuitive and depend on the scale of the design matrix, as illustrated in Section 3.2. The examples show how to put scale-invariant priors on interpretable quantities like DF or $\mathrm{DF}(\boldsymbol{\delta})/(\mathrm{DF}(\boldsymbol{\delta}) + \mathrm{DF}(\boldsymbol{\xi}))$, inducing priors on the less intuitive smoothing parameters. The global mean surface temperature example shows how a prior on DF can control smoothness directly.

The new definition's second use is as an interpretable measure of each effect's contribution to the fit's complexity. This sheds light on competition between effects, which is especially acute in cases where the competing effects have design matrixes with column spaces that overlap substantially, as in both examples. Competition of this sort is a broad topic that is only now being noticed (e.g., Reich, Hodges, and Zadnik 2006); tools like DF can help us describe and understand this phenomenon.

As a byproduct, the AL example presents an instance in which the approximate DF of an effect (Ruppert, Wand, and Carroll 2003, sec. 8.3; Cui 2008, p. 79) performs poorly. It almost certainly is not an accident that this arose in a model with highly collinear effects, in contrast with the examples of Ruppert, Wand, and Carroll (2003). If the approximation fails only with severe collinearity, then it may be possible to develop a diagnostic that indicates when the approximation is accurate, which would be useful given that it computes much faster than exact DF.

The simulation described in Section 3.1 demonstrates a tendency to allocate DF to an effect that was, in fact, absent. In practice, this may not be important, because our methods would be used in the context of a larger effort of model building, and before fitting this model, one would probably drop the heterogeneity term if the data indicated its absence. Moreover, if the purpose of the analysis is better estimation of AL at individual sites, then this is almost certainly not important. Nonetheless, from a theoretical standpoint, this problem needs to be explained. The proximate cause is that smoothers tend not to shrink effects to zero even when they should. (This is the sole virtue of the zero variance estimates that are frequently produced by maximizing the restricted likelihood.) The extent of undershrinkage depends on the prior and would be reduced by, say, a scaled $\mathrm{Beta}(0.1, 0.1)$ prior on the relevant DF, which pushes the posterior more toward the extremes of complete or no shrinkage. Perhaps the larger implication is the need to rethink these extravagantly overparameterized models, which are the problem's ultimate cause. The difficulty almost certainly arises because of the extreme collinearity of the clustering and heterogeneity effects; partitioning degrees of freedom merely describes the consequence of this collinearity.

Finally, how does DF compare with the popular effective number of parameters p_D? Recall that in view of Plummer (2008), we distinguish the problems of model complexity and model comparison. Thus DF's purpose is to describe a model's complexity in terms drawn from simple, well-understood models, making it possible to control complexity in fitting the model to a data set by means of a prior distribution on complexity. Regarding p_D's interpretability, its definition (Spiegelhalter et al. 2002, sec. 2.3) is opaque, and the only concrete interpretation that Spiegelhalter et al. (2002) offer is that in simple cases—where complexity measures already exist—p_D takes values that people generally like. Second, our approach implicitly presumes that a model exists independently of any realized outcome **y**, so its complexity can be defined without a realized **y**. This is true of DF and of most complexity measures (e.g., Hastie and Tibshirani 1990; Ye 1998), but not p_D, the definition of which *requires* a realized **y** through the point estimate $\tilde{\boldsymbol{\theta}}$ and the posterior mean of the deviance $D(\boldsymbol{\theta})$. In special cases, such as normal hierarchical models with all covariances known, p_D does not depend on **y**, and in these cases p_D agrees with DF (Green 2002; Spiegelhalter et al. 2002, secs. 2 and 4). But with unknown variances (i.e., in practical situations), p_D and DF diverge.

A complexity measure defined independently of the outcome **y** will in general be a function of unknowns. This is acceptable; statisticians routinely specify models as functions of unknowns (in Bayesian terms, conditional on unknowns), and describing a model's complexity as a function of its unknowns creates no difficulties. The DF of a *fitted* model is obviously a function of the data, through either plug-in estimates or posterior distributions of the unknown parameters in the covariance matrixes. But a complexity measure defined independently of **y**, like DF, allows a prior distribution on complexity to softly control the complexity of the fit. This is not possible with p_D, because in general it is defined in terms of a specific realized **y**.

It is not clear that p_D can be partitioned corresponding to components of the model, as DF can. Spiegelhalter et al. (2002, secs. 6.3 and 8.1) did partition p_D corresponding to subdivisions of the deviance to create an outlier diagnostic from DIC, and in problems where the likelihood factors, DIC and p_D partition analogously (e.g., for errors-in-covariates models). Green (2002) gives a partition of p_D identical to ours in the special case where all covariance matrices are known, but it is not clear this can be extended to the case where covariances are unknown.

Thus, while the relatively easy computation of p_D is certainly desirable, for our purposes this seems to be outweighed by its conceptual and practical disadvantages.

SUPPLEMENTAL MATERIALS

Appendixes and Periodontal Data: This archive file contains Appendixes (a)–(j) and the periodontal data analyzed in Section 3.1, along with R code to read it and construct the neighbor-pair matrix **Q** used for the analyses. (TECHMS08-161 Supplementary Materials.zip; zip archive)

ACKNOWLEDGMENTS

This work was supported in part by a Merck Company Foundation Quantitative Sciences Fellowship. The work of Carlin was supported in part by National Cancer Institute grant 2-R01-CA095955-05A2. The authors thank Dr. Bryan Michalowicz of the University of Minnesota School of Dentistry for permission to use the periodontal data and post it on the *Technometrics* website.

[Received September 2008. Revised March 2009.]

REFERENCES

Besag, J., York, J. C., and Mollie, A. (1991), "Bayesian Image Restoration, With Two Applications in Spatial Statistics" (with discussion), *Annals of the Institute of Statistical Mathematics*, 43, 1–59. [124]

Best, N. G., Arnold, R. A., Thomas, A., Waller, L. A., and Conlon, E. M. (1999), "Bayesian Models for Spatially Correlated Disease and Exposure Data" (with discussion), in *Bayesian Statistics 6*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, Oxford: Oxford University Press, pp. 131–156. [127]

Cui, Y. (2008), "Smoothing Analysis of Variance for General Designs and Partitioning Degrees of Freedom in Hierarchical and Other Richly Parameterized Models," Ph.D. dissertation, University of Minnesota, Division of Biostatistics. [128,135]

Daniels, M. J. (1999), "A Prior for the Variance in Hierarchical Models," *The Canadian Journal of Statistics*, 27, 569–580. [126]

Eberly, L. E., and Carlin, B. P. (2000), "Identifiability and Convergence Issues for Markov Chain Monte Carlo Fitting of Spatial Models," *Statistics in Medicine*, 19, 2279–2294. [130]

Gelman, A. (2006), "Prior Distributions for Variance Parameters in Hierarchical Models," *Bayesian Analysis*, 1, 515–534. [133]

Gelman, A., and Pardoe, I. (2006), "Bayesian Measures of Explained Variance and Pooling in Multilevel (Hierarchical) Models," *Technometrics*, 48, 241–251. [128]

Gilthorpe, M. S., Zamzuri, A. T., Griffiths, G. S., Maddick, I. H., Eaton, K. A., and Johnson, N. W. (2003), "Unification of the 'Burst' and 'Linear' Theories of Periodontal Disease Progression: A Multilevel Manifestation of the Same Phenomenon," *Journal of Dental Research*, 82, 200–205. [124]

Green, P. (2002), Discussion of "Bayesian Measures of Model Complexity and Fit," by D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. van der Linde, *Journal of the Royal Statistical Society, Ser. B*, 64, 627–628. [128,135,136]

Groetsch, C. W. (1993), *Inverse Problems in the Mathematical Sciences*, Wiesbaden: Vieweg.

Hastie, T. J., and Tibshirani, R. J. (1990), *Generalized Additive Models*, Boca Raton, FL: CRC/Chapman & Hall. [128-130,133,135]

Hodges, J. S., and Sargent, D. J. (2001), "Counting Degrees of Freedom in Hierarchical and Other Richly-Parameterised Models," *Biometrika*, 88, 367–379. [124,125]

Hodges, J. S., Carlin, B. P., and Fan, Q. (2003), "On the Precision of the Conditionally Autoregressive Prior in Spatial Models," *Biometrics*, 59, 317–322. [125]

Hodges, J. S., Cui, Y., Sargent, D. J., and Carlin, B. P. (2007), "Smoothing Balanced Single-Error-Term Analysis of Variance," *Technometrics*, 49, 12–25. [126]

Lu, H., Hodges, J. S., and Carlin, B. P. (2007), "Measuring the Complexity of Generalized Linear Hierarchical Models," *The Canadian Journal of Statistics*, 35, 69–87. [125,126,135]

Natarajan, R., and Kass, R. E. (2000), "Reference Bayesian Methods for Generalized Linear Mixed Models," *Journal of the American Statistical Association*, 95, 227–237. [126]

Ortega, J. M., and Rheinboldt, W. C. (1970), *Iterative Solution of Nonlinear Equations in Several Variables*, New York: Academic Press.

Paciorek, C. J., and Schervish, M. J. (2006), "Spatial Modelling Using a New Class of Nonstationary Covariance Functions," *Environmetrics*, 17, 483–506. [126]

Plummer, M. (2008), "Penalized Loss Functions for Bayesian Model Comparison," *Biostatistics*, 9, 523–539. [127,135]

Prado, R., West, M., and Krystal, A. D. (2001), "Multichannel Electroencephalographic Analyses via Dynamic Regression Models With Time-Varying Lag-Lead Structure," *Journal of the Royal Statistical Society, Ser. C*, 50, 95–109. [127]

Reich, B. J., and Hodges, J. S. (2008a), "Modeling Longitudinal Spatial Periodontal Data: A Spatially Adaptive Model With Tools for Specifying Priors and Checking Fit," *Biometrics*, 64, 790–799. [124]

——— (2008b), "Identification of the Variance Components in the General Two-Variance Linear Model," *Journal of Statistical Planning and Inference*, 38, 1592–1604. [125]

Reich, B. J., Hodges, J. S., and Carlin, B. P. (2007), "Spatial Analyses of Periodontal Data Using Conditionally Autoregressive Priors Having Two Types of Neighbor Relations," *Journal of the American Statistical Association*, 102, 44–55. [124]

Reich, B. J., Hodges, J. S., and Zadnik, V. (2006), "Effects of Residual Smoothing on the Posterior of the Fixed Effects in Disease-Mapping Models," *Biometrics*, 62, 1197–1206. [127,135]

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003), *Semiparametric Regression*, New York: Cambridge University Press. [127-129,131,133-135]

Scheffé, H. (1959), *The Analysis of Variance*, New York: Wiley. [134]

Schott, J. R. (1997), *Matrix Analysis for Statistics*, New York: Wiley. [128]

Spiegelhalter, D. J., Best, D. G., Carlin, B. P., and van der Linde, A. (2002), "Bayesian Measures of Model Complexity and Fit" (with discussion), *Journal of the Royal Statistical Society, Ser. B*, 64, 583–639. [127,135,136]

Vaida, F., and Blanchard, S. (2005), "Conditional Akaike Information Criterion for Mixed Effects Models," *Biometrika*, 92, 351–370. [127]

Weisberg, S. (1985), *Applied Linear Regression*, New York: Wiley. [130]

West, M., and Harrison, J. (1997), *Bayesian Forecasting and Dynamic Models* (2nd ed.), New York: Springer. [126,127,133]

Ye, J. (1998), "On Measuring and Correcting the Effects of Data Mining and Model Selection," *Journal of the American Statistical Association*, 93, 120–131. [135]

Zhao, Y. H. (1999), "Examining Mean Structure, Correlation Structure, and Differences Between Examiners in a Large Periodontal Data Set," M.S. thesis, University of Minnesota, Division of Biostatistics. [130]