Published in final edited form as:

Technometrics. 2010 February 1; 52(1): 124–136. doi:10.1198/TECH.2009.08161.

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

Yue Cui¹, James S. Hodges^{2,*}, Xiaoxiao Kong², and Bradley P. Carlin²

¹ Department of Biometrics, Abbott, Abbott Park, Illinois USA 60064

² Division of Biostatistics, University of Minnesota, Minneapolis, Minnesota USA 55414

SUMMARY

Hodges & Sargent (2001) developed a measure of a hierarchical model's complexity, degrees of freedom (DF), that is consistent with definitions for scatterplot smoothers, interpretable in terms of simple models, and that enables control of a fit's complexity by means of a prior distribution on complexity. DF describes complexity of the whole fitted model but in general it is unclear how to allocate DF to individual effects. We give a new definition of DF for arbitrary normal-error linear hierarchical models, consistent with Hodges & Sargent's, 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 gives 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 to the effective number of parameters p_D of Spiegelhalter et al (2002). Technical appendices and a dataset are available online as supplemental materials.

Keywords

Degrees of freedom; hierarchical model; model complexity; prior distribution

1 Introduction and motivation

Hodges & 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. We use an example to introduce 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. Below, 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, three 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.

*hodges@ccbr.umn.edu.

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. Non-spatial heterogeneity arises from local features of the dentition, e.g., 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 non-spatial random effects to merge the “linear” and “burst” theories of progressive attachment loss, while Reich & Hodges (2008a) used spatial random effects to improve detection of disease progression. The preceding paragraph, however, suggests a model with both spatial clustering and non-spatial heterogeneity in true AL. We consider a model commonly used by epidemiologists (Besag et al 1991). For simplicity, this section models one subject and one measurement at each site; Section 3.1 analyzes all 12 subjects, and the dataset 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 $\boldsymbol{\xi} = (\xi_1, \dots, \xi_N)'$, modelled as $N(\mathbf{0}, \tau_h^2 \mathbf{I}_N)$. Spatial clustering is captured by $\boldsymbol{\delta}$ as follows. Neighbor relations among sites are summarized in an $N \times N$ matrix $\mathbf{Q}$, with non-diagonal 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 we used; other choices are possible (e.g., Reich et al 2007). Then $\boldsymbol{\delta}$ is modeled with an intrinsic conditional autoregressive model, $CAR(\mathbf{Q}, \tau_c^2)$: conditional on τ_c^2 , $\boldsymbol{\delta}$ has (improper) joint density

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

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 non-zero intercept, which is made explicit below. Also note that $\boldsymbol{\delta}' \mathbf{Q} \boldsymbol{\delta} = \sum_{j \sim i} (\delta_i - \delta_j)^2$, where $j \sim i$ means j is i 's neighbor, so $\mathbf{Q}$'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 respectively. 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 $\mathbf{y} = \boldsymbol{\delta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\delta} \sim CAR(\mathbf{Q}, \tau_c^2)$ as above and $\varepsilon_i \sim N(0, \sigma^2)$, and assume $G = 1$. 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 $\mathbf{D} = \text{diag}(d_1, \dots, d_{N-1}, 0)$, $d_1 \geq \dots \geq d_{N-1} > 0$, 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 et al 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 , while for $i < N$, $\hat{\phi}_i = (\mathbf{V}_1' \mathbf{y})_i / (1 + r d_i)$, which is shrunk toward zero by $r d_i > 0$. 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 zero to increasingly greater degrees for given $r = \sigma^2 / \tau_c^2$. Thus for small τ_c^2 , the fitted values $\hat{\delta}$ mostly reflect damped large-scale structure (Reich & Hodges 2008b) and as τ_c^2 increases, damping is reduced in all $\hat{\phi}$ and the fit becomes increasingly wiggly.

After fitting this model, any of several measures could be used to describe the whole model's complexity, but the distinct contributions of the heterogeneity and clustering components are not well understood. This paper proposes a partition of a readily interpretable measure of complexity (Hodges & 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 $\mathbf{X}_1\theta$ and covariance matrix Γ_0 , with $\mathbf{X}_1$ known and θ constrained by a model or prior distribution, $\mathbf{Z}\theta \sim N(\mathbf{0}, \Gamma_1)$, with $\mathbf{Z}$ known. Γ_0 and Γ_1 are usually functions of a few unknowns. H&S reformulated this as a linear model

$$\mathbf{Y} = \begin{pmatrix} \mathbf{y} \\ \mathbf{0} \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{Z} \end{pmatrix} \theta + \begin{pmatrix} \boldsymbol{\varepsilon} \\ \boldsymbol{\delta} \end{pmatrix} = \mathbf{X}\theta + \mathbf{e}, \quad (1)$$

where $\mathbf{e} = (\boldsymbol{\varepsilon}, \boldsymbol{\delta})' \sim N(\mathbf{0}, \Gamma)$ for $\Gamma = \text{diag}(\Gamma_0, \Gamma_1)$, so $\boldsymbol{\varepsilon}$ and $\boldsymbol{\delta}$ are independent *a priori* conditional on unknowns in Γ . (This formulation is equivalent to H&S'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 Γ $\mathbf{X}_1\theta$ has posterior mean or BLUP $\mathbf{X}_1(\mathbf{X}'\Gamma^{-1}\mathbf{X})^{-1}\mathbf{X}'\Gamma_0^{-1}\mathbf{y}$. H&S used linear model theory to rationalize defining the complexity or degrees of freedom (DF) of this model fit given Γ as

$$\rho = \text{tr}[\mathbf{X}_1(\mathbf{X}'\Gamma^{-1}\mathbf{X})^{-1}\mathbf{X}'\Gamma_0^{-1}],$$

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

H&S's DF can be used for several purposes. Model complexity is used by model-comparison criteria, though for reasons discussed below, we do not consider such criteria. The original purposes 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 Γ . (Section 2.6 below gives an accessible introduction to this idea.) Several authors have specified priors this way, e.g., Lu et al (2007), extending uniform shrinkage priors (Daniels 1999, Natarajan & Kass 2000), Paciorek & Schervish (2006), and Hodges et al (2007).

H&S's approach has limitations that restrict its use. Except for special cases, e.g., 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 re-expressed as $\mathbf{V}_1'\boldsymbol{\delta} \sim N(\mathbf{0}, \tau_c^2\mathbf{D}_1^{-1})$, where $\mathbf{D}_1 = \text{diag}(d_1, \dots, d_{N-1})$. The models for $\boldsymbol{\delta}$ and $\boldsymbol{\xi}$ are then combined as in (1):

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix} = \begin{pmatrix} \mathbf{I} & \mathbf{I} \\ -\mathbf{V}'_1 & \mathbf{0} \\ \mathbf{0} & -\mathbf{I} \end{pmatrix} \begin{pmatrix} \delta \\ \xi \end{pmatrix} + \begin{pmatrix} \varepsilon \\ \varphi_c \\ \varphi_h \end{pmatrix}$$

for $\varepsilon \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$, $\varphi_c \sim N(\mathbf{0}, \tau_c^2 \mathbf{D}_1^{-1})$, and $\varphi_h \sim N(\mathbf{0}, \tau_h^2 \mathbf{I}_N)$. H&S's DF is then (Lu et al 2007)

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

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

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

While this suggests how total DF in the fit might be attributed to δ and ξ , any such partition must be rigorously justified (Section 2.5 below does so).

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 , $t = 0, \dots, 124$, 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 & Harrison 1997, Chapter 2). This model has equations for observation error, variation in local mean, and variation in trend respectively:

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

Let μ , β , $\mathbf{n}$, $\mathbf{w}_1$ and $\mathbf{w}_2$ be the vectors of μ_t , β_t , n_t , $w_{1,t}$ and $w_{2,t}$ respectively. Assume $\mathbf{n}$, $\mathbf{w}_1$ and $\mathbf{w}_2$ are mutually independent and $\mathbf{n} \sim N(\mathbf{0}, \sigma_n^2 \mathbf{I}_{T+1})$, $\mathbf{w}_1 \sim N(\mathbf{0}, \sigma_1^2 \mathbf{I}_T)$, and $\mathbf{w}_2 \sim N(\mathbf{0}, \sigma_2^2 \mathbf{I}_{T-1})$, so σ_n^2 , σ_1^2 and σ_2^2 describe respectively 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, while the model for updating μ_t is similar, so the variances σ_1^2 and σ_2^2 play roles like those of the variances in Example 1.

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

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

, 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

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

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 Section 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 et al 2006) and with each other, 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 Example 1's model, Eberly and Carlin (2000) used $\tau_c/(\tau_c + \tau_h)$ as a measure of clustering's contribution to model fit. However, τ_c and τ_h are not commensurate: they are arbitrary, depending on the scaling of the design matrices for δ and ξ . Also, while the CAR's variance parameter τ_c^2 can be interpreted as an index of correlation between neighbor pairs, it does not describe that correlation in any precise sense (Wall 2004). DF, by contrast, is independent of scaling (see Section 2.4 below) 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 $DF_c/(DF_c + DF_h)$ shows the data's evidence about 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 & Harrison (1997, Chapter 2) and Prado et al (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 think about, it is hard to specify priors for them. A prior on interpretable DF avoids this difficulty.

This paper proposes a new definition of DF, consistent with 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 et al 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 it to the examples and

explores its uses. The new definition is consistent with the one in Ruppert et al (2003, Section 8.3) except for DF for residuals (Ruppert et al 2003, Section 3.14). Ruppert et al (2003) also gave an approximation to an effect's DF, saying "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 exact DF. Section 4 considers DF for residuals, in particular why the most common formula (Ruppert et al 2003, Section 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 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, e.g., Vaida & 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 view 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, so that describing model complexity and comparing models are distinct problems. We therefore 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

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\theta}_1 + \mathbf{X}_2\boldsymbol{\theta}_2 + \boldsymbol{\varepsilon}, \quad (2)$$

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, etc. 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 $\mathbf{X}_2 = [\mathbf{X}_{21}, \dots, \mathbf{X}_{2L}]$, $l \leq L$, and

conformably partition $\boldsymbol{\theta}_2 = (\boldsymbol{\theta}_{21}', \dots, \boldsymbol{\theta}_{2L}')'$. Model the j^{th} cluster of smoothed parameters $\boldsymbol{\theta}_{2j}$ as $\boldsymbol{\theta}_{2j} | \boldsymbol{\Gamma}_{2j} \sim N(\mathbf{0}, \boldsymbol{\Gamma}_{2j})$, 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 $\boldsymbol{\varepsilon}$ 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 $\mathbf{y} = \mathbf{X}_1\boldsymbol{\theta}_1 + \mathbf{S} + \boldsymbol{\varepsilon}$, where $\mathbf{X}_1\boldsymbol{\theta}_1$ represents regressors, $\mathbf{X}_2 = \mathbf{I}_n$, $\boldsymbol{\theta}_2 = \mathbf{S} \in R^n$ with $\boldsymbol{\Gamma}_2 = \text{cov}(\mathbf{S})$ non-diagonal and usually a function of unknowns, and $\text{cov}(\boldsymbol{\varepsilon})$ is diagonal.

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

2.2 Motivation for the new definition

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

$$\mathbf{y} = \mathbf{X}_2\boldsymbol{\theta} + \boldsymbol{\varepsilon}, \quad (3)$$

where θ is smoothed using the model or prior $\theta \sim N(\mathbf{0}, \Gamma_2)$, and $\varepsilon \sim N(\mathbf{0}, \Gamma_0)$, for positive definite Γ_0, Γ_2 . H&S wrote the model on θ as “constraint cases”, and combined it with (3) to give

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

where ε^* is multinormal with mean $\mathbf{0}$ and block diagonal covariance $\Gamma = \text{diag}(\Gamma_0, \Gamma_2)$. H&S's DF is

$$\begin{aligned} \rho &= \text{tr}[\mathbf{X}_2(\mathbf{X}'_2\Gamma^{-1}\mathbf{X}_2)^{-1}\mathbf{X}'_2\Gamma_0^{-1}] \\ &= \text{tr}[\mathbf{X}_2(\mathbf{X}'_2\Gamma_0^{-1}\mathbf{X}_2 + \Gamma_2^{-1})^{-1}\mathbf{X}'_2\Gamma_0^{-1}]. \end{aligned} \quad (4)$$

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

$$(\mathbf{X}'_2\Gamma_0^{-1}\mathbf{X}_2 + \Gamma_2^{-1})^{-1} = \Gamma_2 - \Gamma_2\mathbf{X}'_2(\mathbf{X}_2\Gamma_2\mathbf{X}'_2 + \Gamma_0)^{-1}\mathbf{X}_2\Gamma_2. \quad (5)$$

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

$$\begin{aligned} \rho &= \text{tr}[\mathbf{X}_2\Gamma_2\mathbf{X}'_2\Gamma_0^{-1} - \mathbf{X}_2\Gamma_2\mathbf{X}'_2(\mathbf{X}_2\Gamma_2\mathbf{X}'_2 + \Gamma_0)^{-1}\mathbf{X}_2\Gamma_2\mathbf{X}'_2\Gamma_0^{-1}] \\ &= \text{tr}[\mathbf{X}_2\Gamma_2\mathbf{X}'_2(\mathbf{X}_2\Gamma_2\mathbf{X}'_2 + \Gamma_0)^{-1}(\mathbf{X}_2\Gamma_2\mathbf{X}'_2 + \Gamma_0 - \mathbf{X}_2\Gamma_2\mathbf{X}'_2)\Gamma_0^{-1}] \\ &= \text{tr}[\mathbf{X}_2\Gamma_2\mathbf{X}'_2(\mathbf{X}_2\Gamma_2\mathbf{X}'_2 + \Gamma_0)^{-1}], \end{aligned}$$

which has the form: trace [ratio of {modelled 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 θ_{2j} are independent of each other conditional on their covariances θ_{2j} , and the variance in $\mathbf{y}$ arising from the effect

represented by θ_{2j} is $\mathbf{X}_{2j}\Gamma_{2j}\mathbf{X}'_{2j}$. Along with variation from the error ε , the total modelled variance for $\mathbf{y}$ is $\sum_1^l \mathbf{X}_{2j}\Gamma_{2j}\mathbf{X}'_{2j} + \Gamma_0 = \mathbf{X}_2\Gamma_2\mathbf{X}'_2 + \Gamma_0$. Thus, in this special case when all effects are smoothed (θ_1 is null) and all Γ_j are positive definite, the reformulated DF for $\mathbf{X}_{2j}$ is

$$\text{tr}[\mathbf{X}_{2j}\Gamma_{2j}\mathbf{X}'_{2j}(\mathbf{X}_2\Gamma_2\mathbf{X}'_2 + \Gamma_0)^{-1}]$$

Green (2002) used a different approach to derive an equivalent decomposition of the effective number of parameters p_D when Γ_0 and Γ_2 are completely specified, in which case p_D is identical to H&S's DF. However, Green's derivation does not appear to extend to cases in which Γ_0 or Γ_2 are functions of unknowns. The development above and below is also reminiscent of Gelman & 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 are 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, but 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 $\lambda \Gamma_1 \in R^{n \times n}$ as θ_1 's covariance matrix, where Γ_1 is unspecified but positive definite and λ is a positive scalar. In the limit as λ goes to $+\infty$, Γ_1 disappears.

For the standard model (2), then, define DF for nonsmoothed effects, denoted by $DF(\mathbf{X}_1)$, as

$$DF(\mathbf{X}_1) = \lim_{\lambda \rightarrow +\infty} \text{tr}[\mathbf{X}_1 \lambda \Gamma_1 \mathbf{X}_1' (\mathbf{X}_1 \lambda \Gamma_1 \mathbf{X}_1' + \mathbf{X}_2 \Gamma_2 \mathbf{X}_2' + \Gamma_0)^+] \\ = \text{tr}[P_{\mathbf{X}_1}] = \text{rank}(\mathbf{X}_1).$$

Appendix (b) gives a proof. For the smoothed effect θ_{2j} , DF is defined analogously:

$$DF(\mathbf{X}_{2j}) = \lim_{\lambda \rightarrow +\infty} \text{tr}[\mathbf{X}_{2j} \Gamma_{2j} \mathbf{X}_{2j}' (\mathbf{X}_1 \lambda \Gamma_1 \mathbf{X}_1' + \mathbf{X}_2 \Gamma_2 \mathbf{X}_2' + \Gamma_0)^+] \\ = \text{tr}\{\mathbf{X}_{2j} \Gamma_{2j} \mathbf{X}_{2j}' [(\mathbf{I} - P_{\mathbf{X}_1})(\mathbf{X}_2 \Gamma_2 \mathbf{X}_2' + \Gamma_0)(\mathbf{I} - P_{\mathbf{X}_1})]^+\}.$$

Appendix (c) gives a proof. Similarly, for the error term ϵ ,

$$DF(\epsilon) = \lim_{\lambda \rightarrow +\infty} \text{tr}[\mathbf{I} \Gamma_0 \mathbf{I}' (\mathbf{X}_1 \lambda \Gamma_1 \mathbf{X}_1' + \mathbf{X}_2 \Gamma_2 \mathbf{X}_2' + \Gamma_0)^+] \\ = \text{tr}\{\Gamma_0 [(\mathbf{I} - P_{\mathbf{X}_1})(\mathbf{X}_2 \Gamma_2 \mathbf{X}_2' + \Gamma_0)(\mathbf{I} - P_{\mathbf{X}_1})]^+\}.$$

This follows by the argument in Appendix (c). In general this differs from residual degrees of freedom as defined in Ruppert et al (2003, Section 3.14) or Hastie & Tibshirani (1990, Section 3.5).

By this definition, the DF of a smoothed effect $\mathbf{X}_{2j}$ or error ϵ is the fraction of variation contributed by $\mathbf{X}_{2j}$ or ϵ out of variation not accounted for by the unsmoothed effects $\mathbf{X}_1$. Because $DF(\mathbf{X}_{2j})$ can be laborious to compute, Cui (2008, p. 79) derived the approximation

$DF_p(\mathbf{X}_{2j}) = \text{tr}\{\mathbf{X}_{2j} \Gamma_{2j} \mathbf{X}_{2j}' [(\mathbf{I} - P_{\mathbf{X}_1})(\mathbf{X}_2 \Gamma_2 \mathbf{X}_2' + \Gamma_0)(\mathbf{I} - P_{\mathbf{X}_1})]^+\}$, which is equivalent to the approximation in Ruppert et al (2003, pp. 175–176) when both are defined. Section 3.2 briefly considers 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 Γ_0 is positive definite,

$DF(\mathbf{X}_1) + \sum_{j=1}^J DF(\mathbf{X}_{2j}) + DF(\epsilon) = \text{dim}(\mathbf{y})$, the dimension of $\mathbf{y}$. Appendix (e) gives a proof. For the definition of DF in residuals given by Ruppert et al (2003) and Hastie & Tibshirani

(1990), $DF(\mathbf{X}_1) + \sum_1^J DF(\mathbf{X}_{2j}) + DF(\text{residuals}) < \dim(\mathbf{y})$, with the deficiency arising from $DF(\text{residuals})$.

(DF.c)—The DF of an effect has a reasonable upper bound: $DF(\mathbf{X}_{2j}) \leq \text{rank}((\mathbf{I} - P_{\mathbf{X}_1})\mathbf{X}_{2j}) = \text{rank}([\mathbf{X}_1, \mathbf{X}_{2j}]) - \text{rank}(\mathbf{X}_1) \leq \text{rank}(\mathbf{X}_{2j})$ for $j = 1, \dots, J$. 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. (Section 2.6 discusses prior distributions on DF.) Suppose in the standard model (2), the covariance of each smoothed effect θ_{2j} and the covariance of the error term ε are each characterized by a single parameter:

$\Gamma_{2j} = \sigma_j^2 \Gamma_{2j}^0$ and $\Gamma_0 = \sigma_0^2 \Gamma_0^0$, where the σ_j^2 are unknown scalars and $\Gamma_{2j}^0, \Gamma_0^0$ are known and

positive definite. Let $\mathbf{B} \in R^{p \times p}$ be nonsingular, and $\mathbf{H}_j$ be a matrix such that $\mathbf{H}_j \Gamma_{2j}^0 \mathbf{H}_j' = \Gamma_{2j}^0$,

e.g., Γ_{2j}^0 is the identity and $\mathbf{H}_j$ is orthogonal. Then the posterior of $\mathbf{X}\theta$ arising from

independent priors on $(DF(\mathbf{X}_{21}), \dots, DF(\mathbf{X}_{2J}))$ 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 $\mathbf{X}_{2j}^* = t_j \mathbf{X}_{2j} \mathbf{H}_j$ 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 the spatial map represented by $\mathbf{Q}$ is connected, so $G = 1$. (This is not necessary.) The CAR model for the

spatial clustering effects δ models $\mathbf{V}_1' \delta$ as $N(\mathbf{0}, \tau_c^2 \mathbf{D}_1^{-1})$, while $\mathbf{1}_N' \delta$ is not smoothed, so re-express 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

$\theta_1 = \mathbf{1}_N' \delta / N$, $\theta_{21} = \mathbf{V}_1' \delta$, $\theta_{22} = \xi$, $\Gamma_{21} = \tau_c^2 \mathbf{D}_1^{-1}$, $\Gamma_{22} = \tau_h^2 \mathbf{I}_N$, and $\Gamma_0 = \sigma^2 \mathbf{I}_N$. The constraint cases become $\theta_{21} \sim N(\mathbf{0}, \tau_c^2 \mathbf{D}_1^{-1})$ and $\theta_{22} \sim N(\mathbf{0}, \tau_h^2 \mathbf{I}_N)$. Appendix (h) derives these expressions for $DF(\delta)$ and $DF(\xi)$:

$$\begin{aligned} DF(\delta) &= \text{rank}(\mathbf{X}_1) + \text{tr}\{\mathbf{X}_{21} \Gamma_{21} \mathbf{X}_{21}' [(\mathbf{I}_N - P_{\mathbf{X}_1})(\mathbf{X}_{21} \Gamma_{21} \mathbf{X}_{21}' + \mathbf{X}_{22} \Gamma_{22} \mathbf{X}_{22}' + \Gamma_0)(\mathbf{I}_N - P_{\mathbf{X}_1})]^{-1}\} \\ &= 1 + \sum_{i=1}^{N-1} \frac{\frac{\tau_c^2}{\sigma^2} / d_i}{\frac{\tau_c^2}{\sigma^2} / d_i + \frac{\tau_h^2}{\sigma^2} + 1}; \\ DF(\xi) &= \text{tr}\{\mathbf{X}_{22} \Gamma_{22} \mathbf{X}_{22}' [(\mathbf{I}_N - P_{\mathbf{X}_1})(\mathbf{X}_{21} \Gamma_{21} \mathbf{X}_{21}' + \mathbf{X}_{22} \Gamma_{22} \mathbf{X}_{22}' + \Gamma_0)(\mathbf{I}_N - P_{\mathbf{X}_1})]^{-1}\} \\ &= \sum_{i=1}^{N-1} \frac{\frac{\tau_h^2}{\sigma^2}}{\frac{\tau_c^2}{\sigma^2} / d_i + \frac{\tau_h^2}{\sigma^2} + 1}; \end{aligned}$$

Thus $DF(\delta) + DF(\xi)$ 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 $DF(\delta)$ increases and $DF(\xi)$ decreases. When there is more than one island ($G > 1$), it is easy to show that DF for clustering and heterogeneity are the sums of the respective DF for each island.

Example 2. Linear growth model—Write the 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:

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

In terms of the standard model, the unsmoothed effects are $\theta_1 = (\mu_0, \beta_0)'$, the smoothed effects $\theta_2 = (\mathbf{w}_1', \mathbf{w}_2')'$, the smoothing covariances $\Gamma_{21} = \sigma_1^2 \mathbf{I}_T$ and $\Gamma_{22} = \sigma_2^2 \mathbf{I}_{T-1}$, and the error covariance $\Gamma_0 = \sigma_n^2 \mathbf{I}_{T+1}$. $DF(\mathbf{w}_1)$, $DF(\mathbf{w}_2)$, and $DF(n)$ follow straightforwardly, but the expressions do not simplify, so 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 $\sigma_2^2 > 0$, $DF(\mathbf{w}_1)$ and $DF(\mathbf{w}_2)$ 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 $\sigma_2^2 > 0$, $\mathbf{w}_1$ and $\mathbf{w}_2$ compete with each other. Table 1 shows how they compete for various (σ_1^2, σ_2^2) ; noise variance σ_n^2 is fixed at 1 and T at 124. When $\sigma_2 = 0$ (Table 1a), $DF(\mathbf{w}_1)$ 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), $DF(\mathbf{w}_1)$ is reduced as σ_2 grows because larger σ_2 allows $\mathbf{w}_2$ to compete more aggressively, so $DF(\mathbf{w}_2)$ grows. The analogous result occurs when σ_1 increases for fixed σ_2 .

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 & 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 Example 1's model, the vector of effect-specific DF, $(DF(\delta), DF(\xi))$, 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 $(DF(\delta), DF(\xi))$. But it is equally legitimate to place a prior on $(DF(\delta), DF(\xi))$, to induce a prior on the unknown variance ratios $(\tau_c^2/\sigma^2, \tau_h^2/\sigma^2)$. Such a prior has the advantage of being specified on the interpretable complexity measure instead of the usually more obscure variances or variance ratios.

This can be stated more precisely. In the standard model (2), suppose each smoothed effect θ_{2j} and the error term ε have covariances characterized by one parameter, e.g., $\Gamma_{2j} = \sigma_j^2 \Gamma_{2j}^0$ and $\Gamma_0 = \sigma_0^2 \Gamma_0^0$, where the σ_j^2 are unknown scalars and $\Gamma_{2j}^0, \Gamma_0^0$ are known and positive definite. Assume further that $\mathbf{X}_{2j} \notin R(\mathbf{X}_1)$. Then $DF(\cdot)$ is a 1-1 mapping between $\mathbf{q} = (DF(\mathbf{X}_{21}), \dots, DF(\mathbf{X}_{2l}))'$ on $\mathbf{q}$'s range and $\mathbf{s} = (\log(\sigma_1^2/\sigma_0^2), \dots, \log(\sigma_l^2/\sigma_0^2)) \in R^l$, so 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 $\mathbf{X}_{2j}\theta_{2j}$ 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., $(DF(\delta)/DF(\delta) + DF(\xi))$ in Example 1 (discussed below). If an l -variate function of $\mathbf{q}$, $\mathbf{u} = (u_1(\mathbf{q}), \dots, u_l(\mathbf{q}))$ is a 1-1 mapping, then a prior on $\mathbf{u}$ induces a unique prior on $\mathbf{s}$.

When Γ_{2j} 's form is more complex than assumed above, a prior on DF partially specifies a prior on Γ_{2j} , and a complete prior specification requires a prior on other functions of Γ_{2j} . For example, if $\Gamma_{2j} = \text{diag}(\sigma_{j1}^2, \sigma_{j2}^2)$, then a uniform prior on $DF(\mathbf{X}_{2j})$ induces a prior on a scalar function of $(\sigma_{j1}^2/\sigma_0^2, \sigma_{j2}^2/\sigma_0^2)$. A complete specification requires a prior on another function of $(\sigma_{j1}^2/\sigma_0^2, \sigma_{j2}^2/\sigma_0^2)$, e.g., on $\sigma_{j1}^2/\sigma_{j2}^2$.

3 Examples: Priors on DF, partitioning DF, competing effects

3.1 Periodontal measurements: clustering and heterogeneity model

We analyze AL data as described in Section 1 for twelve subjects without missing measurements taken from a larger dataset of 410 subjects. The dataset is available on line with the supplemental materials. Each site in each subject was measured once by each of two examiners, so each subject provided 84 AL measurements on 42 sites. Zhao (1999) investigated differences between examiners using all 410 subjects. She 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 the present purpose we simply treat the two exams on each patient as providing two iid measurements at each site.

Let y_{ijk} be the k^{th} measurement of site j for subject i , $i = 1, \dots, 12$, $j = 1, \dots, 42$, $k = 1, 2$. Model y_{ijk} as $y_{ijk} = \mu + \alpha_i + \delta_{ij} + \xi_{ij} + \varepsilon_{ijk}$, where μ is the grand mean and α_i is subject i 's random effect, modelled as an independent $N(0, \sigma_a^2)$ draw. The vector $\delta_i = (\delta_{i1}, \dots, \delta_{i42})$

captures spatial clustering for subject i . We assume these are independent across i and model each δ_i as $CAR(\mathbf{Q}, \sigma_{c,i}^2)$, with neighbor pairs as in Figure 1. Heterogeneity effects ξ_{ij} are modelled as $N(0, \sigma_{h,i}^2)$ assumed iid within subject. The ε_{ijk} capture unsystematic measurement error and are modeled as independent $N(0, \sigma_0^2)$, σ_0^2 being common to all subjects.

Subject i has smoothing variances $\sigma_{c,i}^2$ and $\sigma_{h,i}^2$ and thus $DF(\delta_i)$ and $DF(\xi_i)$, 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 $f_i = DF(\delta_i)/(DF(\delta_i) + DF(\xi_i))$, the fraction, for subject i , of the fitted complexity attributed to clustering. A flat prior on f_i shows no preference between clustering and heterogeneity without constraining the total DF in subject i 's fit, $DF(\delta_i) + DF(\xi_i)$, so it could be considered natural. We used independent flat priors on f_i , $DF(\delta_i) + DF(\xi_i)$, and $DF(\alpha)$. Table 2 shows the posterior means of the f_i (the posterior medians are similar).

In Table 2, overall DF for spatial clustering has posterior mean 82.00 out of 492; subject-specific values range from 1.66 to 16.75 out of 41. Overall DF for heterogeneity has posterior mean 158.26 out of 503; subject-specific values range from 1.60 to 28.12 out of 42. The posterior mean of f_i ranges from 0.18 to 0.52 with median 0.36, compared to the prior mean 0.50, so for most subjects the data indicate that heterogeneity should receive more fitted complexity than clustering. Compared to $\sigma_{c,i}/(\sigma_{c,i} + \sigma_{h,i})$ (Eberly & Carlin 2000), f_i has a direct interpretation and accounts for all other sources of variation, i.e., because measurement error variance σ_0^2 is common to all subjects, f_i depends indirectly on $\sigma_{c,j}^2/\sigma_0^2$ and $\sigma_{h,j}^2/\sigma_0^2$ 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 $E(f_i|\mathbf{y})$ (subject 11) and the largest (subjects 7 and 10). Based on Section 1's interpretation, when DF are low for both clustering and heterogeneity, the data should have little large-scale trend or unpatterned noise respectively. When DF are large for both, we should see big differences in local level between different parts of the mouth, and 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 similar posterior DF for clustering, about 6 DF, while subject 11 has much larger DF for heterogeneity, about 28 DF compared to about 6 DF for subject 7. Figure 3's data show similarly modest large-scale trend for these two subjects but much greater unpatterned variation for subject 11. Now compare subjects 7 and 10, who have the same $E(f_i|\mathbf{y})$ while subject 10's posterior expected DF for both components are smaller by a factor of about 4. As expected, subject 10's data show much less trend and variation than subject 7's.

The random effects α , δ , and ξ compete with each other: the column spaces for α and δ are orthogonal to each other but both are contained in the column space for ξ . To give a taste of how this affects partitioning of DF, Table 3 shows results from simulated data like the periodontal dataset just analyzed. For simplicity, all 12 simulated subjects have the same σ_c^2 and σ_h^2 in both the data and the model. We fixed σ_0^2 at 1 and sampled 100 artificial datasets from each of 6 sets of true variance ratios ($\sigma_\alpha^2/\sigma_0^2, \sigma_c^2/\sigma_0^2, \sigma_h^2/\sigma_0^2$), drawing new α_i , δ_i , and ξ_i for each artificial dataset. The analyses used a flat prior on $(DF(\alpha), DF(\delta), DF(\xi))$; flat priors on $DF(\alpha)$, $DF(\delta) + DF(\xi)$ and $DF(\delta)/(DF(\delta) + DF(\xi))$ give similar results. The DF allocated to subjects, $DF(\alpha)$ are similar in all cases considered, and are not discussed further.

In Table 3, consider the columns labelled “True DF”; these are DF as a function of the true variances. The model with both clustering and heterogeneity (the 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 proportionally more when the true variances are 1 (42% reduction) than when the true variances are 0.25 (21%). The presence of clustering has a similar effect on the True DF for heterogeneity (29% and 19% reductions, 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 labelled “Est[imated] DF”, which for each row in the table is the posterior median DF for each artificial dataset, averaged over the 100 artificial datasets. 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. Section 5 discusses the implications of this finding.

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 $(\sigma_n^2, \sigma_1^2, \sigma_2^2)$: independent Gamma(0.001, 0.001) on each $1/\sigma_j^2$, 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, 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 $DF(\mathbf{w}_1) + DF(\mathbf{w}_2)$ 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 re-scaled design matrices. When $\mathbf{X}$ ’s columns have been multiplied by 100, the posterior means of $DF(\mathbf{w}_1)$ and $DF(\mathbf{w}_2)$ sum to 21.5. It was simply luck that in the “natural” scaling, the gamma prior gave a reasonable result.

Priors specified on DF, by contrast, 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 $(DF(\mathbf{w}_1), DF(\mathbf{w}_2))$ 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 $DF(\mathbf{w}_1)$ and $DF(\mathbf{w}_2)$ arising from flat priors on them constrained so total smoothed $DF < 6$. Both posteriors are skewed, but in different directions. For the local mean $E(DF(\mathbf{w}_1)|\mathbf{y}) = 0.86$, while for the local slope $E(DF(\mathbf{w}_2)|\mathbf{y}) = 4.58$.

This example also gives an instance in which the approximate DF (Section 2.3; Ruppert et al 2003, Section 8.3) performs poorly. For the flat priors on $(DF(\mathbf{w}_1), DF(\mathbf{w}_2))$ with total DF constrained to be less than 6, the posterior mean of exact and 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 less than 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 $\mathbf{w}_2$.

Other choices are possible for DF-based priors. For example, West & Harrison (1997) usually fix 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) gives an influential critique and some alternatives. These alternatives are all scale-dependent, like the gamma prior. We do not, however, propose a flat prior on DF as a default: for this model and dataset, a flat prior on $(DF(\mathbf{w}_1), DF(\mathbf{w}_2))$ 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 Γ_0 set to $\sigma_\varepsilon^2 \mathbf{I}$, can be re-expressed as

$$\begin{aligned} \mathbf{y} &= \mathbf{f} + \boldsymbol{\varepsilon} \\ \boldsymbol{\varepsilon} &\sim N(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I}) \\ \mathbf{f} &\sim N(\mathbf{X}_1 \boldsymbol{\theta}_1, \mathbf{X}_2 \Gamma_2 \mathbf{X}_2'), \end{aligned}$$

where $\mathbf{f}$ is a true but unknown function we wish 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 et al (2003), where $\mathbf{f}$ is the true smooth function relating a predictor x and dependent variable $\mathbf{y}$. It appears that Ruppert et al (2003) require $\mathbf{X}_1$ to have full rank; this section assumes $\mathbf{X}_1$ has full rank.

Treating σ_ε^2 and Γ_2 as known, the fitted values from this model – the BLUPs or conditional posterior means – are $\hat{\mathbf{y}} = \mathbf{S}\mathbf{y}$ for

$$\mathbf{S} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{X}_2 \end{bmatrix} \left[\begin{bmatrix} \mathbf{X}_1' \\ \mathbf{X}_2' \end{bmatrix} \begin{bmatrix} \mathbf{X}_1 & \mathbf{X}_2 \end{bmatrix} + \sigma_\varepsilon^2 \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \Gamma_2^{-1} \end{bmatrix} \right]^{-1} \begin{bmatrix} \mathbf{X}_1' \\ \mathbf{X}_2' \end{bmatrix};$$

$\mathbf{S}$ is a linear smoother. The DF in the fitted values $\hat{\mathbf{y}}$ is $tr(\mathbf{S})$ under all published definitions.

The most common definition for residual DF in linear mixed models is derived as follows (Ruppert et al 2003, Section 3.14; Hastie & Tibshirani 1990, Section 3.5). Assuming $\mathbf{f}$ is fixed, the mean squared error – the expectation of the residual sum of squares given Γ_2, Γ_0 – is

$$\begin{aligned} MSE(\mathbf{y}|\Gamma_2, \Gamma_0, \mathbf{f}) &= E_\varepsilon[(\hat{\mathbf{y}} - \mathbf{y})'(\hat{\mathbf{y}} - \mathbf{y})|\Gamma_2, \Gamma_0, \mathbf{f}] \\ &= E_\varepsilon[\mathbf{y}'(\mathbf{S} - \mathbf{I})'(\mathbf{S} - \mathbf{I})\mathbf{y}|\Gamma_2, \Gamma_0, \mathbf{f}] \\ &= \sigma_\varepsilon^2 tr((\mathbf{S} - \mathbf{I})'(\mathbf{S} - \mathbf{I})) + \mathbf{f}'(\mathbf{S} - \mathbf{I})'(\mathbf{S} - \mathbf{I})\mathbf{f} \\ &= \sigma_\varepsilon^2(n + tr(\mathbf{S}'\mathbf{S}) - 2tr(\mathbf{S})) + \mathbf{f}'(\mathbf{S} - \mathbf{I})'(\mathbf{S} - \mathbf{I})\mathbf{f}. \end{aligned} \tag{6}$$

The term $\mathbf{f}'(\mathbf{S} - \mathbf{I})'(\mathbf{S} - \mathbf{I})\mathbf{f}$ arises from bias. “Assuming the bias term ... is negligible” (Ruppert et al 2003, p. 83), then by analogy with ordinary linear models, the DF in the

residuals is $n + \text{tr}(\mathbf{S}'\mathbf{S}) - 2\text{tr}(\mathbf{S})$, and $\{\text{residual sum of squares}\}/(n + \text{tr}(\mathbf{S}'\mathbf{S}) - 2\text{tr}(\mathbf{S}))$ is an unbiased estimate of σ_ε^2 . Since DF in the fit is $\text{tr}(\mathbf{S})$,

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

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

Taking the mixed-effect model (2) literally, $\mathbf{f}$ is random, so we can remove the conditioning on $\mathbf{f}$ in (6). To do so, we need a 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

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

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

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

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

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

as defined in Section 2; the missing degrees of freedom 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 under 6 and fixing the variance ratios at their posterior medians, residual DF is 115.7 under the widely-used definition and 117.6 under the new definition. However, for the AL data, fixing the variance ratios at their posterior medians, 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 effect models, specifically those with random effects that do not meet the definition in, say, Scheffé (1959, p. 238: 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 et al (2003, Section 4.9), present the clearest example of this difficulty. The theory of penalized splines presumes the smooth function $\mathbf{f}$ is fixed but unknown (e.g., Ruppert et al 2003, p. 58); the spline is fit using a random-effect model because this gives a

minimization problem formally identical to the minimization problem in the original definition of penalized splines, but 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 fixed if unknown and is formally represented as a random effect for convenience only. Thus, the mean squared error given Γ_0 , Γ_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, so “the books don’t 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 bias is small. For uses of this machinery described by Ruppert et al (2003) and Hastie & Tibshirani (1990), in which the functions $\mathbf{f}$ are quite smooth, bias likely is small and this presumption seems reasonable as, indeed, it 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 et al (2003), so this standard derivation will be extended to cases where bias is not necessarily small, e.g., for the AL data. Thus, a theoretically more tidy treatment would seem desirable; the present paper’s new definition of DF may provide such a treatment.

From a Bayesian viewpoint, Γ_2 can describe either variation in θ_2 , for old-fashioned random effects, or uncertainty about θ_2 , for new-fangled random effects, and no problem arises because a probability distribution, once specified, obeys the same rules regardless of its rationale. However, except when designing experiments, a Bayesian viewpoint does not consider expectations over varying $\mathbf{y}$, so the usual derivation of residual DF has no interest. Thus, a Bayesian interpretation cannot justify removing the conditioning on $\mathbf{f}$ in the usual derivation of residual DF.

Section 2’s new definition, while not explicitly Bayesian, does treat the covariance matrices Γ_{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, i.e., without singling out error variation for distinctive treatment as in the usual derivation of residual DF.

5 Discussion

We have presented a new conception of degrees of freedom for models using random effects and normal probability distributions, anticipated partly 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 dataset 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 to define DF for non-normal random-effect models, possibly avoiding the linear and normal approximations explicit in Lu et al (2007) and implicit in Ruppert et al (2003, p. 212).

The examples illustrated two uses of the new definition. The first is inducing 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 Section 3.2 illustrated. The examples showed how to put scale-invariant priors on interpretable quantities like DF or $DF(\delta)/(DF(\delta) + DF(\zeta))$, inducing priors on the less

intuitive smoothing parameters. The global mean surface temperature example showed 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 matrices with column spaces that overlap substantially, as in both examples. Competition of this sort is a large topic that is only now being noticed (e.g., Reich et al 2006); tools like DF can help us describe and understand this phenomenon.

As a byproduct, the AL example gave an instance in which the approximate DF of an effect (Ruppert et al 2003, Section 8.3; Cui 2008, p. 79) performs poorly. It is almost certainly not an accident that this arose in a model with highly collinear effects, in contrast with the examples in Ruppert et al (2003). If the approximation fails only with severe collinearity, 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.

Section 3.1's simulation found 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 etc., and before fitting this model one would probably drop the heterogeneity term if the data indicated it was absent. Also, if the purpose of the analysis is better estimation of AL at individual sites, this is almost certainly not important. From a theoretical point of view, however, 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 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 a need to re-think 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 to 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 dataset by means of a prior distribution on complexity. Regarding p_D 's interpretability, its definition (Spiegelhalter et al 2002, Section 2.3) is opaque and the only concrete interpretation 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 $\mathbf{y}$, so its complexity can be defined without a realized $\mathbf{y}$. This is true of DF and of most complexity measures (e.g., Hastie & Tibshirani 1990, Ye 1998), but not p_D , the definition of which *requires* a realized $\mathbf{y}$ through the point estimate $\hat{\theta}$ and the posterior mean of the deviance $D(\theta)$. In special cases, such as normal hierarchical models with all covariances known, p_D does not depend on $\mathbf{y}$ and in these cases p_D agrees with DF (Spiegelhalter et al 2002, Sections 2, 4; Green 2002). But with unknown variances – i.e., in practical situations – p_D and DF diverge.

A complexity measure defined independently of the outcome $\mathbf{y}$ 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 it creates no difficulty to

describe a model's complexity as a function of its unknowns. The DF of a *fitted* model is obviously a function of the data, either through plug-in estimates or posterior distributions of the unknown parameters in the covariance matrices. But a complexity measure defined independently of $\mathbf{y}$, like DF, allows a prior distribution on complexity to control softly the complexity of the fit. This is not possible with p_D because in general it is defined in terms of a specific realized $\mathbf{y}$.

It is not clear p_D can be partitioned corresponding to components of the model, as DF can. Spiegelhalter et al (2002, Sections 6.3, 8.1) do 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 p_D 's relatively easy computation is certainly desirable, for our purposes this seems outweighed by its conceptual and practical disadvantages.

Acknowledgments

This work was supported in part by a Merck Company Foundation Quantitative Sciences Fellowship. 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 *Technometrics*'s web site. The fourth author's work was supported in part by National Cancer Institute grant 2-R01-CA095955-05A2.

References

- Besag J, York JC, Mollie A. Bayesian image restoration, with two applications in spatial statistics (with discussion). *Ann Inst Stat Math* 1991;43:1–59.
- Cui, Y. PhD dissertation. University of Minnesota Division of Biostatistics; 2008. Smoothing analysis of variance for general designs and partitioning degrees of freedom in hierarchical and other richly parameterized models.
- Daniels MJ. A prior for the variance in hierarchical models. *Can J Stat* 1999;27:569–580.
- Eberly LE, Carlin BP. Identifiability and convergence issues for Markov chain Monte Carlo fitting of spatial models. *Stat Med* 2000;19:2279–2294. [PubMed: 10960853]
- Gelman A. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis* 2006;1:515–534.
- Gelman A, Pardoe I. Bayesian measures of explained variance and pooling in multilevel (hierarchical) models. *Technometrics* 2006;48:241–251.
- Gilthorpe MS, Zamzuri AT, Griffiths GS, Maddick IH, Eaton KA, Johnson NW. Unification of the “burst” and “linear” theories of periodontal disease progression: A multilevel manifestation of the same phenomenon. *J Dent Res* 2003;82:200–205. [PubMed: 12598549]
- Green P. Discussion of Spiegelhalter et al (2002). *J Roy Stat Soc, Ser B* 2002;64:627–628.
- Groetsch, CW. *Inverse Problems in the Mathematical Sciences*. Vieweg: Wiesbaden; 1993.
- Hastie, TJ.; Tibshirani, RJ. *Generalized Additive Models*. Boca Raton FL: CRC/Chapman & Hall; 1990.
- Hodges JS, Carlin BP, Fan Q. On the precision of the conditionally autoregressive prior in spatial models. *Biometrics* 2003;59:317–322. [PubMed: 12926716]
- Hodges JS, Cui Y, Sargent DJ, Carlin BP. Smoothing balanced single-error-term analysis of variance. *Technometrics* 2007;49:12–25.
- Hodges JS, Sargent DJ. Counting degrees of freedom in hierarchical and other richly-parameterised models. *Biometrika* 2001;88:367–379.
- Lu H, Hodges JS, Carlin BP. Measuring the complexity of generalized linear hierarchical models. *Can J Stat* 2007;35:69–87.

- Natarajan R, Kass RE. Reference Bayesian methods for generalized linear mixed models. *J Amer Stat Assn* 2000;95:227–237.
- Ortega, JM.; Rheinboldt, WC. *Iterative Solution of Nonlinear Equations in Several Variables*. New York: Academic Press; 1970.
- Paciorek CJ, Schervish MJ. Spatial modelling using a new class of nonstationary covariance functions. *Environmetrics* 2006;17:483–506. [PubMed: 18163157]
- Plummer M. Penalized loss functions for Bayesian model comparison. *Biostatistics* 2008;9:523–539. [PubMed: 18209015]
- Prado R, West M, Krystal AD. Multichannel electroencephalographic analyses via dynamic regression models with time-varying lag-lead structure. *J Roy Stat Soc, Ser C* 2001;50:95–109.
- Reich BJ, Hodges JS. Modeling longitudinal spatial periodontal data: A spatially adaptive model with tools for specifying priors and checking fit. *Biometrics* 2008a;64:790–799. [PubMed: 18177463]
- Reich BJ, Hodges JS. Identification of the variance components in the general two-variance linear model. *J Stat Plan Inf* 2008b;138:1592–1604.
- Reich BJ, Hodges JS, Carlin BP. Spatial analyses of periodontal data using conditionally autoregressive priors having two types of neighbor relations. *J Amer Stat Assn* 2007;102:44–55.
- Reich BJ, Hodges JS, Zadnik V. Effects of residual smoothing on the posterior of the fixed effects in disease-mapping models. *Biometrics* 2006;62:1197–1206. [PubMed: 17156295]
- Ruppert, D.; Wand, MP.; Carroll, RJ. *Semiparametric Regression*. New York: Cambridge; 2003.
- Schott, JR. *Matrix Analysis for Statistics*. New York: Wiley; 1997.
- Scheffé, H. *The Analysis of Variance*. New York: Wiley; 1959.
- Spiegelhalter DJ, Best DG, Carlin BP, Van Der Linde A. Bayesian measures of model complexity and fit (with discussion). *J Roy Stat Soc, Ser B* 2002;64:583–639.
- Vaida F, Blanchard S. Conditional Akaike information criterion for mixed effects models. *Biometrika* 2005;92:351–370.
- Wall MW. A close look at the spatial structure implied by the CAR and SAR models. *J Stat Plan Inf* 2004;121:311–324.
- Weisberg, S. *Applied Linear Regression*. Wiley; New York: 1985.
- West, M.; Harrison, J. *Bayesian Forecasting and Dynamic Models*. 2. New York: Springer; 1997.
- Ye J. On measuring and correcting the effects of data mining and model selection. *J Amer Stat Assn* 1998;93:120–131.
- Zhao, YH. MS thesis. Division of Biostatistics, University of Minnesota; 1999. Examining mean structure, correlation structure, and differences between examiners in a large periodontal data set.

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.

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

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 $E(f_i|y) = 0.18, 0.52$, and 0.52 respectively.

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 .

Table 1

DF for $\mathbf{w}_1$ and $\mathbf{w}_2$ in the linear growth model in different scenarios ($\sigma_n = 1$)

Table 1a. Assuming $\sigma_j = 0$ for $j = 1$ or 2 .		
φ	DF($\mathbf{w}_1$) if $\sigma_1 = \varphi, \sigma_2 = 0$	DF($\mathbf{w}_2$) if $\sigma_1 = 0, \sigma_2 = \varphi$
0.01	0.1	3.4
0.1	4.8	13.1
1	54.3	47.4
10	120.6	116.4

σ_2	$\sigma_1 = 0.1$		$\sigma_1 = 1$	
	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

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. The maximum possible DF for δ_i and ξ_i are, respectively, 41 and 42 for individual subjects and 492 and 503 for all subjects combined.

source	$DF(\delta_i)$	$DF(\xi_i)$	$E(f_i/y)$
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

True DF and average posterior median DF for clustering and heterogeneity, for various true $(\sigma_a^2/\sigma_0^2, \sigma_c^2/\sigma_0^2, \sigma_h^2/\sigma_0^2)$, with 100 simulated datasets per scenario. “Est DF” is the average of the 100 posterior medians; “MC SE” is the Monte Carlo standard error of that estimate.

Table 3

$\frac{\sigma_a^2}{\sigma_0^2}, \frac{\sigma_c^2}{\sigma_0^2}, \frac{\sigma_h^2}{\sigma_0^2}$	True DF		Est DF		MC SE	
	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