Spatial smoothing of autocorrelations to control the degrees of freedom in fMRI analysis

K.J. Worsley

March 10, 2005

Department of Mathematics and Statistics, McGill University, 805 Sherbrooke St. West, Montreal, Québec, Canada H3A 2K6, and McConnell Brain Imaging Centre, Montreal Neurological Institute, 3801 University Street, Montreal, Québec, Canada H3A 2B4.
tel: 1-514-398-3842, fax: -3899, web: http://www.math.mcgill.ca/keith

---

In the statistical analysis of fMRI data, the parameter of primary interest is the effect of a contrast; of secondary interest is its standard error, and of tertiary interest is the standard error of this standard error, or equivalently, the degrees of freedom (df). In a ReML (Restricted Maximum Likelihood) analysis, we show how spatial smoothing of temporal autocorrelations increases the effective df (but not the smoothness of primary or secondary parameter estimates), so that the amount of smoothing can be chosen in advance to achieve a target df, typically 100. This has already been done at the second level of a hierarchical analysis by smoothing the ratio of random to fixed effects variances (Worsley *et al.*, 2002); we now show how to do it at the first level, by smoothing autocorrelation parameters. The proposed method is extremely fast and it does not require any image processing. It can be used in conjunction with other regularization methods (Gautama & Van Hulle, 2004) to avoid unnecessary smoothing beyond 100 df. Our results on a typical 6 minute, TR=3, 1.5T fMRI data set show that 8.5mm smoothing is needed to achieve 100 df, and this results in roughly a doubling of detected activations.

---

1 Introduction

One of the simplest models for the first level of fMRI data analysis is the linear model with AR(p) temporal correlation structure (Bullmore *et al.*, 1996; Locascio *et al.*, 1997; Woolrich *et al.*, 2001; Worsley *et al.*, 2002; Marchini & Smith, 2003). The AR(p) parameters are estimated separately at each voxel, then spatially smoothed to reduce their variability, at the cost of slightly increasing their bias (see Figure 1). Until recently, the amount of smoothing was chosen heuristically (e.g. 15mm by the `FMRISTAT` software), but Gautama & Van Hulle (2004) have now introduced a method to estimate the amount of smoothing in a principled fashion. Their method is based on choosing the amount of smoothing to best predict the autocorrelations.

In this paper we show how the amount of smoothing influences the effective degrees of freedom (df) of ReML (Restricted Maximum Likelihood) estimators (Harville, 1974). The

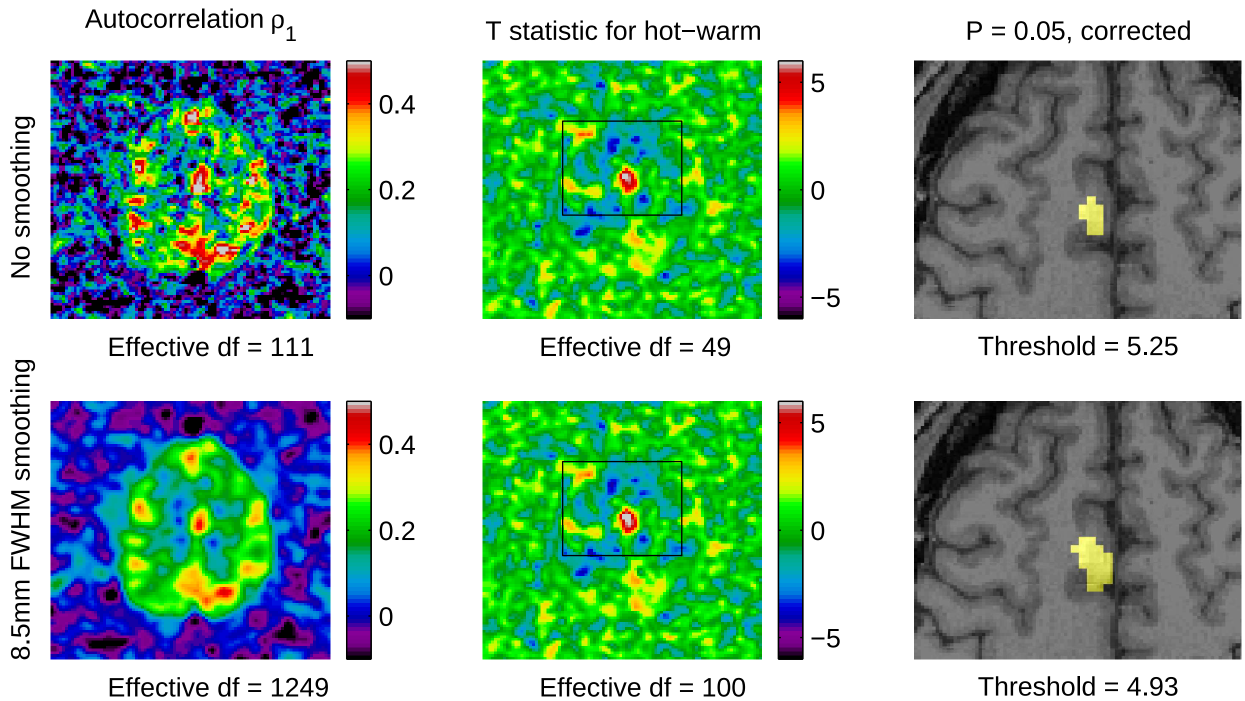

Figure 1: Temporal autocorrelation (lag 1) of fMRI data from a pain perception experiment, without (top row) and with (bottom row) spatial smoothing, the corresponding T statistics for a hot-warm effect, and the detected activation (magnified region framed on the T statistic). Note that with modest smoothing, the effective df increases, the resulting $P = 0.05$ threshold decreases, and roughly twice as much activation is detected.

theory uses the same techniques as Kenward & Roger (1997) and Kiebel *et al.* (2003). Kenward and Roger were concerned with ReML estimators (close to what `FMRISTAT` and `FSL` use), whereas Kiebel applied these techniques to least-squares estimators (close to what `SPM` uses). We show how the df can be approximated *before* the analysis is carried out, so that in conjunction with other regularization methods (e.g. Gautama & Van Hulle, 2004), we can choose the amount of smoothing in advance to achieve a targeted df. Curiously enough, this df is different for different contrasts, and different again for an F statistic that combines contrasts, so if more than one inference is desired from a single smoothing, then we suggest being conservative and taking the maximum. The proposed method is extremely fast and, unlike the method of Gautama and Van Hulle (2004), it does not require any image processing.

2 Method

We adopt a linear model for the mean of the fMRI data, with an AR(p) model for the temporal correlation structure (6). This is fitted by the pre-whitening method used by Worsley *et al.* (2002), similar to that used by the `FSL` software, and close to ReML, which is summarized as follows. The model is first fitted by least-squares to find least squares residuals. These residuals are used to calculate the temporal autocorrelations, and a simple method is used to correct for the bias incurred by using residuals from the fitted rather than the true model. The temporal autocorrelations are spatially smoothed to reduce their variability, then inserted into the Yule-Walker equations to derive the AR(p) model coefficients. The fitted AR(p) model is used to whiten the data and the covariates, and the model is re-fitted. Effects of contrasts in the coefficients, and their estimated standard errors S (13), are calculated, leading to T and F statistics, and our statistical inference.

The effect of the spatial smoothing on the effective df is as follows. Suppose X is the $n \times m$ design matrix of the linear model whose columns are the covariates, and let c be an m-vector of contrasts in those columns whose effects we are interested in. Let

$$x = (x_1, \ldots, x_n)' = X(X'X)^{-1}c \tag{1}$$

be the least-squares contrast in the *observations*, and let τ_j be its lag j autocorrelation

$$\tau_j = \sum_{i=j+1}^{n} x_i x_{i-j} / \sum_{i=1}^{n} x_i^2. \tag{2}$$

Let $FWHM_{\text{data}}$ be the effective FWHM of the fMRI data, and let $FWHM_{\text{filter}}$ be the FWHM of the Gaussian filter used for spatial smoothing of the temporal autocorrelations. Let

$$f = \left(1 + 2\frac{FWHM^2_{\text{filter}}}{FWHM^2_{\text{data}}}\right)^{-D/2} \tag{3}$$

where D is the number of dimensions. Then our main result, proved in the Appendix, is that the effective df of the contrast is

$$\tilde{\nu} \approx \nu / (1 + 2f \sum_{j=1}^{p} \tau_j^2) \tag{4}$$

where $\nu = n - m$ is the usual least-squares residual df. For an F statistic that simultaneously tests k columns of the $m \times k$ contrast matrix C, the effective numerator df is the same as (4) but with x replaced by the normalized matrix of the least-squares contrasts in the observations

$$x = X(X'X)^{-1}C(C'(X'X)^{-1}C)^{-1/2}, \tag{5}$$

so that $x'x$ is the $k \times k$ identity matrix, and with the autocorrelation τ_j replaced by the average of the k temporal autocorrelations of the columns of x. Finally the effective df of the smoothed autocorrelation itself, defined as the number of independent observations that would produce the same variance without smoothing, is ν/f.

Temporal correlation of the covariates decreases effective df, but since $f \leq 1$, spatial smoothing ameliorates this effect. Reversing this formula (4), we can calculate the amount of smoothing required to achieve a desired df. Note that this will depend on the contrast, so we suggest being conservative by taking the maximum of the amounts of smoothing.

Note that we can never get more than the least-squares df without smoothing the residual variance as well, in which case the factor f would be applied to all the terms in the denominator of (4). Of course we do not wish to this because the residual variance contains too much anatomical structure, and so smoothing could result in serious biases.

The conditions for the result (4) to hold are that the sample size n must be large and the temporal correlations must be small. To see how well this approximation holds up when these conditions are relaxed, we carry out some simulations in the next section.

3 Results

3.1 Simulations

The above theoretical effective df (4) is derived under the assumption that the sample size is large and the temporal autocorrelations ρ_j are small. In practice sample sizes of at least 100 are common, so small sample size is not a serious problem. Fortunately 1.5T fMRI data is well described by an AR(1) model (Worsley *et al.*, 2002) with lag 1 temporal autocorrelations $\rho_1 \sim 0.3$ in grey matter (see e.g. Figure 1). To assess the robustness of our theoretical results to non-zero ρ_1, we simulated null AR(1) fMRI noise data with ρ_1 varying from 0 to 0.4 (N$(0,1)$ at each of 64^3 voxels, $FWHM_{\mathrm{data}} = 5$ voxels spatial smoothing, $n = 120$ temporally correlated frames).

We chose as a paradigm the block design from a pain perception experiment fully described in Worsley *et al.* (2002): 9s rest, 9s hot, 9s rest, 9s hot stimuli repeated 10 times, the same one used by Gautama & Van Hulle (2004). This was convolved with a canonical difference of gammas HRF. Cubic time trends were added to account for drift, giving $m = 6$ linear model parameters and $\nu = 114$ least-squares residual df. Four contrasts were chosen: hot, hot+warm, hot-warm, and the cubic drift term.

After fitting each model by least-squares, the bias-corrected autocorrelation $\widehat{\rho}_1$ was spatially smoothed by different amounts. We then estimated $\mathrm{E}(S^2)$ and $\mathrm{Var}(S^2)$ by taking the sample mean and variance of the image of variance estimates S^2 (13) over all 64^3 voxels. The effective df was then estimated by (17).

Figure 2 shows the simulated effective df $\tilde{\nu}$ plotted against the ratio $FWHM_{\mathrm{filter}}/FWHM_{\mathrm{data}}$, together with the theoretical effective df (4). As expected, effective df increases with increasing smoothing, reaching the least-squares df for very high smoothing. Interestingly, the effective df is lowest for the smoothest contrasts.

The most interesting question is how this varies with the temporal autocorrelation of the data, since our theoretical result assumes this is small. Fortunately we can see that the variability is not too great, particularly up to $\rho_1 = 0.3$, the most common value encountered in 1.5T data. For 3T data AR(1) models fit reasonably well with higher ρ_1, but fortunately our results suggest that the true df is slightly higher than the theoretical df, which means that inference using the theoretical df will be slightly conservative. Note however that this conclusion is reversed for the cubic drift contrast, but these types of smooth contrast will be of little interest for most fMRI experiments precisely because they will be confounded with drift.

Finally we illustrate how the theoretical effective df (4) can be used for targeting a particular df. Figure 2 shows the amount of smoothing required to achieve 100 df. It depends on the contrast: for the highest frequency contrast (hot+warm), a filter FWHM of 0.81 times that of the data FWHM will gaurantee 100 df; for the cubic drift term, 1.49 times the data FWHM will be needed.

3.2 Application

We applied this result to the same fMRI data as Gautama & Van Hulle (2004), fully described in Worsley *et al.* (2002). The paradigm and model is the same as above, but the first three frames were discarded due to scanner instabilities, so that $n = 117$ and $\nu = 111$. After motion correction, each frame of BOLD data was converted to percent of whole brain BOLD average. We took as the spatial correlation of the data $FWHM_{\mathrm{data}} = 6$mm which is the amount of smoothing applied during motion correction. Note however that this is probably an

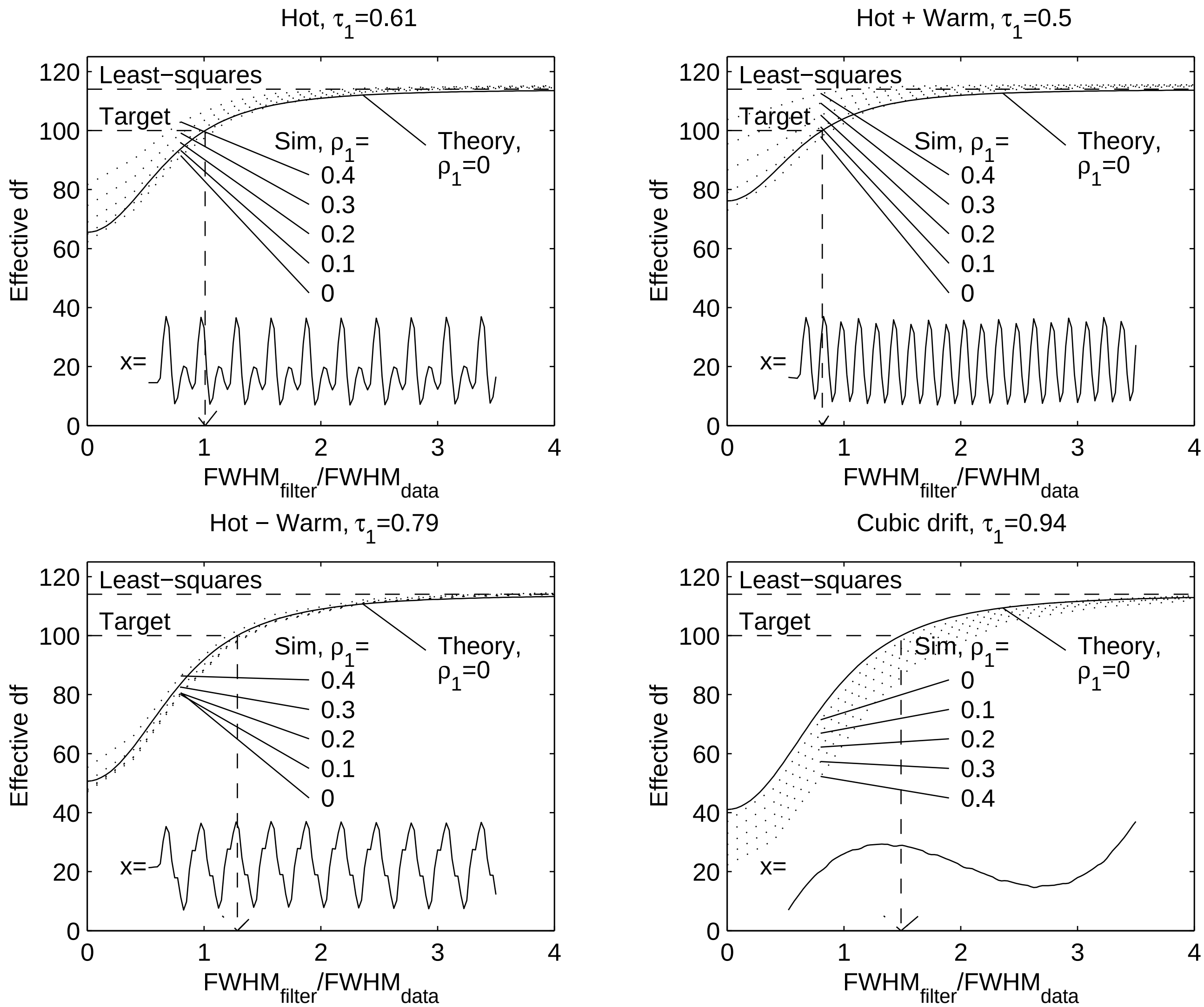

Figure 2: Validation of the theoretical effective df $\tilde{\nu}$ (solid line) by comparison with simulations (dotted lines). Four different contrasts were considered; the time course of each contrast in the observations, x, is plotted in the lower part (τ_1 is its lag 1 autocorrelation). Effective df is plotted against the spatial smoothing of the temporal autocorrelations, relative to the spatial smoothing of the fMRI data. Smoothing increases the effective df, up to the least-squares df $\nu = n - m = 120 - 6 = 114$ (dashed line). We can see that the theoretical result, derived assuming the autocorrelation ρ_1 is zero, is a reasonable approximation even if the autocorrelation is not zero. Using this, we can find the amount of smoothing needed to target a particular df, here 100 (dashed arrow).

underestimate - fMRI data is more spatially smooth in grey matter (Worsley, 2002), averaging closer to 8mm FWHM over the whole brain.

Without spatial smoothing, the effective df was $\tilde{\nu} = 49$ for the hot-warm contrast. Note that because different slices were acquired at different times, the covariates were slightly different, so this represents an average over all slices. Reversing this relationship, we calculated $FWHM_{\text{filter}} = 8.5$mm needed to target 100 df. Note that this is much smaller than the 15mm smoothing used in Worsley *et al.* (2002), and much closer to the $\sim$ 7mm smoothing found by Gautama & Van Hulle, which wou1d result in $\sim$ 95 df. Finally the effective df of the temporal correlation increased from $\nu = 111$ unsmoothed to $\nu/f = 1249$ smoothed.

Figure 1 shows the lag 1 temporal autocorrelation with and without smoothing, and the resulting T statistics. The T statistics look very similar, but because the T statistic with smoothed temporal autocorrelation has higher df, it is not quite as variable, and so its $P = 0.05$ threshold (corrected) is lower (4.93 vs. 5.25). This results in more detected activation: the volume of the T image above threshold increased from 2.5cc without smoothing to 4.7cc with smoothing, roughly a doubling of detected activations. Note again that the T image itself is *not* smoothed; it is the temporal correlation structure that is smoothed.

4 Discussion

Spatial smoothing is not new to fMRI statistical analysis; many authors have advocated spatial smoothing of the data prior to analysis in order to enhance the signal to noise ratio, or as a by-product of motion correction, or to cover up inter-subject registration errors. In this paper we do not smooth the original data, nor the effect of a contrast, nor its estimated standard error; instead we suggest smoothing temporal autocorrelations. The purpose of the smoothing is to decrease errors in the estimated standard errors, or in other words, to increase the effective degrees of freedom (df). This results in less variable T statistic images, in turn giving lower P=0.05 thresholds, and finally resulting in more detected activations. It also has the added benefit of making the effective df less dependent on the underlying temporal correlation structure, hence less spatially variable, thus making statistical inference less problematic.

This raises the interesting issue of what regularization criterion should be used for secondary parameters (sd's). Regularizing to improve predictability of autocorrelations, either empirically by cross-validation (Gautama & Van Hulle, 2004), or by various information criteria (AIC, BIC) may not be directly relevant to our main purpose which is inference about primary parameters (contrasts). The reason for this is that regularizing to decrease the variability of the sd (increase its df) has little effect on inference about contrasts beyond 100 df. But regularizing beyond 100 df may introduce bias into the sd, which would then affect inference about contrasts. In other words, smoothing beyond 100 df is unnecessary; we may gain predictability of the autocorrelations, which would serve no directly useful purpose, but we would run the risk of increasing bias, which could be detrimental. The proposed method might be useful in conjunction with other regularization methods to impose an upper limit (say 100 df) on the amount of smoothing.

On the other hand, when regularizing a contrast, we clearly would like to have as low variability as possible, provided we don't pay too much of a price in increased bias, so any of the above information criteria are directly relevant. Purdon *et al.* (2001) go one step further and regularize the entire model, including all primary, secondary and tertiary parameters, by smoothing the log likelihood. The amount of smoothing is chosen to minimize an information criterion similar to AIC.

This paper gives a principled method of choosing the amount of smoothing. The idea is to choose the amount of smoothing to ensure at least 100 df. Why 100 df? The formula for the effective df still depends on a number of assumptions, such as Gaussian spatial correlation, and small voxel size relative to FWHM. If these assumptions are not met, there could be substantial errors in the effective df. But by targeting a high df, like 100, substantial errors in the df will not have much effect on inference. The reason is that beyond 50 df, the T statistic is very close to a Gaussian. Even if the effective df is in error by ±50%, i.e. the true df is 50 or 200, it will not substantially affect our inference.

This strategy may not result in any smoothing at all; if in the example above we had twice as many observations over a 12 minute time period then the effective df would be more than 100 without any smoothing, so no smoothing would be necessary. If on the other hand we had a short sequence with less than 100 least-squares df, then no amount of smoothing can increase this above 100 (without smoothing the variance itself). In this case, we recommend choosing the amount of smoothing to target a high proportion, say 90%, of the least-squares df. This strategy has been implemented in the `fmrilm` function of the `FMRISTAT` package, available from `http://www.math.mcgill.ca/keith/fmristat`.

Most FMRI papers in the literature are group studies. Our proposed method, along with any other method whose main purpose is to regularizes the sd, will have little impact on second level analyses beyond possibly improving efficiency, particularly if there is a substantial random subject effect. Of course, there are always exceptions where improving the quality of inadequate data in individuals may be important: presurgical planning or other activation-based ROI selection applications such as various types of "connectivity" analyses.

Finally we note that spatial smoothing has been usefully employed at the second level of an fMRI analysis, again to increase df. Worsley *et al.* (2002) suggest spatially smoothing the ratio of random to fixed effects variance, deriving a formula for determining the resulting effective df. Typically the problem here is more severe, because the random effects df can be very low (∼3), which necessitates much more smoothing (∼19mm) to target 100 df.

In both cases there is an attendant side-effect: too much smoothing will introduce bias, which might result in more false positives. But as in most statistical analyses there is a trade-off; we prefer to tolerate a small amount of bias in the hopes of discovering more true activation.

A Appendix

A.1 Effective df

The linear model can be written in matrix notation as

$$Y \sim \mathrm{N}_n(X\beta, V(\theta)), \tag{6}$$

where Y is an n-vector of fMRI observations at a single voxel; X is an $n \times m$ matrix whose columns are the covariates, temporal drift, or other explanatory variables; β is an m-vector of unknown coefficients; and V is an $n \times n$ variance matrix parameterized by an unknown q-vector θ. The method of choice for fitting such a model is ReML (Harville, 1974), which chooses parameters β, θ to minimize

$$(Y - X\beta)'\Lambda(Y - X\beta) - \log|\Lambda| + \log|X'\Lambda X(X'X)^{-1}|, \tag{7}$$

where $\Lambda = V^{-1}$ (upside down V). The resulting approximate covariances of the parameter estimators for large n are

$$\begin{aligned} \mathrm{Var}(\widehat{\beta}) &\approx (X'\Lambda X)^{-1}, && (8)\\ \mathrm{Var}(\widehat{\theta}) &\approx [\mathrm{tr}(R_V \dot{V}_j R_V \dot{V}_k)/2]^{-1}, && (9)\\ \mathrm{Cov}(\widehat{\beta}, \widehat{\theta}) &\approx 0, && (10) \end{aligned}$$

where $R_V = \Lambda - \Lambda X(X'\Lambda X)^{-1}X'\Lambda$, $\dot{V}_j = \partial V/\partial \theta_j$, and $[a_{jk}]$ denotes the matrix whose jkth element is a_{jk}. Usually we are only interested in an effect

$$E = c'\widehat{\beta}, \qquad (11)$$

where c is a known m-vector of contrasts. From (8), its variance for large n is

$$\mathrm{Var}(E) \approx c'(X'\Lambda X)^{-1}c. \qquad (12)$$

The usual plug-in estimator of this variance is

$$S^2 = \widehat{\mathrm{Var}}(E) = c'(X'\widehat{\Lambda}X)^{-1}c \qquad (13)$$

where $\widehat{\Lambda} = V(\widehat{\theta})^{-1}$. For large n this is approximately unbiased, i.e.

$$\mathrm{E}(S^2) \approx c'(X'\Lambda X)^{-1}c. \qquad (14)$$

We can find an approximate variance of S^2 by the classical method of a first-order Taylor-series expansion of (13) as a function of $\widehat{\theta}$ about its true value:

$$\mathrm{Var}(S^2) \approx (\dot{S}^2)'\mathrm{Var}(\widehat{\theta})\dot{S}^2, \qquad (15)$$

where

$$\dot{S}^2 = \frac{\partial S^2}{\partial \widehat{\theta}} = [c'(X'\Lambda X)^{-1}X'\Lambda \dot{V}_j \Lambda X(X'\Lambda X)^{-1}c]. \qquad (16)$$

Finally the effective df for the contrast can be found by the usual Satterthwaite method:

$$\tilde{\nu} = 2\mathrm{E}(S^2)^2/\mathrm{Var}(S^2), \qquad (17)$$

identical to that of Kenward & Roger (1997). Note that the effective df depends on the contrast c as well as the unknown variance parameters θ. This leads to another unfortunate consequence: the effective df varies spatially, since the variance parameters vary spatially, making statistical inference on the image awkward. Spatial smoothing of the temporal correlation structure will reduce variability in $\widehat{\theta}$, increase the effective df, reduce the spatial variability in df, and make statistical inference easier. To quantify this, we shall consider a special model, the AR(p) model.

A.2 AR(p) model

A particular case of V is the AR(p) model, where $q = p + 1$ and the inverse of V can be well approximated for large n (apart from edge effects) by a linear model

$$\Lambda = V^{-1} = \sum_{j=0}^{p} D_j \theta_j \qquad (18)$$

where $D_0 = I$ is the identity, and for $j \geq 1$, D_j is an $n \times n$ matrix whose $\pm j$th off-diagonal is 1/2 and the rest is 0, and $\theta = (\theta_0, \ldots, \theta_p)$. V^{-1} has the form (18) because the elements of V^{-1} are partial correlations, conditional on the rest of the observations, and partial correlations at lags greater than p, conditional on the intervening observations, are zero. Note that $\theta_0, \theta_1, \ldots, \theta_p$ are not the variance and autocorrelations $\sigma^2, \rho_1, \ldots, \rho_p$ but some awkward parametrization of them - see (29). An interesting feature of this model is that the estimated values of θ depend on the data only through the sample variance and temporal autocorrelations out to lag p. This can be seen by substituting the AR(p) model (18) directly into the ReML criterion (7) and observing that $(Y - X\beta)' D_j (Y - X\beta)/n$ are the temporal autocovariances.[1]

Spatial smoothing of the temporal autocorrelations will reduce their variability, thus reducing the variability of $\widehat{\theta}$, and hopefully increasing the effective df $\tilde{\nu}$. Unfortunately the exact effect of smoothing will depend on the true variance parameters, which are not known in advance of fitting the model. As a first approximation, it seems reasonable to try to estimate the effective df for some simple model, hopefully not too far from the true model, then check to see how robust this estimator is to departures from the simple model. This is the approach we take here.

The simplest model is when the observations are uncorrelated, that is, $V = I\sigma^2$, so that $\theta = (1/\sigma^2, 0, \ldots, 0)$. In this case, if we still (unnecessarily) fit an AR(p) model, the resulting variance of $\widehat{\theta}$ from (9) is

$$\begin{aligned} \mathrm{Var}(\widehat{\theta}) &\approx [\mathrm{tr}(R_I D_j R_I D_k)\sigma^4/2]^{-1} \\ &\approx \mathrm{diag}(1, 2, \ldots, 2)\ 2/(\nu\sigma^4) \end{aligned} \tag{19}$$

if we further assume that n is large relative to m, so that $R_I \approx I$, and where $\nu = \mathrm{tr}(R_I) = n - m$ is the usual (least-squares) df. We now have a simple rough approximation to the variability of $\widehat{\theta}$ which does not depend on the true (unknown) temporal correlations. We now proceed to use this to estimate the effective df. First let

$$x = (x_1, \ldots, x_n)' = X(X'X)^{-1}c \tag{20}$$

be the least-squares contrast in the *observations*, and let τ_j be its lag j autocorrelation

$$\tau_j = x' D_j x / x'x = \sum_{i=j+1}^{n} x_i x_{i-j} / \sum_{i=1}^{n} x_i^2. \tag{21}$$

To calculate the effective df as above, the last thing we need is the derivative of (18)

$$\dot{V}_j = -V D_j V. \tag{22}$$

Then from (14), (16), (19), (21) and(22),

$$\mathrm{E}(S^2) \approx x'x\sigma^2, \tag{23}$$

$$\dot{S}^2 = -(1, \tau_1, \ldots, \tau_p) x'x\sigma^4, \tag{24}$$

$$\mathrm{Var}(S^2) \approx (2/\nu)\sigma^4 (x'x)^2 (1 + 2\sum_{j=1}^{p} \tau_j^2), \tag{25}$$

$$\tilde{\nu} \approx \nu / (1 + 2\sum_{j=1}^{p} \tau_j^2). \tag{26}$$

[1]For those familiar with Generalized Linear Models (McCullagh and Nelder, 1983), this is a simple consequence of the fact that the canonical link for the Wishart distribution of the data $(Y - X\beta)(Y - X\beta)'$ is the inverse link function (18), and when the link is canonical then the inner products of the data with the regressors D_j are minimal sufficient for the maximum likelihood estimators.

This last is exactly what we are looking for - a simple workable expression for the effective df that does not depend on the observations, just the covariates and their contrast. We now look at the effect of spatial smoothing on this.

A.3 Autocorrelations

Recall that $\widehat{\theta}$ depends on the data only through the sample variance and the autocorrelations up to lag p. Worsley *et al.* (2002) suggest spatially smoothing the autocorrelations (but not the variance) to reduce their variability *before* fitting the model. The estimator employed is almost but not quite the ReML estimator, but nevertheless we shall try approximating the variance of the estimator by the approximate ReML variances above.

To get an idea of the effect of smoothing, we shall assume that the sample autocorrelations are locally homogeneous, so that smoothing has little effect on their mean, only their variance. This implies that $\mathrm{E}(S^2)$ and $\dot{S}^2$ are largely unaffected, and the only effect is on $\mathrm{Var}(\widehat{\theta})$. Again the exact dependence of $\widehat{\theta}$ on the sample variance and autocorrelations is complicated, but fortunately there is a very simple relationship if the true correlation structure is close to independence as we assumed above, that is, if $\theta_1, \ldots, \theta_p$ are all small. First let

$$r = (r_1, \ldots, r_n)' = R_I Y \sim \mathrm{N}(0, R_I V R_I) \tag{27}$$

be the least-squares residuals, let $\widehat{\sigma}^2 = r'r/\nu$ be their sample variance, and let $\widehat{\rho}_j$ be their lag j autocorrelation

$$\widehat{\rho}_j = r' D_j r / r'r = \sum_{i=j+1}^{n} r_i r_{i-j} \Big/ \sum_{i=1}^{n} r_i^2. \tag{28}$$

For large n, $R_I V R_I \approx V$ so that the sample variance and autocorrelations $\widehat{\rho}_j$ will be approximately unbiased for the true lag j autocorrelation ρ_j. If n is not small, we can apply the bias correction of Worsley *et al.* (2002). Then

$$\widehat{\theta}_j \approx \begin{cases} 1/\widehat{\sigma}^2, & j = 0 \\ -2\widehat{\rho}_j/\widehat{\sigma}^2, & j = 1, \ldots, p. \end{cases} \tag{29}$$

There are several ways of seeing this. One could work through the solution to the Yule-Walker equations to get estimators of the AR(p) parameters, insert them into the lower triangular matrix $(\widehat{V}^{-1/2})'$, then calculate $\widehat{\Lambda} = (\widehat{V}^{-1/2})'\widehat{V}^{-1/2}$. A more direct method is to note that since $\widehat{\theta}_1, \ldots, \widehat{\theta}_p$ will all be small then, using a first-order Taylor-series expansion,

$$\widehat{V} = (I\widehat{\theta}_0 + \sum_{j=1}^{p} D_j \widehat{\theta}_j)^{-1} \approx I/\widehat{\theta}_0 - \sum_{j=1}^{p} D_j \widehat{\theta}_j / \widehat{\theta}_0^2 \tag{30}$$

from which we infer (29) immediately. We can now see from (29), again if n is large so that variability in $\widehat{\sigma}$ is small relative to its mean, that $\widehat{\theta}_j$ is approximately proportional to $\widehat{\rho}_j$. This combined with the fact, from (19), that the components of $\widehat{\theta}$ are approximately uncorrelated, implies any reduction in variability of the autocorrelations will invoke a corresponding proportional reduction in variability of $\widehat{\theta}_j$, $j = 1, \ldots, p$.

A.4 Spatial smoothing

All that remains is to relate the amount of spatial smoothing of the autocorrelations to their variability. A little random field theory supplies the answer. We first show that autocovariances, like variances, are $1/\sqrt{2}$ times less smooth then their component random fields. In what follows, we shall add an argument (s) to indicate spatial location s, where necessary. Consider term i of the autocovariance at lag j at spatial location s, and its spatial covariance with another term k at spatial location t. From the covariance of elements of a Wishart random matrix:

$$\begin{aligned}\mathrm{Cov}(r_i(s)r_{i-j}(s), r_k(t)r_{k-j}(t)) &= \mathrm{Cov}(r_i(s), r_k(t))\mathrm{Cov}(r_{i-j}(s), r_{k-j}(t)) \\ &\quad + \mathrm{Cov}(r_i(s), r_{k-j}(t))\mathrm{Cov}(r_{i-j}(s), r_k(t)). \end{aligned} \tag{31}$$

We shall assume that the spatio-temporal correlation structure is locally separable, that is, the same spatial correlation function $h(t)$ at every time, and the same temporal correlation at every voxel:

$$\mathrm{Cov}(r(s), r(t)) = \mathrm{Var}(r)h(s-t). \tag{32}$$

Then from (31) and (32), the spatial covariance of the numerator of $\widehat{\rho}_j$ is

$$\mathrm{Cov}(\sum_{i=j+1}^{n} r_i(s)r_{i-j}(s), \sum_{k=j+1}^{n} r_k(t)r_{k-j}(t)) \propto h(s-t)^2. \tag{33}$$

Since the denominator is approximately constant for large n, this immediately implies that

$$\mathrm{Cor}(\widehat{\rho}_j(s), \widehat{\rho}_j(t)) \approx h(s-t)^2. \tag{34}$$

i.e. the spatial correlation function of $\widehat{\rho}_j$ is $h(t)^2$. If we model the errors as white noise smoothed with a Gaussian kernel of FWHM $FWHM_{\mathrm{data}}$, then $h(t)$ is Gaussian with FWHM multiplied by $\sqrt{2}$. Squaring a Gaussian divides its FWHM by $\sqrt{2}$, so $h(t)^2$ is Gaussian with FWHM $FWHM_{\mathrm{data}}$. Again for large n, sample autocorrelations are approximately Gaussian so, working backwards, their effective FWHM is approximately $FWHM_{\mathrm{data}}/\sqrt{2}$.

Smoothing a Gaussian random field with FWHM $FWHM_{\mathrm{data}}/\sqrt{2}$ with a filter with FWHM $FWHM_{\mathrm{filter}}$ multiplies its variance by a factor of

$$f = \left(1 + 2\frac{FWHM_{\mathrm{filter}}^2}{FWHM_{\mathrm{data}}^2}\right)^{-D/2} \tag{35}$$

where D is the number of dimensions (Worsley *et al.*, 2002). In the light of (29), and again for large n, we conclude that $\mathrm{Var}(\widehat{\theta}_1, \ldots, \widehat{\theta}_p)$ is multiplied by f, hence from (15), (17), (23) and (24) we get the result we want: for large n, and when the errors are approximately uncorrelated, the effective df is

$$\tilde{\nu} \approx \nu/(1 + 2f\sum_{j=1}^{p} \tau_j^2). \tag{36}$$

The effective df of the smoothed AR(p) coefficients may be defined as the number of independent observations that would produce the same variance without smoothing. Since the effect of smoothing the autocorrelations is to multiply the variances of the AR(p) coefficients by f as noted above, then their effective df is the residual df divided by f, namely ν/f.

A.5 F statistics

The effective df of F statistics is more complicated. Several asymptotically equivalent ones have been proposed in the statistics literature, but perhaps the most commonly used is that proposed by Kenward & Roger (1997) which has the property that it gives the correct answer in the special case of Hotelling's T^2, which has an exact F distribution for all sample sizes. Since our aim is to get a rough approximation for large sample sizes and small temporal correlation, we shall adopt one of the simpler methods. The idea is to find the effective df for a range of contrasts that span the contrast space, then take the average. The details are as follows.

Suppose we are interested in simultaneously testing k contrasts. Let C be a $m \times k$ matrix whose columns are the contrasts. Then the F statistic is

$$F = E'(C'(X'\widehat{\Lambda}X)^{-1}C)^{-1}E, \tag{37}$$

where $E = C'\widehat{\beta}$ is the k-vector of effects. The F statistic does not clearly separate into a numerator and denominator, as in the case of the T statistic, so it is not easy to get at the denominator df. The numerator df is governed by E, so instead we shall condition on fixed E, which is roughly independent of S by (10), work out the effective df of what is left, then take expectations over E. Using the same methods as above, it is not hard to show that the effective df conditional on E is the same expression as (36) but with x replaced by

$$x = X(X'X)^{-1}C(C'(X'X)^{-1}C)^{-1}E. \tag{38}$$

The approximate large sample distribution of E when the temporal correlations are small is $E \sim \mathrm{N}(0, (C'(X'X)^{-1}C)^{-1})$, so the null expectation of the autocovariance of x is

$$\mathrm{E}(x'D_jx) = \mathrm{tr}((C'(X'X)^{-1}C)^{-1}C'(X'X)^{-1}X'D_jX(X'X)^{-1}C) \tag{39}$$

and simply k for the variance ($j = 0$). We thus get the same expression as (36) for the unconditional effective df but with τ_j replaced by the average autocorrelation

$$\tau_j = \mathrm{E}(x'D_jx)/k. \tag{40}$$

References

Bullmore, E., Brammer, M., Williams, S., Rabe-Hesketh, S., Janot, N., David, A., Mellers, J., Howard, R. & Sham, P. (1996). Statistical methods of estimation and inference for functional MR image analysis. *Magnetic Resonance in Medicine*, **35**:261-277.

Gautama, T. & Van Hulle, M.M. (2004). Optimal spatial regularisation of autocorrelation estimates in fMRI analysis. *NeuroImage*, **23**:1023-1216.

Harville, D.A. (1974). Bayesian inference for variance components using only error contrasts. *Biometrika*, **61**:383-385.

Kenward, M.G. & Roger, J.H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, **53**:983-997.

Kiebel, S.J., Glaser, D.E. & Friston, K.J. (2003). A heuristic for the degrees of freedom of statistics based on multiple variance parameters. *NeuroImage*, **20**:591-600.

Locascio, J., Jennings, P., Moore, C. & Corkin, S. (1997). Time series analysis in the time domain and resampling methods for studies of functional magnetic resonance brain imaging. *Human Brain Mapping*, **5**:168-193.

Marchini, J. & Smith, S. (2003). On bias in the estimation of autocorrelations for fMRI voxel time-series analysis. *NeuroImage*, **18**:83-90.

McCullagh, P. & Nelder, J.A. (1983). *Generalized Linear Models.* London: Chapman and Hall.

Purdon, P.L., Solo, V., Weisskoff, R.M. & Brown, E. (2001). Locally regularized spatiotemporal modeling and model comparison for functional MRI. *NeuroImage* **14**:912-923.

Woolrich, M., Ripley, B., Brady, M. & Smith, S. (2001). Temporal autocorrelation in univariate linear modeling of FMRI data. *NeuroImage* **14**:13701386.

Worsley, K.J., Liao, C., Aston, J.A.D., Petre, V., Duncan, G.H., Morales, F. & Evans, A.C. (2002). A general statistical analysis for fMRI data. *NeuroImage*, **15**:1-15.

Worsley, K.J. (2002). Non-stationary FWHM and its effect on statistical inference of fMRI data. *NeuroImage*, 15:S346.