Parametric Inference for FMRI Time Series with Correlated Errors

John Carew

jcarew@stat.wisc.edu

Departments of Statistics and

Biostatistics and Medical Informatics

University of Wisconsin – Madison

Outline

- The linear regression model and SPM.
- Correlated errors: A concrete example.
- Smoothing splines and generalized cross-validation.
- Applying the spline model to fMRI.
- Results from real data and simulation.
- Discussion.

A Statistical Model

What is a statistical model?

- Formally, a statistical model is a collection of probability distributions.
- A parametric model is a statistical model and a parameterization that maps a space of labels to the statistical model.
- Under certain conditions, elements of the label space are called parameters. These parameters index probability distributions i.e., elements of the statistical model.

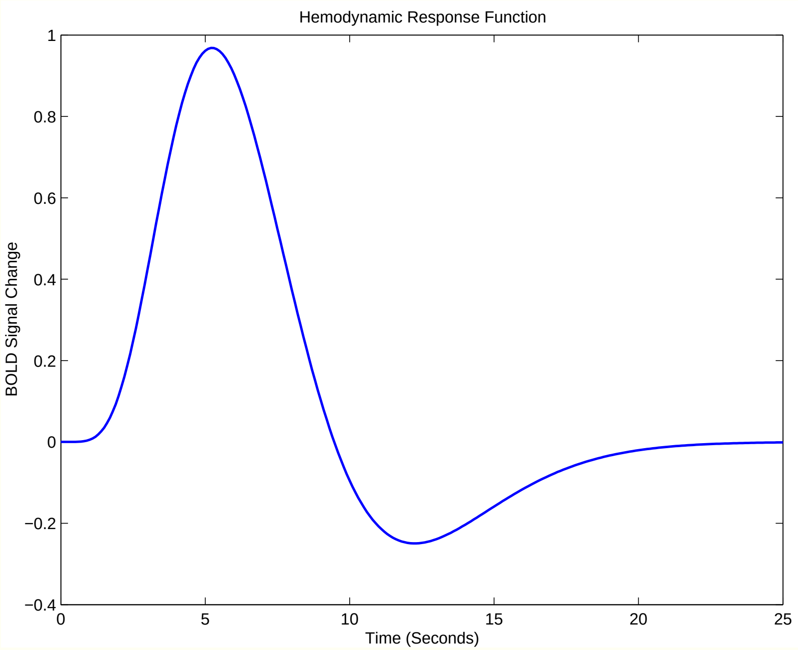
We will focus on parametric models where the goal is signal detection. This involves estimating a parameter from data and estimating the variability of the parameter estimate. First, we consider the signal.

Hemodynamic Response Function

We assume that the BOLD signal $x(t)$ is linked to a stimulus by a convolution of the stimulus $s(t)$ by a hemodynamic response function (HRF) $h(t)$:

$$x(t) = \int_0^\infty h(u)s(t-u)\,du.$$

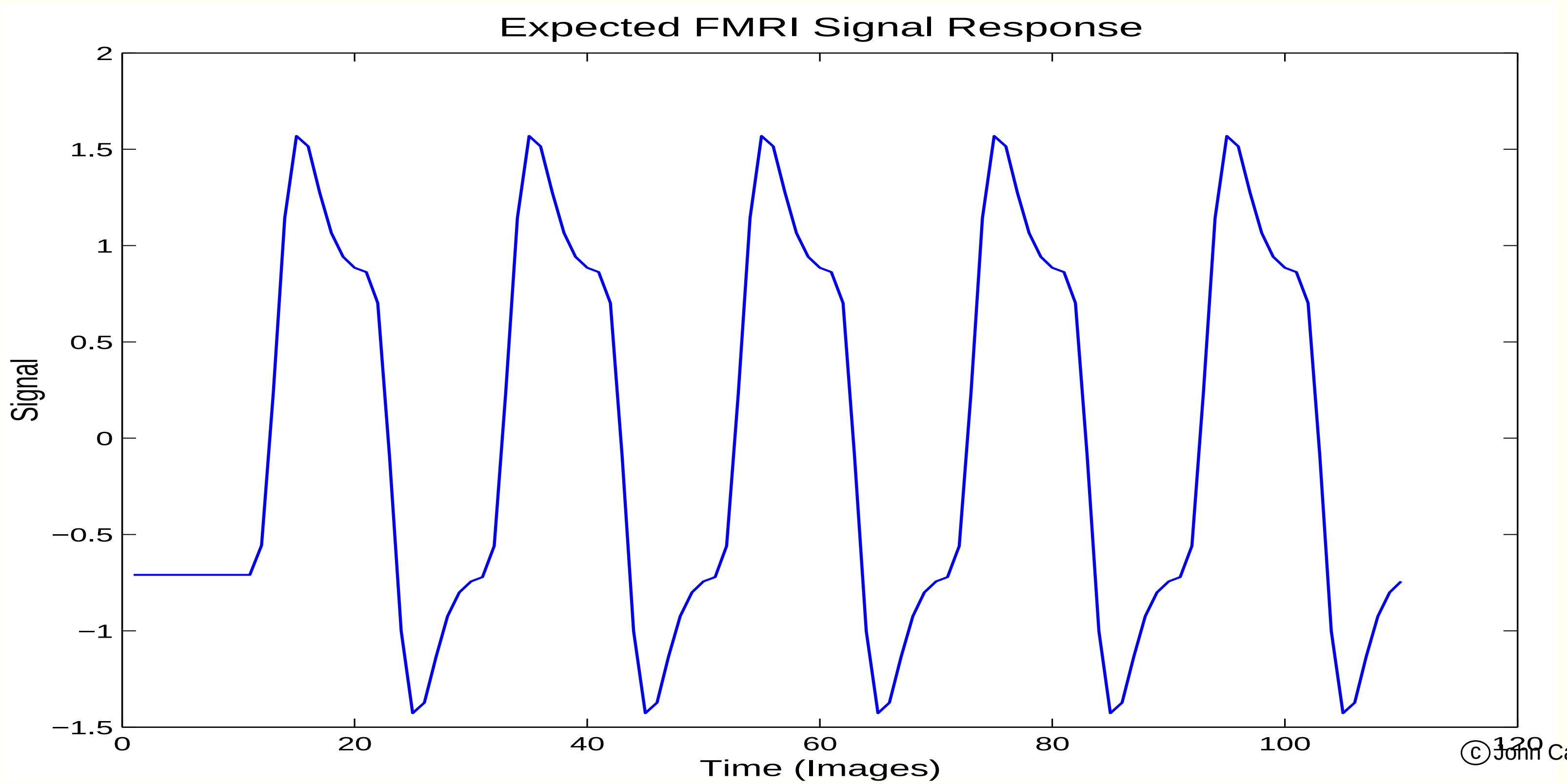
A common model for h is the difference of two gamma functions:



Expected BOLD Signal

To determine the expected BOLD signal from a stimulus $s(t)$, we must first represent the stimulus.

$$s(t_i) = \begin{cases} 1 & \text{if stimulus/task at } t_i \\ 0 & \text{otherwise} \end{cases}$$



Linear Regression Model

A common model for fMRI time series is a linear regression model (e.g., Worsley and Friston, 1995).

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{K}\epsilon \tag{1}$$

- $\mathbf{y} = (y_1, \ldots, y_n)^T$ is an $n \times 1$ matrix of samples of the fMRI signal
- $\mathbf{X}$ is an $n \times p$ design matrix with columns that contain signals of interest and nuisance signals
- β is a $p \times 1$ matrix of unknown parameters
- $\mathbf{K}$ is an unknown matrix
- $\epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$

Fitting the Linear Model

A least squares estimator of β is given by

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \tag{2}$$

Then, the variance of the estimator is

$$\mathrm{var}(\hat{\beta}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \,\mathrm{var}(\mathbf{y})\, \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}. \tag{3}$$

If $\mathrm{var}(\mathbf{y}) = \sigma^2\mathbf{I}$, then (3) simplifies. If it takes some general form (positive definite and symmetric), then the expression remains rather complicated.

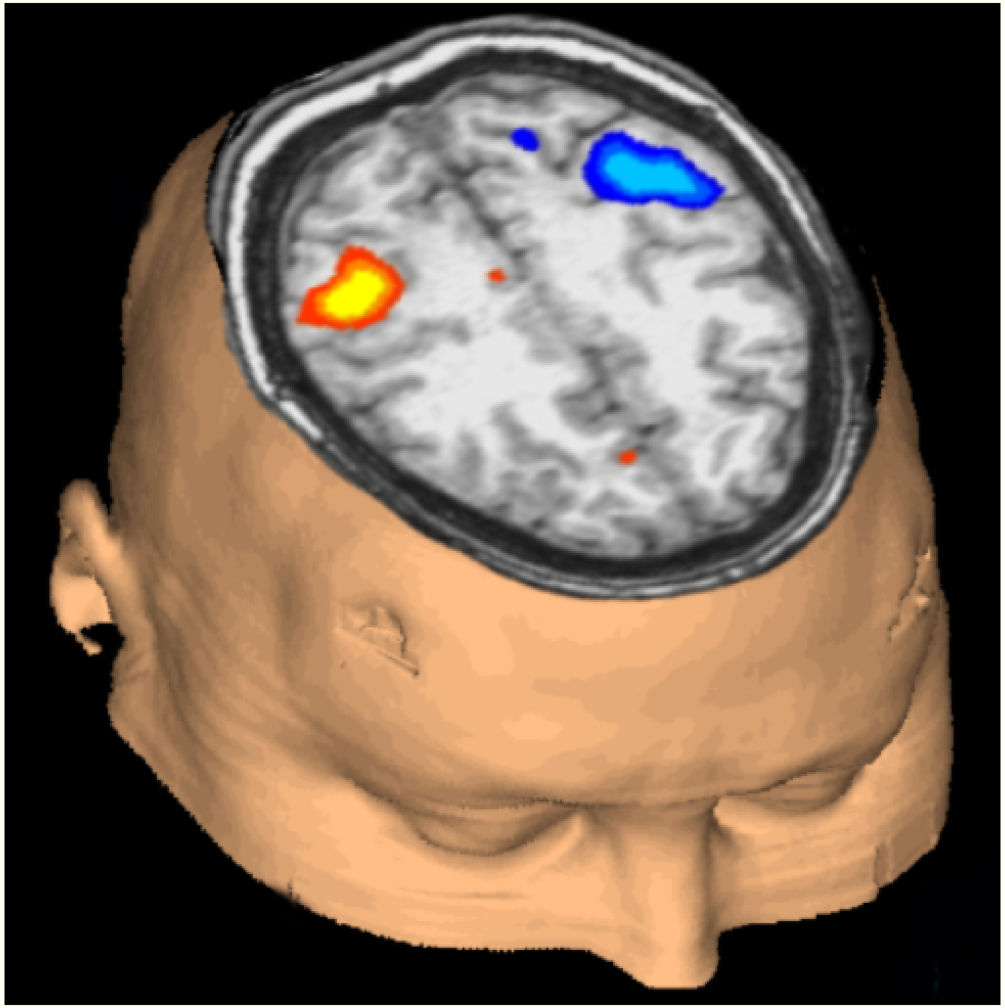
Problems with Regression Model

Suppose we fit the linear model to an fMRI time series and find autocorrelation in the residuals. What happens?

- $\hat{\beta}$ is OK since it only needs $E[\epsilon] = 0$ to be unbiased
- The variance seems suspect since we use the residuals to estimate the variance,
  - Assuming $\epsilon \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$ will lead to an incorrect estimate of $\mathrm{var}(\hat{\beta})$.
  - Alternatively we can assume (or attempt to estimate from data) a variance structure for ϵ. However if our estimate or assumption is incorrect, we will still get a biased variance estimate.
- Biased variance estimates impact the inferential step in our analysis–the, “Statistical Parametric Mapping.”

Statistical Maps

A common statistical map for signal detection is a voxel-by-voxel map of the test statistic under $H_0 : \beta = 0$, i.e., no signal. As an example: $T > 5$, finger tapping (Blue $=$ right, Yellow $=$ left).



Back to Regression Problems

How can an incorrect variance estimate affect, for instance, the SPM that was on the previous slide?

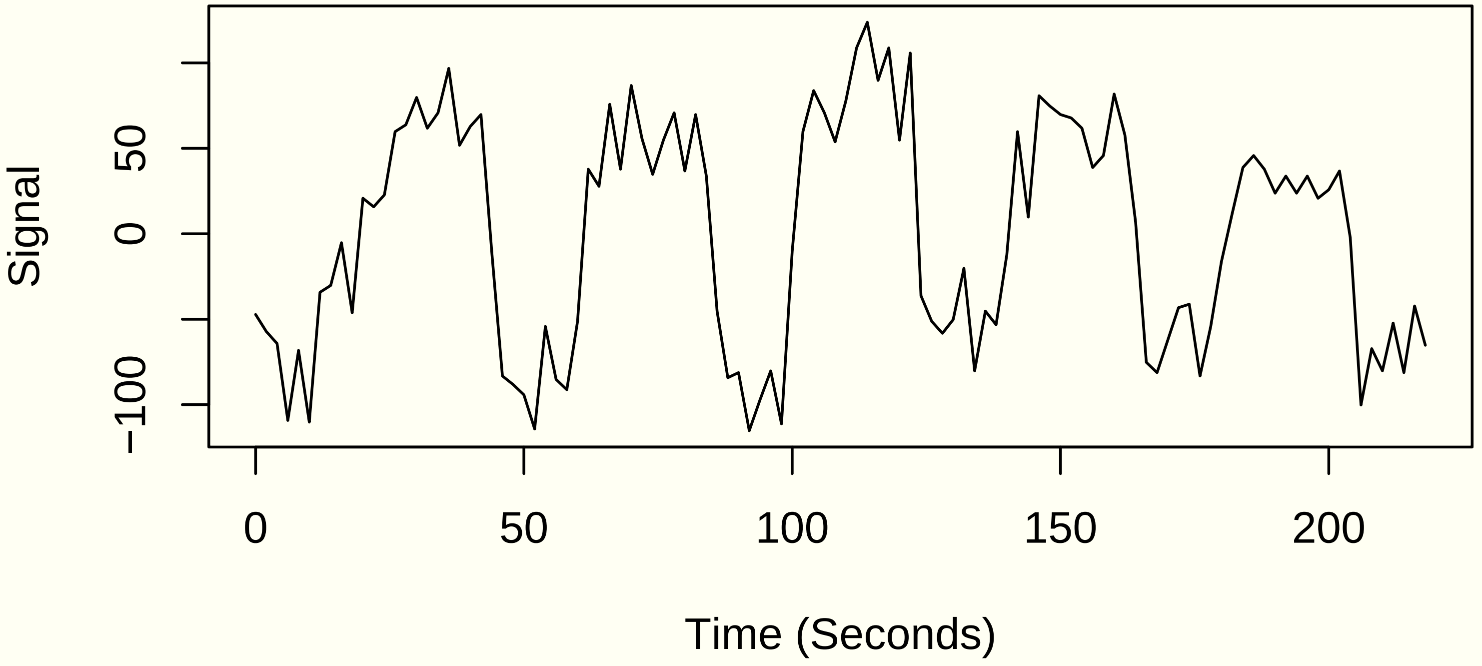
- At each voxel, a test statistic, T is computed by

$$T = \frac{\hat{\beta}}{\sqrt{\mathrm{var}(\hat{\beta})}}. \tag{4}$$

- Thus, if the estimated variance is too low, T will be too large. The significance will be overstated, possibly leading to a false positive.
- Conversely, if the estimated variance is too high, T will be too small. The significance will be understated. This will lead to a loss of power to reject the null hypothesis when it is false.

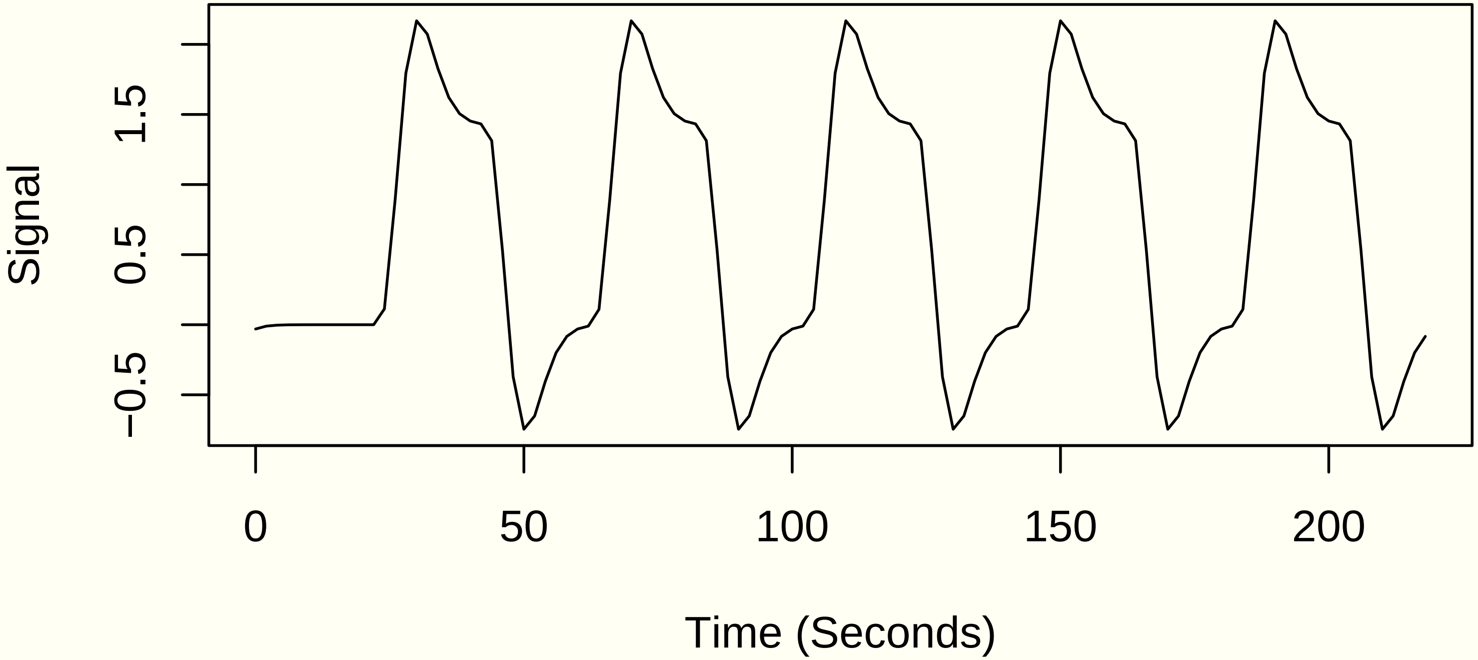
A Concrete Example

Let's fit a linear model to an fMRI time series that presumably contains a signal related to a task performed during the scan session.



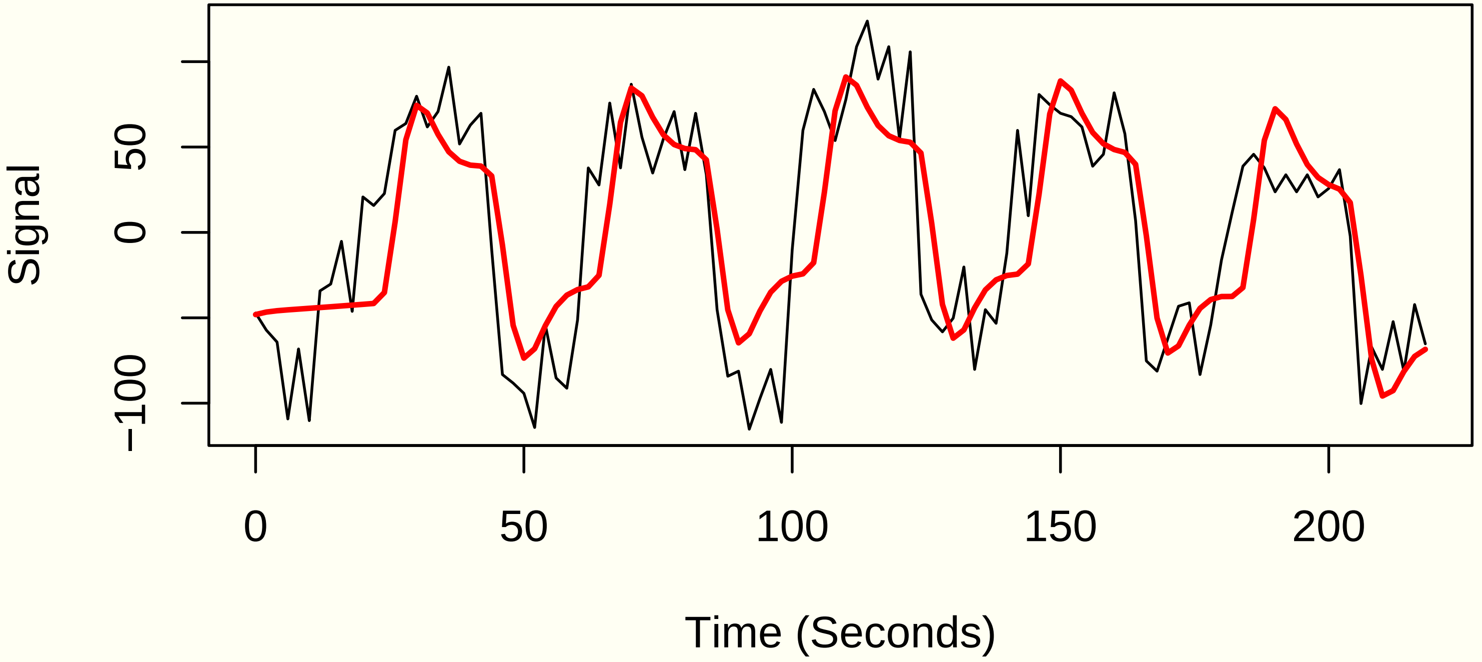
Example: Expected Response

This is the expected response–or the signal we wish to detect in our time series.



Example: Model Fit

A fit of the signal (plus polynomial drift) to the time series shows, visually, a strong signal component. The fit does show some problems (e.g., wrong hemodynamic delay).

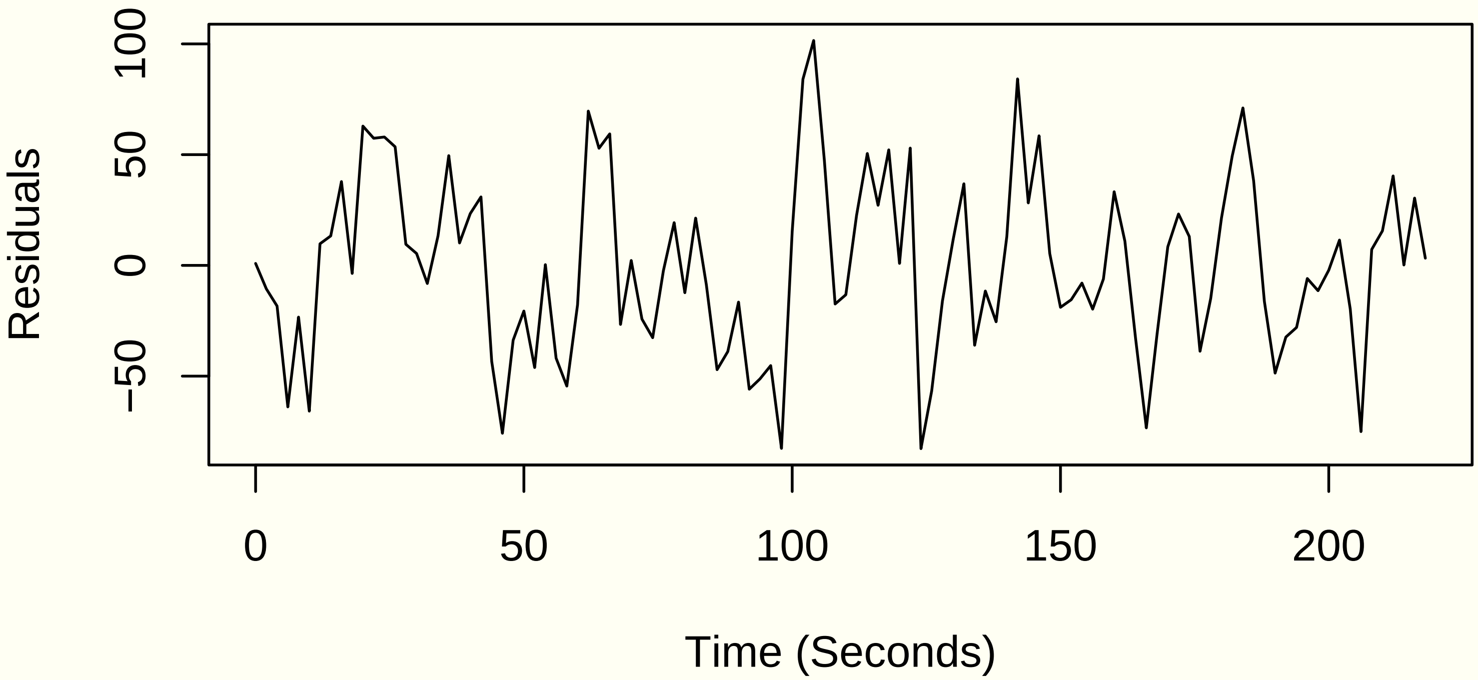


Example: Summary of Fit

	Estimate	Std.Error	t-value	p-value
(Intercept)	-34.006	4.658	-7.300	5.76e-11
signal	52.588	4.114	12.784	< 2e-16
poly(tind, 3)1	-12.391	40.135	-0.309	0.7581
poly(tind, 3)2	-102.628	40.638	-2.525	0.0131
poly(tind, 3)3	-28.310	40.121	-0.706	0.4820

Example: Residuals

A plot of the residuals (observed values minus fitted values) shows systematic components. This suggests correlated errors.



Example: Detecting Correlated Errors

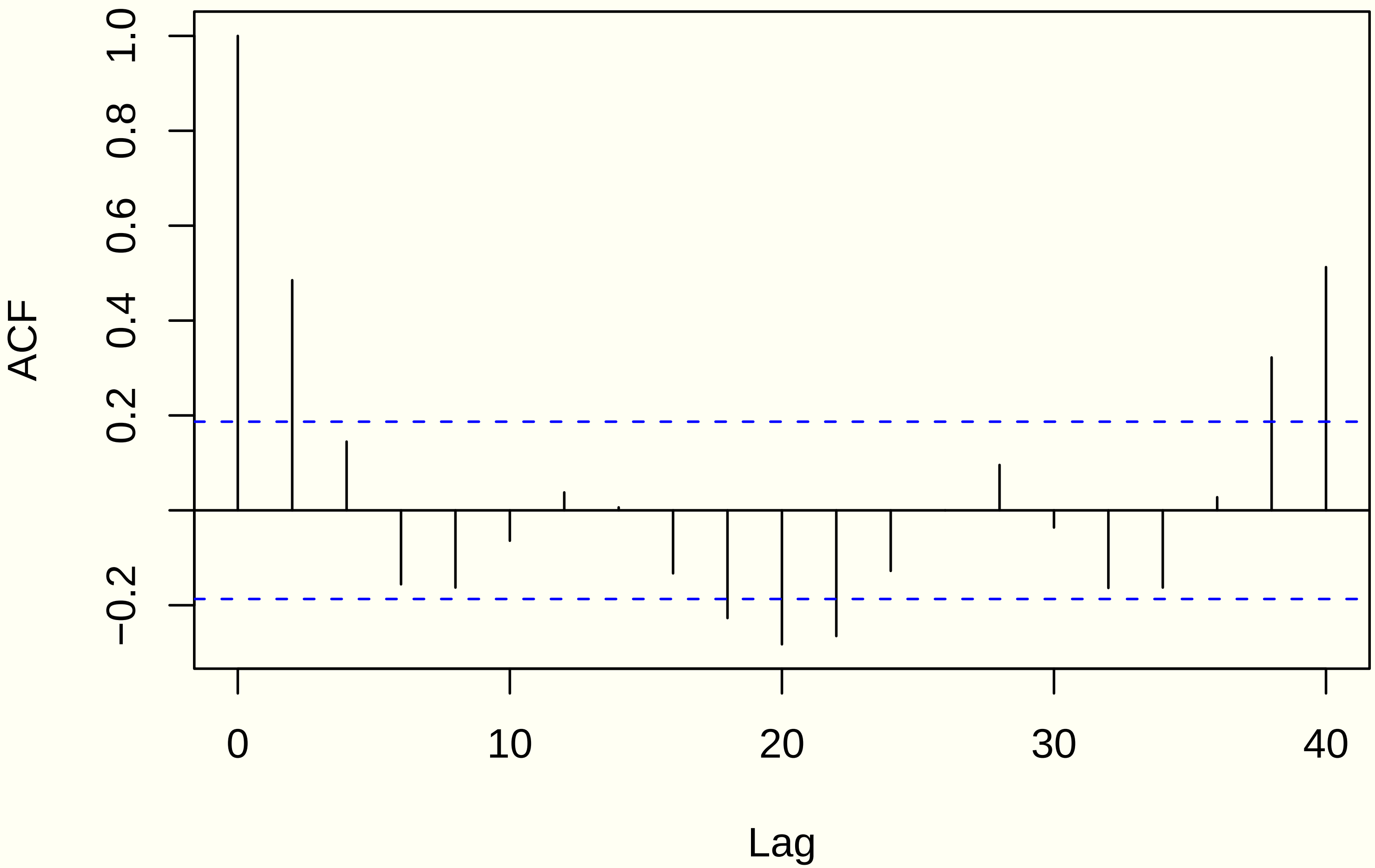
How can we detect correlated errors? One tool is the sample autocovariance function (for lag k):

$$c_k = \frac{1}{N} \sum_{t=1}^{N-k} (x_t - \bar{x})(x_{t+k} - \bar{x})$$

You may recognize the similarity between the sample autocovariance and the sample covariance for realizations of two random variables. The sample autocovariance is the covariance between observations of a time series at a distance k.

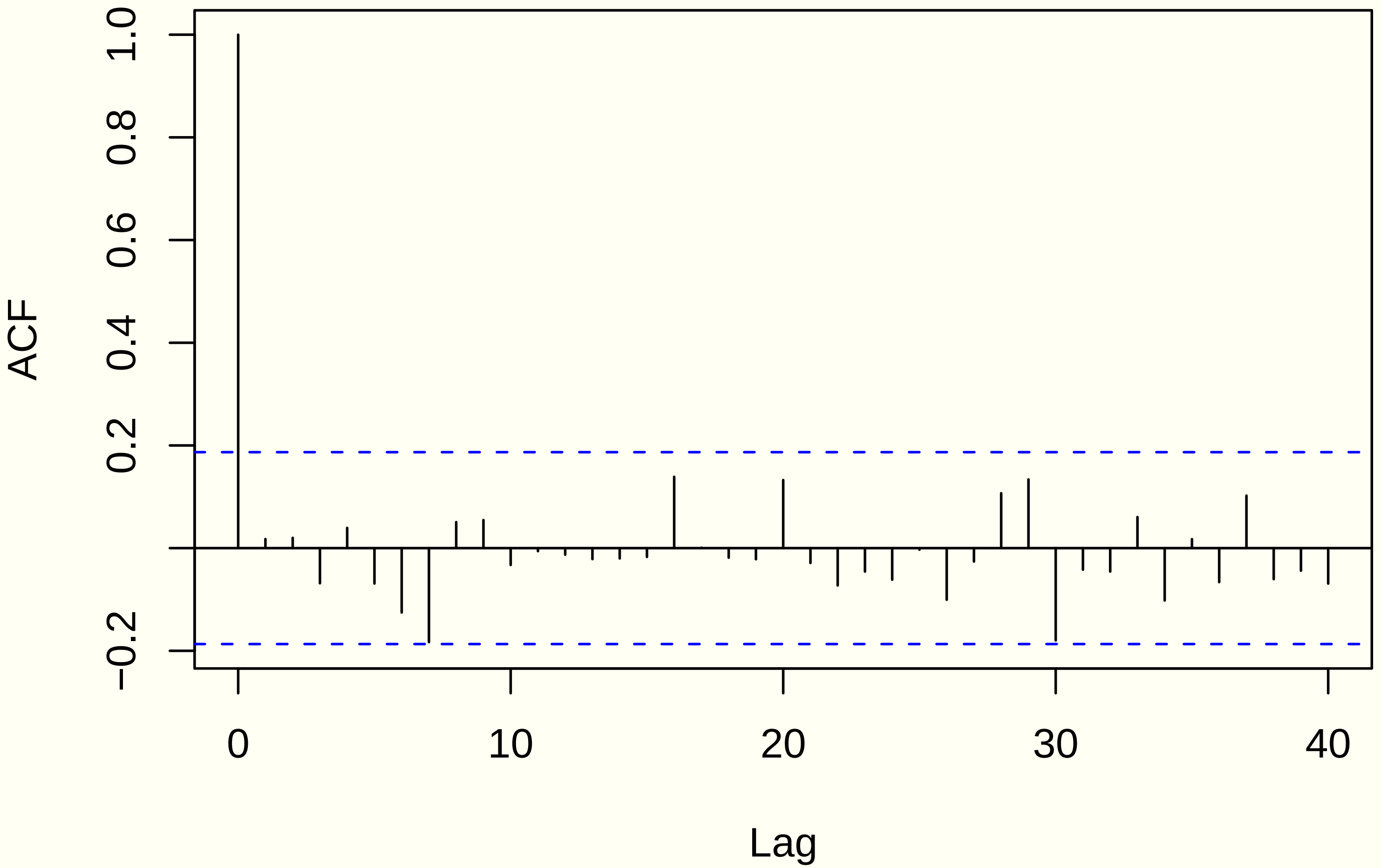
Example: Sample ACF

This plot shows significant correlations in the residuals. What is the source of these correlations?



Example: White Noise

Compare to 110 samples of a Gaussian random variable:



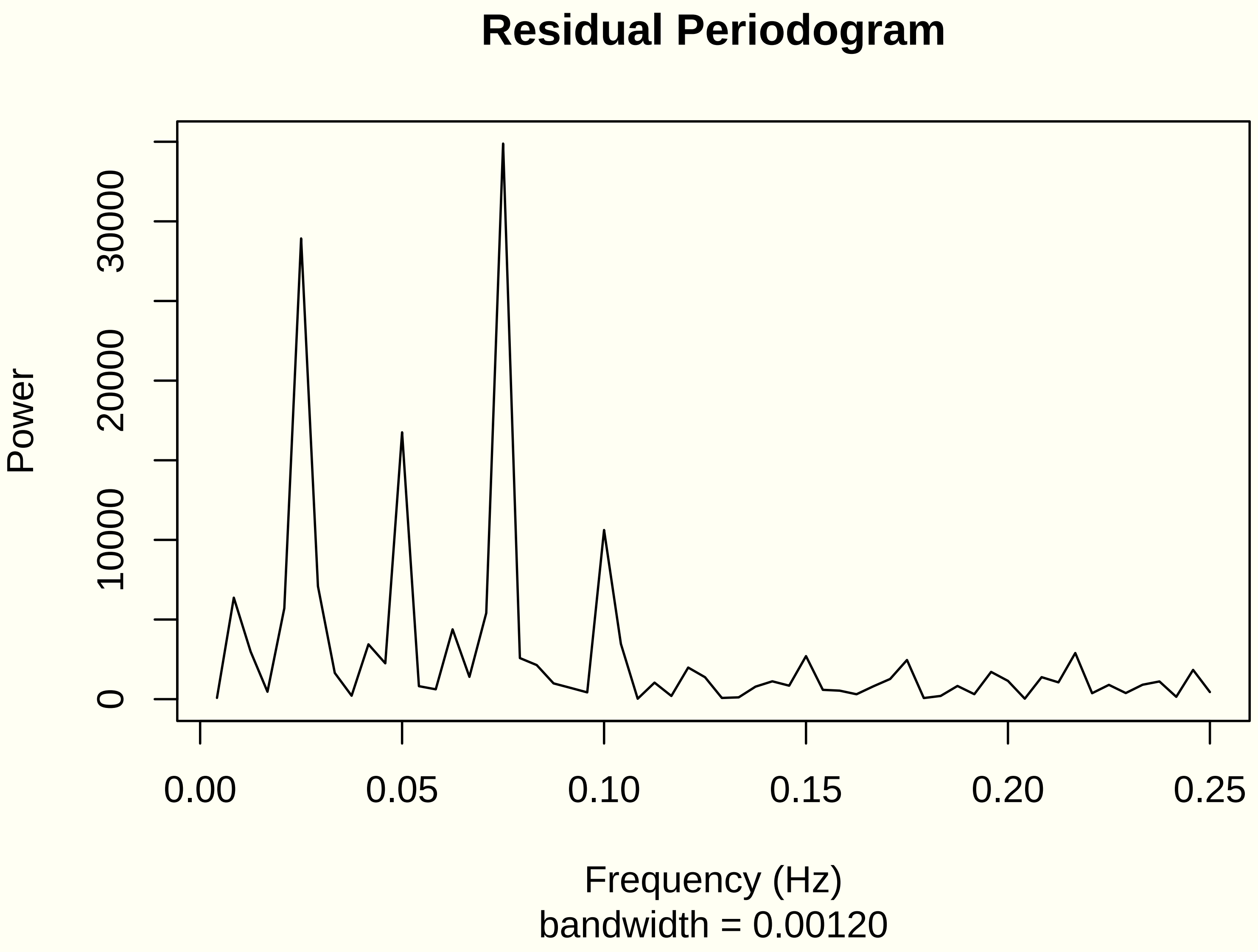
Example: Periodogram

The periodogram (sometimes called power spectrum) is another tool for investigating the correlation structure of a time series. It conveys the same information as the ACF, but in a slightly different way.

$$I(\omega) = \frac{1}{N}\left[\left\{\sum_{t=1}^{N} x_t \sin(\omega t)\right\}^2 + \left\{\sum_{t=1}^{N} x_t \cos(\omega t)\right\}^2\right]$$

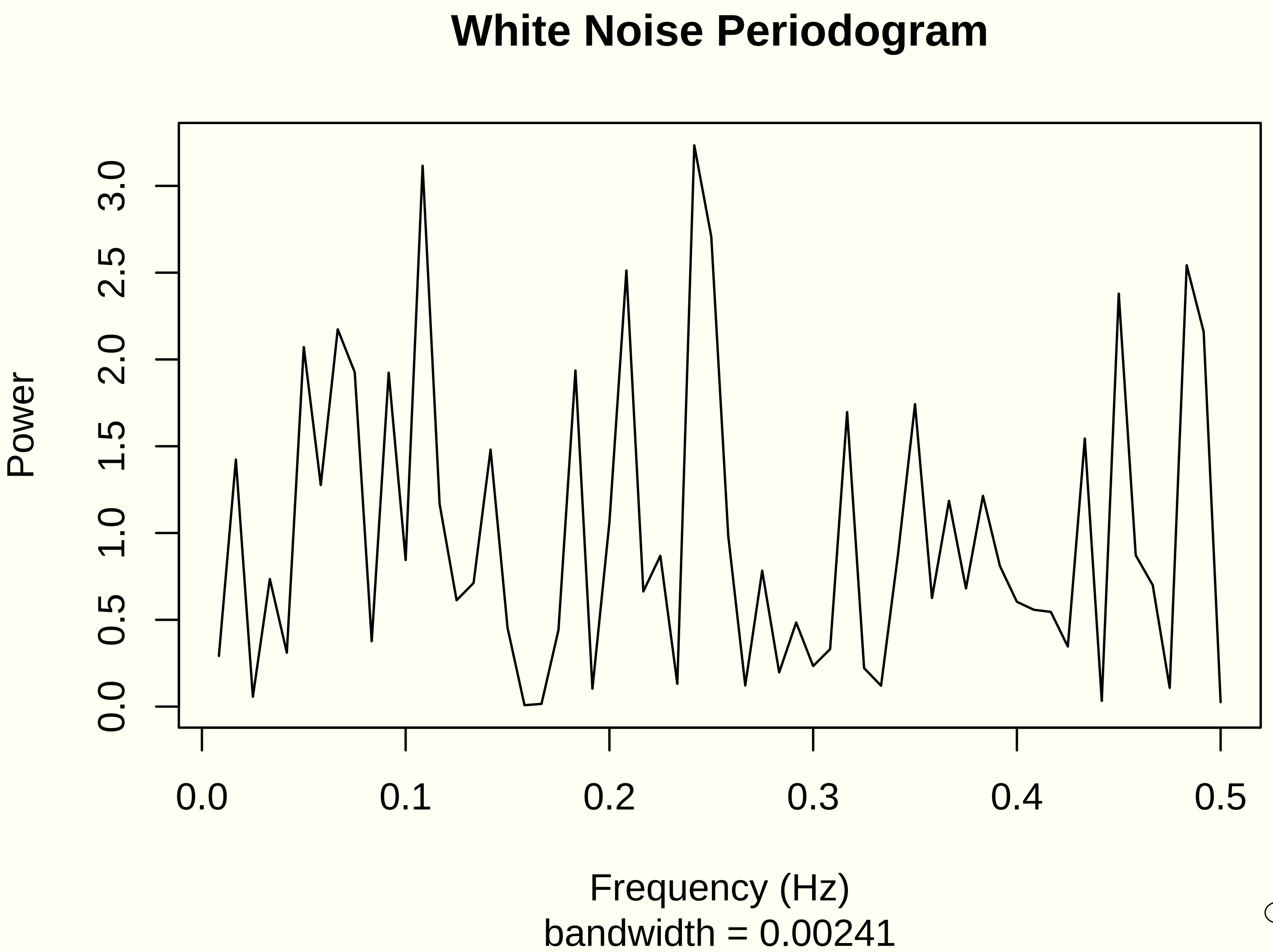
Note: a commonly unknown fact is that the periodogram is not a consistent estimator of the spectral density.

Example: Residual Periodogram



Example: Noise Periodogram

Compare to Gaussian noise:



Example: Re-fit with AR(1)

The regression model was re-fit assuming and AR(1) noise structure.

	Value	Std.Error	t-value	p-value
signal	41.427	6.971	5.943	<.0001

Compare this to the standard OLS assuming independent errors.

signal	52.588	4.114	12.784	<2e-16

Both the estimate and the standard error change. The main message is that to falsely assume independent errors can dramatically overstate the significance of the signal.

Addressing the Variance Problem

Two general approaches are commonly used in the fMRI community:

- Temporal smoothing (Worsley and Friston, 1995): Attempt to "condition" the correlation structure to some known form.
- Pre-whitening (Bullmore et al., 1996): Fit the model under iid error assumptions, model the correlation structure of the residuals, whiten the raw data by removing the estimated residual structure, and finally refit the model. If a good estimate of the residual autocorrelations can be computed, pre-whitening is BLUE.
- It has been argued (e.g., Friston et al., 2000) that smoothing is preferred over prewhitening.

Smoothing Approach

Let $\mathbf{S}$ be a linear transformation. The matrix $\mathbf{S}$ is applied to the linear model to give

$$\mathbf{S}\mathbf{y} = \mathbf{S}\mathbf{X}\beta + \mathbf{S}\mathbf{K}\epsilon. \tag{5}$$

(Note: If $\mathbf{K}$ is known, the approach of whitening is to choose $\mathbf{S} = \mathbf{K}^{-1}$.) Now, a least squares estimator is

$$\hat{\beta} = (\mathbf{S}\mathbf{X})^{+}\mathbf{S}\mathbf{y}. \tag{6}$$

Let $\mathbf{V} = \mathbf{K}\mathbf{K}^T$ be the unknown variance (prior to smoothing). With smoothing, the assumed variance $\mathbf{V}_a = \mathbf{S}\mathbf{S}^T \approx \mathbf{S}\mathbf{V}\mathbf{S}^T$, the true variance. We need to pick $\mathbf{S}$ in a way that will reduce the bias in the variance estimates.

How to Select $\mathbf{S}$?

- The bias as a function of $\mathbf{S}$ is difficult to directly minimize.
- With this limitation, we must determine empirically if a given smoother gives an acceptable level of bias.
- An approach used in the fMRI community, e.g., SPM99, is to use a kernel smoother with a fixed bandwidth comparable to the HRF.
- An alternative approach is to use an objective criterion to pick the “best” smoother over a range of sensible parameters.

The Smoothing Spline Approach

- We fit a spline to each voxel time series and use the spline smoothing matrix as $\mathbf{S}$ for fitting a smoothed linear model to the voxel.
- Spline model for y_i:

$$y_i = f(t_i) + \epsilon_i, \tag{7}$$

  where f is a smooth function, $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, and t_i for $i = 1, \ldots, n$ are equally-spaced times when fMR images are acquired.
- References on the method: Green and Silverman (1994) is a “gentle” introduction. The definitive reference is Wahba (1990) *Spline Models for Observational Data* SIAM, Philadelphia.

Fitting a Smoothing Spline

An estimator of $f(t_i)$ is obtained from

$$f(\hat{t}_i) = \arg \min_{f \in C^2[t_1,t_n]} \left(\sum_{i=1}^{n} (y_i - f(t_i))^2 + \lambda \int_{t_1}^{t_n} (f''(x))^2 \, dx \right). \qquad (8)$$

The unique solution is a natural cubic spline–a piecewise cubic polynomial with “knots” at each sample point. When $\lambda = \infty$, $\hat{f}$ is a linear approximation. When $\lambda = 0$, i.e., no smoothing, $\hat{f}$ interpolates the y_i with a piecewise cubic polynomial. The estimator is linear. See Carew et al. (2003) for details on how to fit.

Model Selection: Choosing λ by GCV

- One method for selecting the optimal smoothing parameter is generalized cross-validation (GCV) (Craven and Wahba, 1979).
- Given λ, the GCV score is

$$V(\lambda) = \frac{1}{n} \cdot \frac{\sum_{i=1}^{n} (y_i - f(\hat{t}_i))^2}{(1 - n^{-1}\mathrm{tr}\mathbf{A}(\lambda))^2}. \quad (9)$$

  Matrix $\mathbf{A}(\lambda) = \Gamma(\mathbf{I} + \lambda\mathbf{D})^{-1}\Gamma^T$ maps the data to their fitted values.
- The GCV score is asymptotically a predictive mean square error criterion. This means that for large n, the λ that minimizes the GCV score will give a spline estimate that minimizes the mean square error between the estimate and the true, unknown function.

Bias of Variance Estimator

Given $\mathbf{S}$ computed with GCV-spline, the variance of a contrast of $\hat{\beta}$ and the bias of the variance estimator can be computed with equations given in Friston et al. (2000):

$$\mathrm{var}(\mathbf{c}^T\hat{\beta}) = \sigma^2\mathbf{c}^T(\mathbf{SX})^+\mathbf{SVS}^T(\mathbf{SX})^{+T}\mathbf{c} \tag{10}$$

and

$$\begin{aligned}\mathrm{bias}(\mathbf{S},\mathbf{V}) &= \frac{\mathrm{var}(\mathbf{c}^T\hat{\beta}) - \mathrm{E}[\widehat{\mathrm{var}(\mathbf{c}^T\hat{\beta})}]}{\mathrm{var}(\mathbf{c}^T\hat{\beta})} \\ &= 1 - \frac{\mathrm{tr}[\mathbf{LSVS}^T]\mathbf{c}^T(\mathbf{SX})^+\mathbf{SV}_a\mathbf{S}^T(\mathbf{SX})^{+T}\mathbf{c}}{\mathrm{tr}[\mathbf{LSV}_a\mathbf{S}^T]\mathbf{c}^T(\mathbf{SX})^+\mathbf{SVS}^T(\mathbf{SX})^{+T}\mathbf{c}},\end{aligned} \tag{11}$$

where $\mathbf{L} = \mathbf{I} - \mathbf{SX}(\mathbf{SX})^+$ is the residual forming matrix and $\mathbf{c}$ is a contrast vector for hypothesis testing of the components of $\hat{\beta}$.

Bias, Continued

An estimate of $\mathrm{var}(\mathbf{c}^T\hat{\beta})$ is obtained by replacing $\mathbf{V}$ with its assumed value, $\mathbf{V}_a$, and σ^2 with its estimate

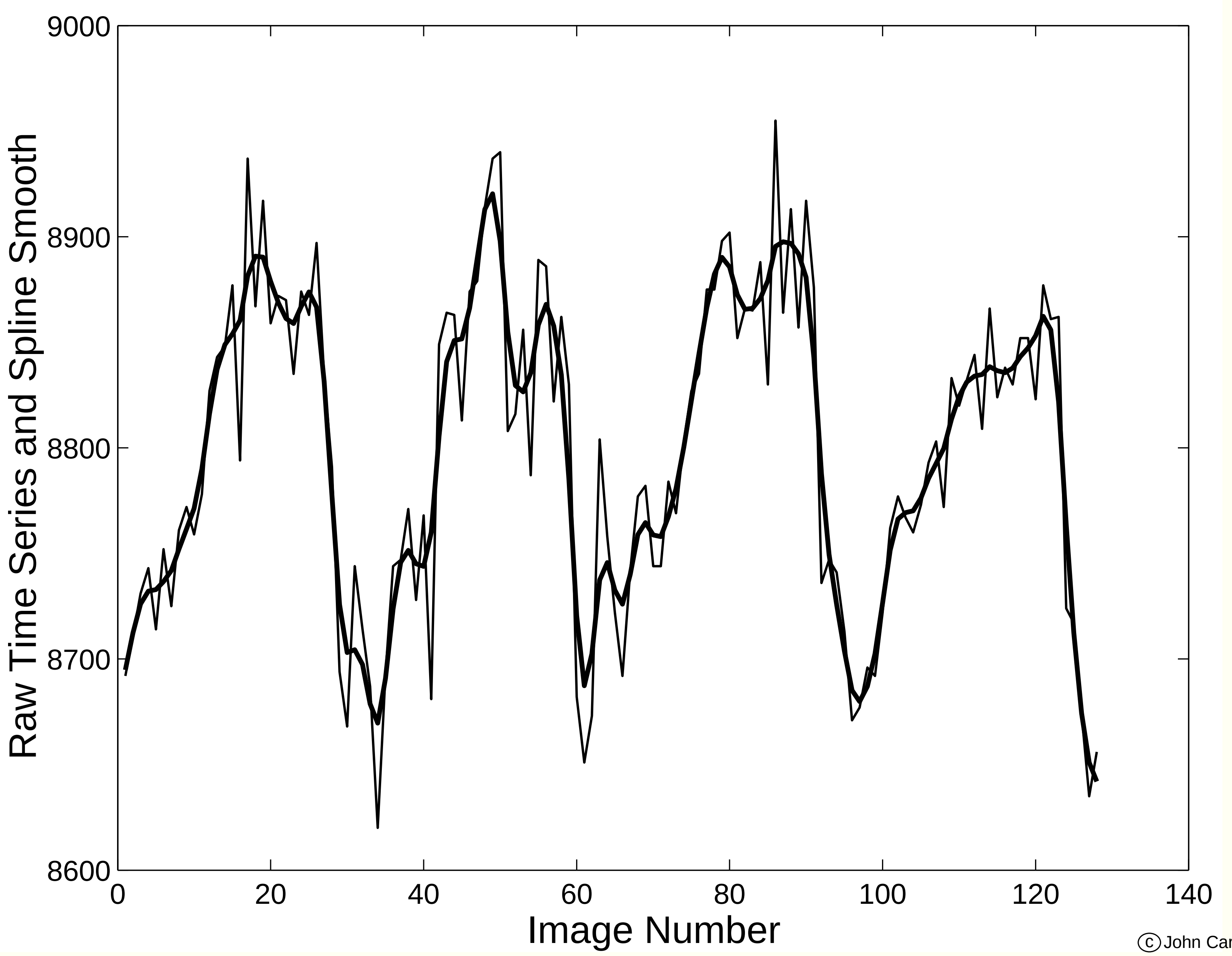
$$\hat{\sigma}^2 = \frac{(\mathbf{LSy})^T\mathbf{LSy}}{\mathrm{tr}(\mathbf{LV}_a)}, \tag{12}$$

given in Worsley and Friston (1995).

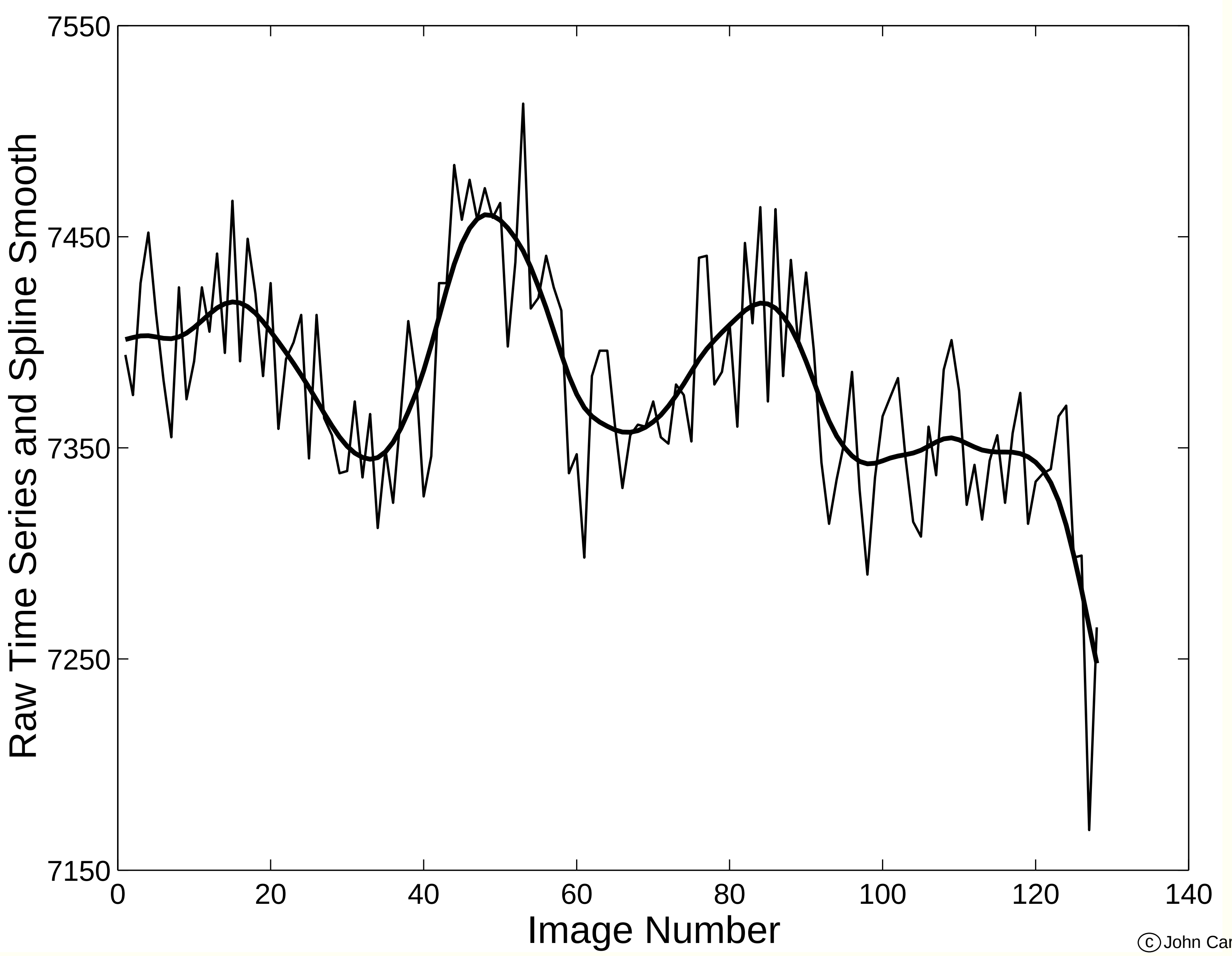
FMRI Experiment

- 1.5T GE-EPI, TR 2s, 64$\times$64 pixels, 22 7mm coronal slices.
- Four symmetric blocks of photic stimulation (30s on, 30s off).
- Time series were analyzed with the smoothed linear model where $\mathbf{S} \in \{\mathbf{I}, \mathrm{SPM-HRF}, \mathrm{GCV-Spline}\}$.

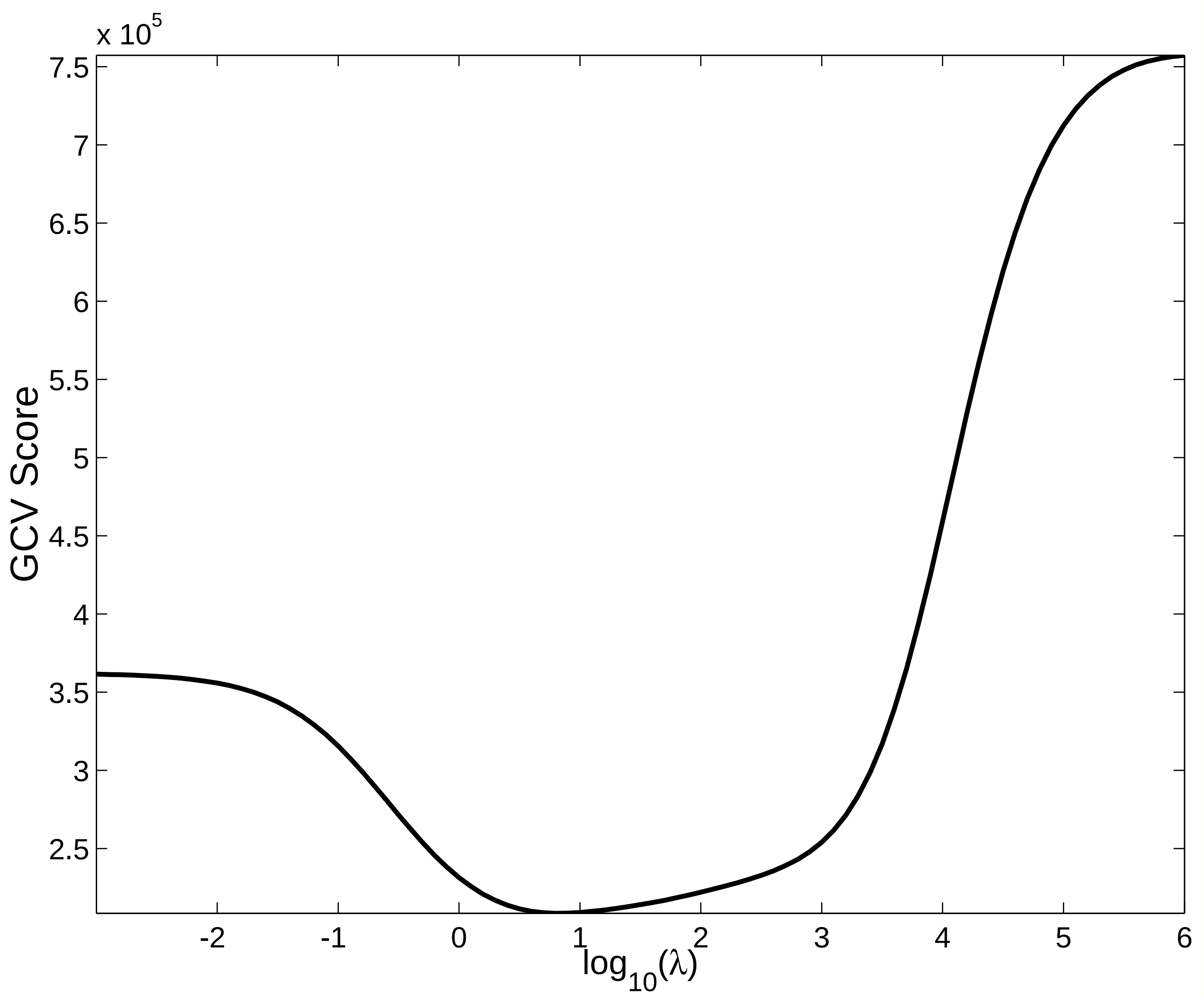
Results from Real Data



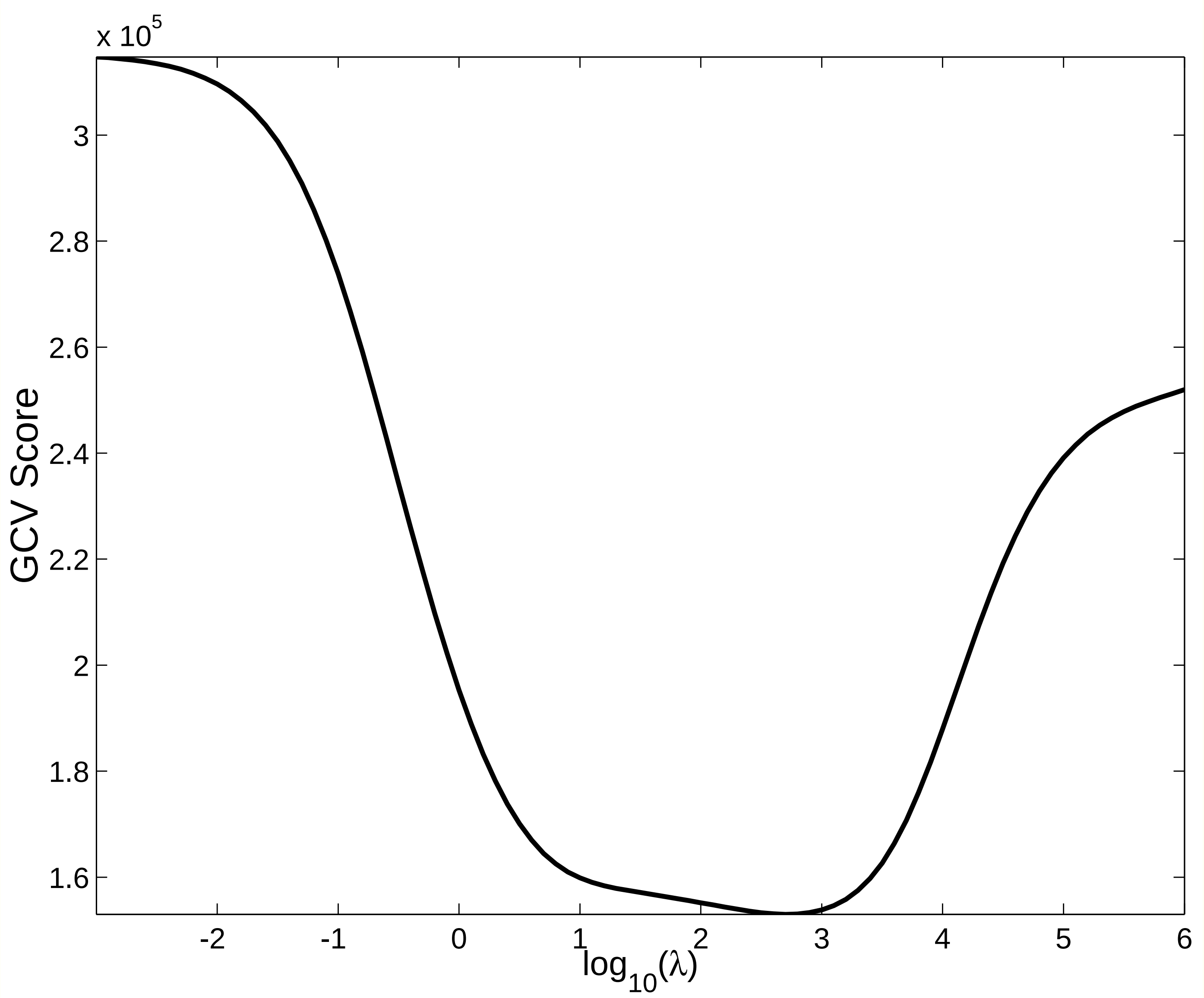
Results from Real Data



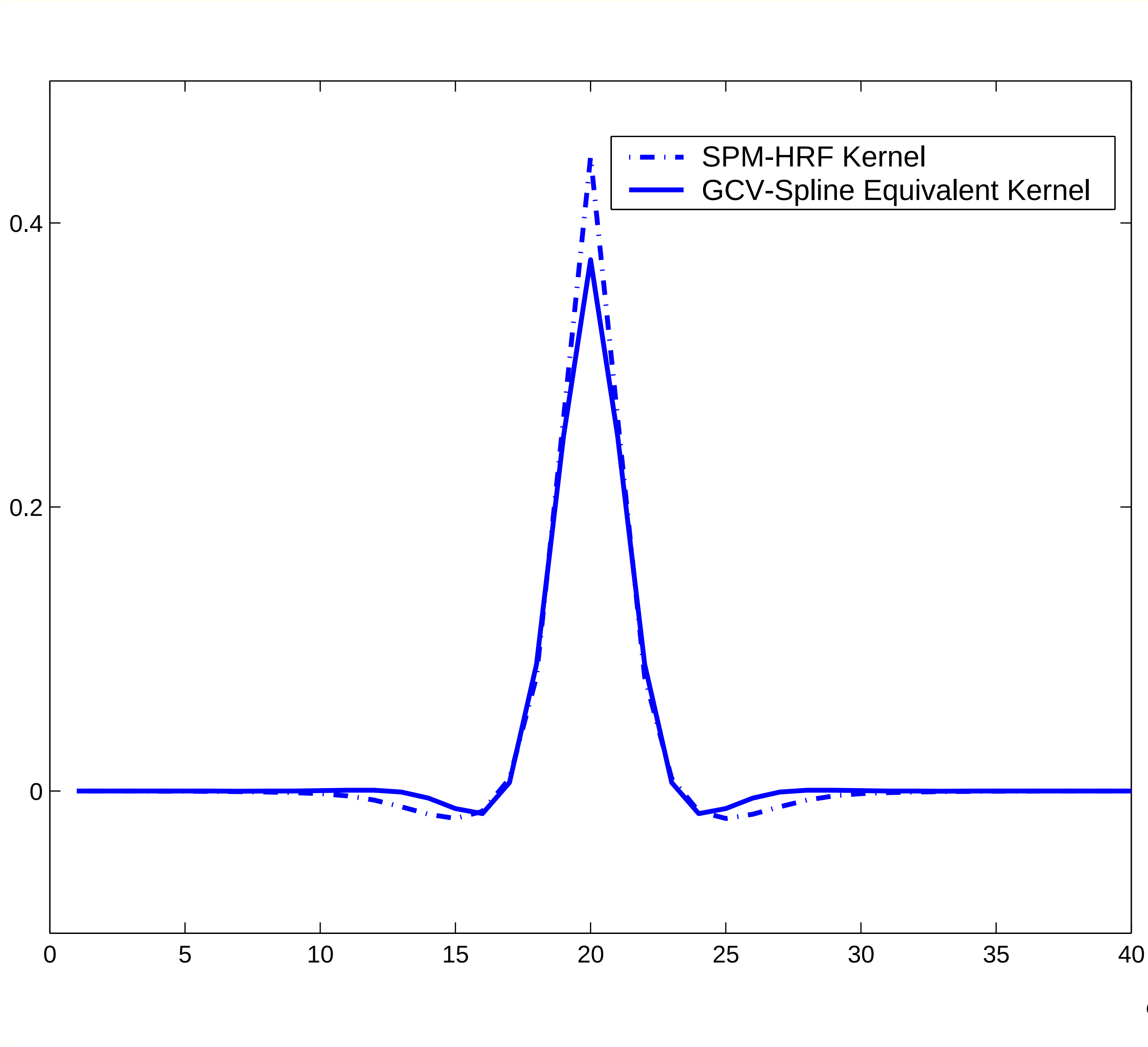
GCV Score $V(\lambda)$



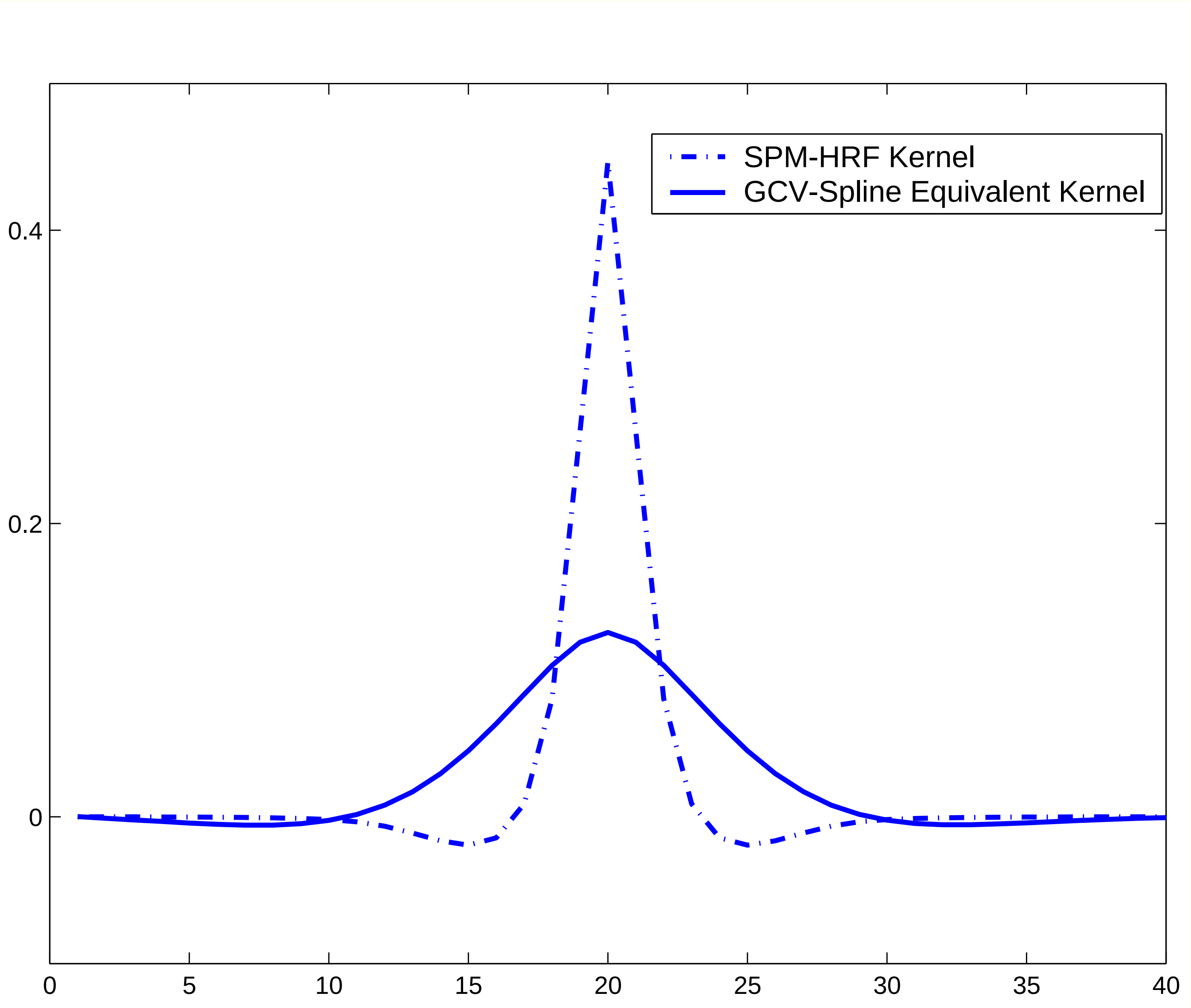
GCV Score $V(\lambda)$



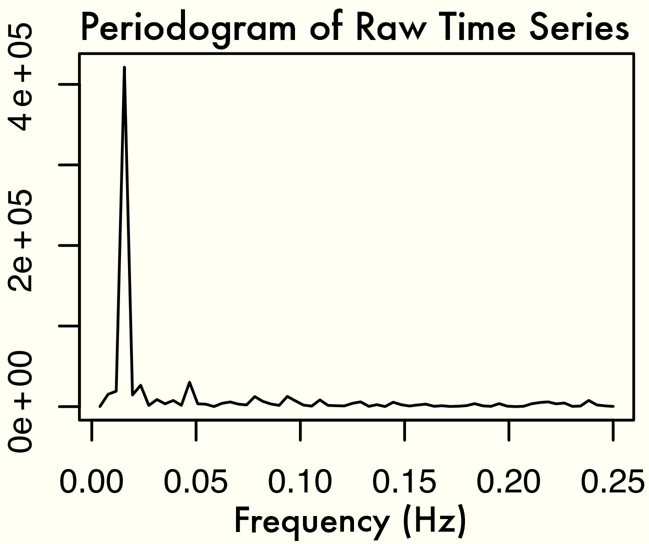
Equivalent Kernels/Basis Functions



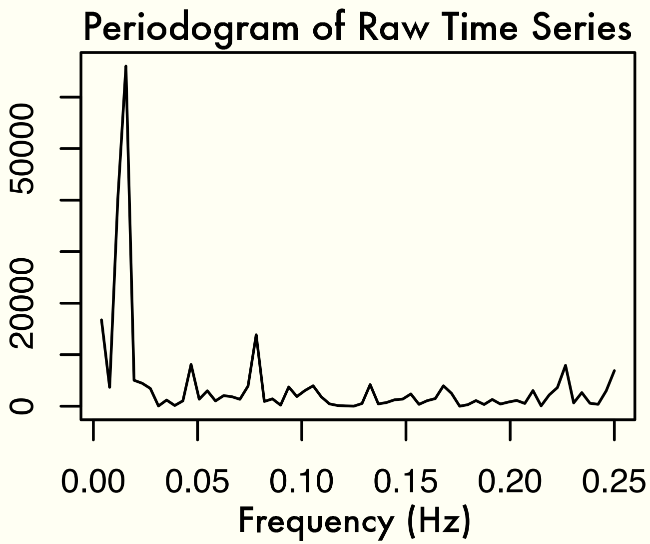
Equivalent Kernels/Basis Functions



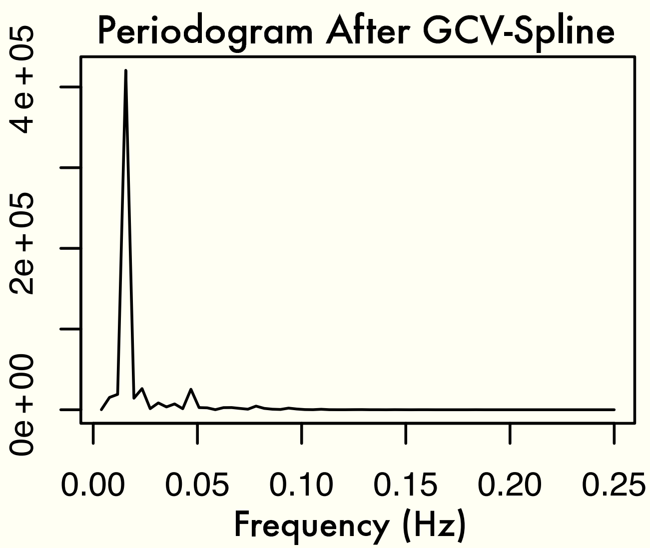
Results from Real Data



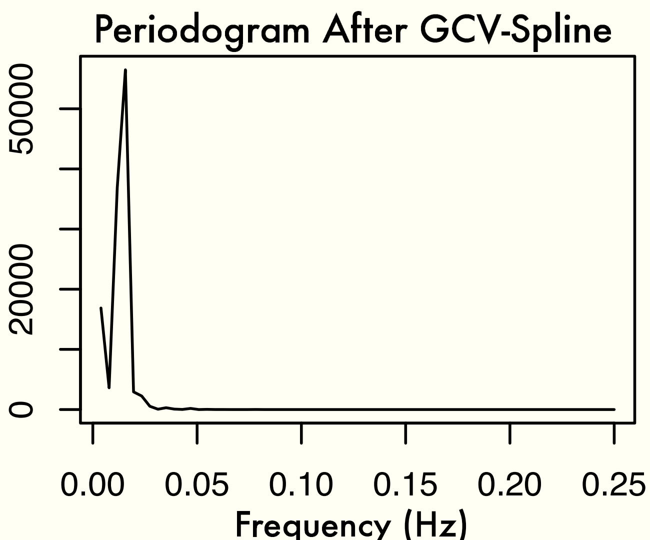
(g)



(h)

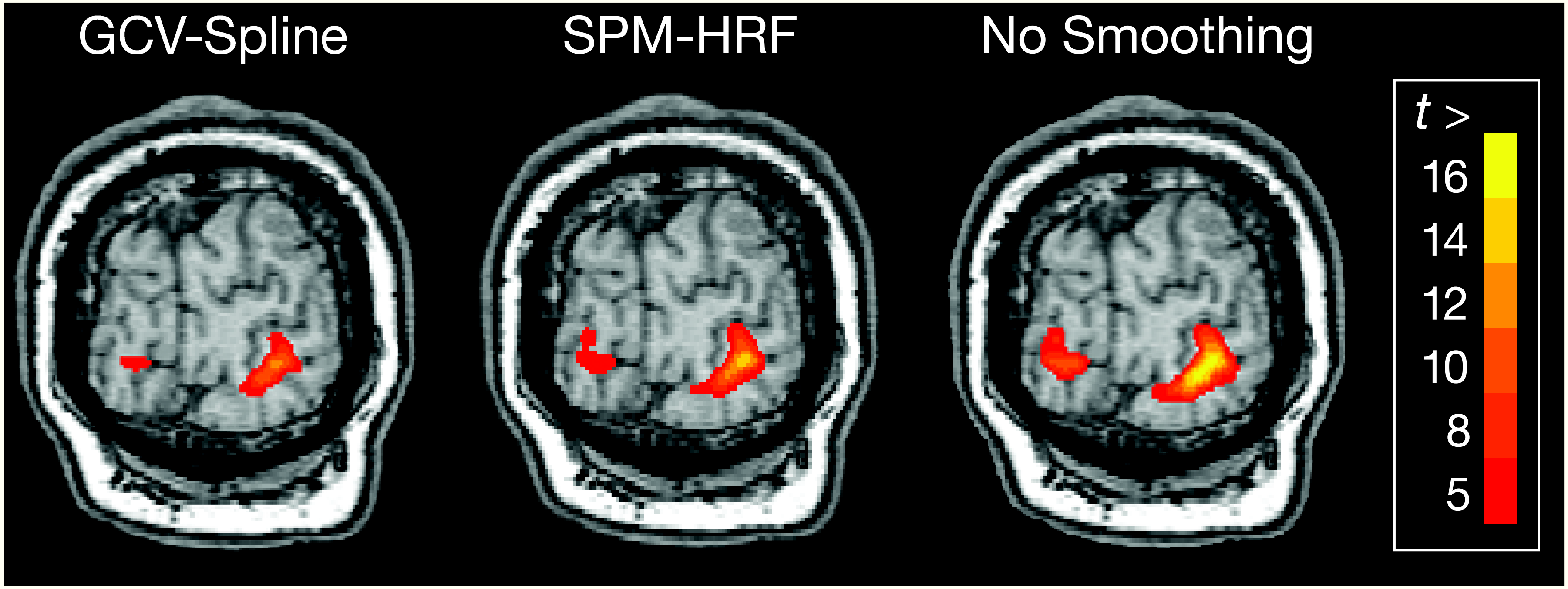


(i)



(j)

Results from Real Data

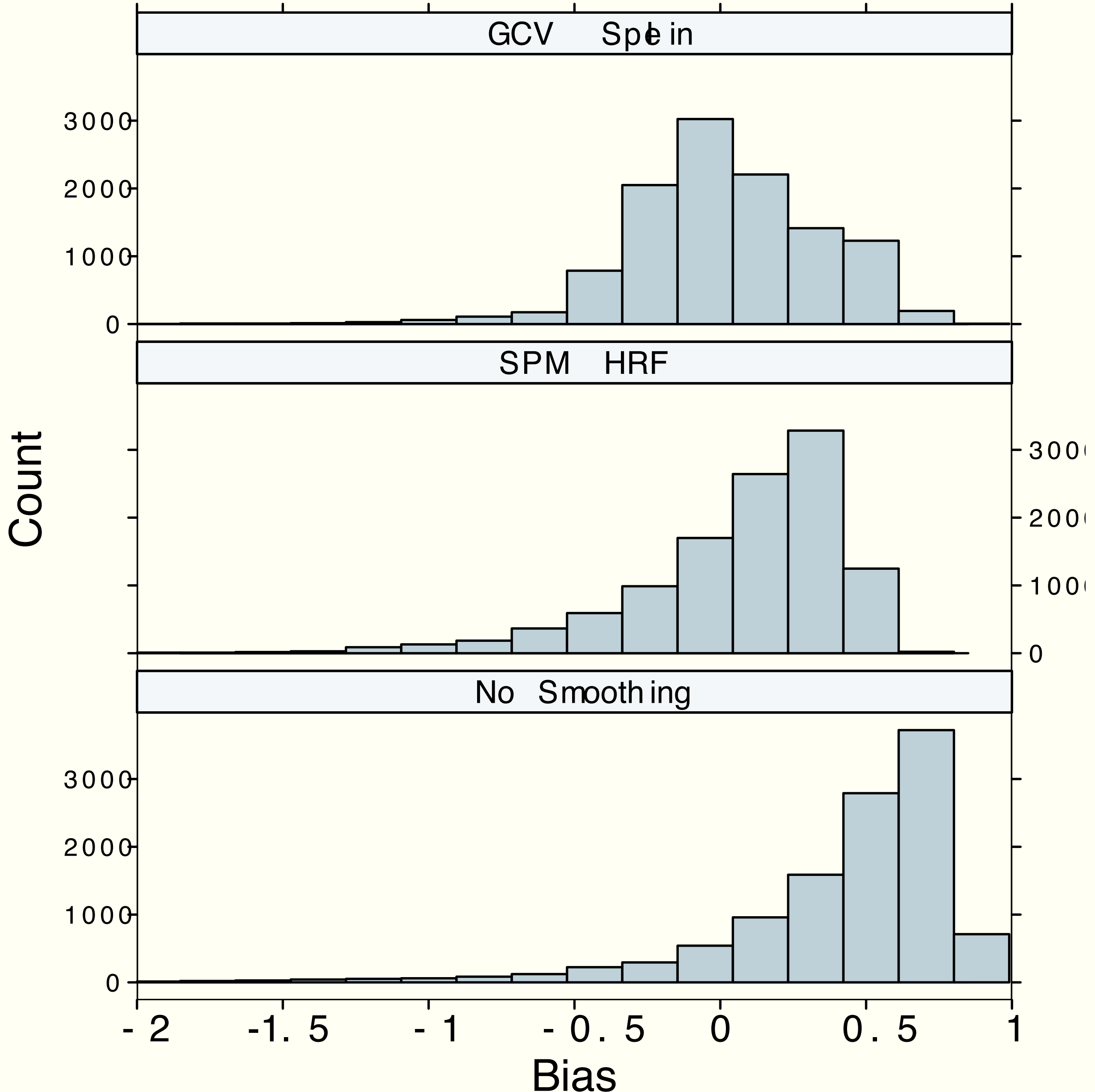


Simulation Study

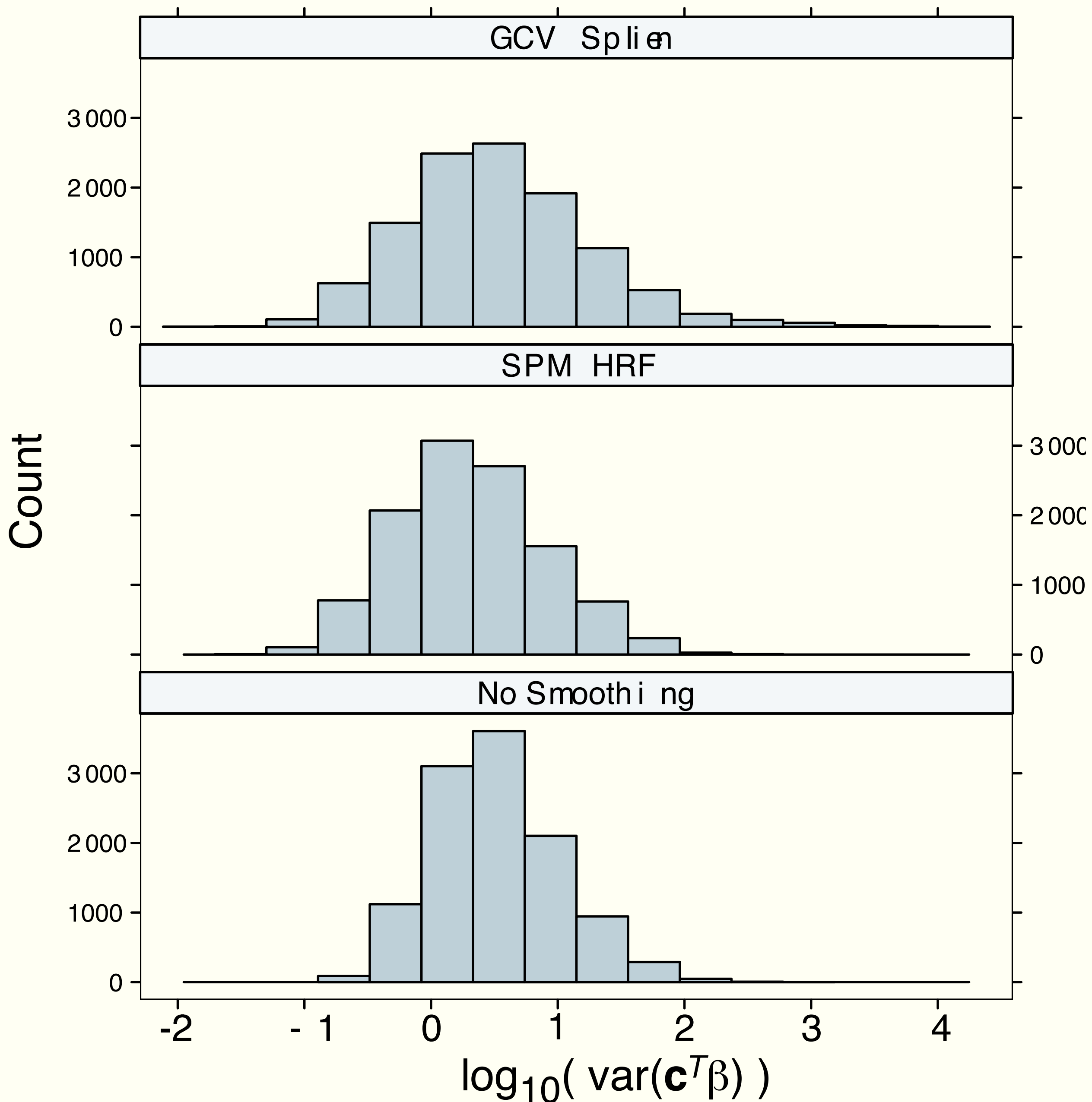
The goal is to create a data set with realistic, known error variance structures to evaluate different smoothing strategies.

- AR(8) model was fit to residuals of each voxel time series when $\mathbf{S} = \mathbf{I}$ to estimate $\mathbf{K}$ for each voxel.
- The estimated $\mathbf{K}$ were used to induce correlations in samples from a Gaussian density; a boxcar signal was added to produce a simulated series.
- The simulated series were analyzed with the three methods used for the fMRI data.
- Since the true variance structure is know (by construction) for each voxel, the bias of the variance estimator can be directly computed.

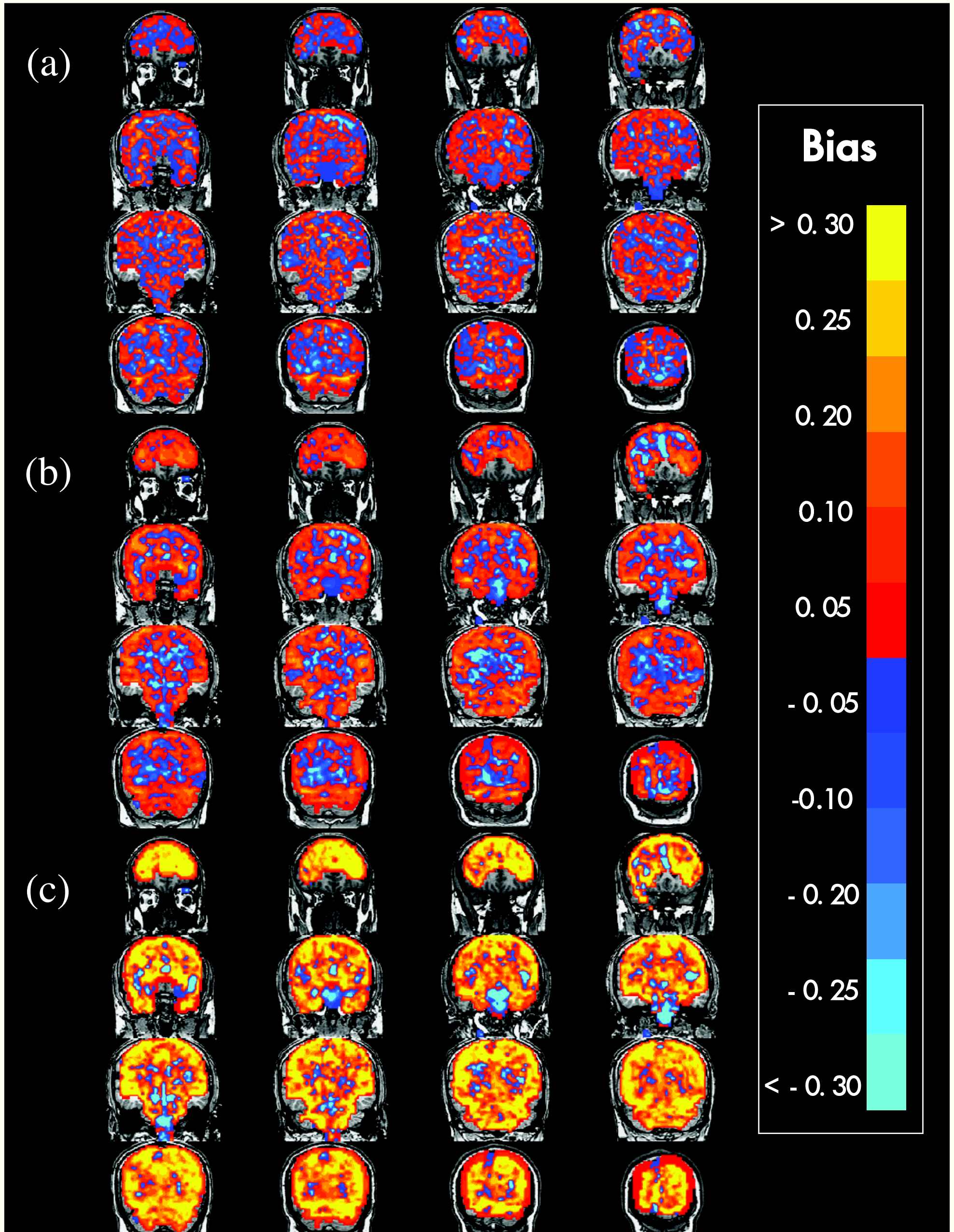
Simulation Results–Bias



Simulation Results–Variance



Simulation Results Bias Images



Summary of Results

- GCV selects appropriate degrees of smoothing of fMRI time series.
- On average, more smoothing is selected with GCV than the fixed SPM-HRF kernel.
- The bias simulation show that the GCV-spline method is, on average, unbiased.
- Spatial maps show that voxels with positive bias are primarily located in gray matter.

Discussion

- How might a fixed kernel smoother of greater bandwidth than the SPM-HRF compare to GCV?
- A more complete study might examine only gray matter voxels.
- Could over-fitting of the variance structure with an AR(8) lead to unrealistic simulated data?
- How might this method perform for event-related fMRI designs?
- To the scientist, how important is the need to make more valid inferences?

Conclusions

- It is critical that the assumptions behind every statistical procedure are checked—this can be challenging when fitting $10,000+$ models. See `www.sph.umich.edu/fni-stat/SPMd`
- Spline smoothing with GCV can be used to find an optimal smoother for an fMRI time series of block design.
- Empirically, we conclude that the GCV method selects the degree of smoothing that leads to, on average, unbiased variance estimates.
- By selecting an appropriate smoother for each time series, a substantial reduction in bias can be attained when compared to assuming that all time series require the same degree of smoothing.

Acknowledgments

- I thank Grace Wahba, Xianhong Xie, Rick Nordheim, and Beth Meyerand for their important contributions to this work.
- My support is provided under training grant T32 EY07119 from the National Institutes of Health.