Simultaneous Confidence Bands for the Coefficient Function in Functional Regression

Philip T. Reiss, New York University

Available at: https://works.bepress.com/phil_reiss/5/
Simultaneous Confidence Bands for the Coefficient Function in Functional Regression

Philip T. Reiss
New York University
phil.reiss@nyumc.org

Joint Statistical Meetings
Denver, Colorado
August 7, 2008

Joint work with Todd Ogden, Columbia

Research supported in part by grant number 5F31MH073379-02, NIMH
A motivating example: Serotonin receptors in depression

- Major depressive disorder affects $\approx 9.5\%$ of the U.S. population each year.

- Serotonin (5-HT), a neurotransmitter, is believed to play an important role in the disorder, mediated in part by distribution of 5-HT$_{1A}$ receptors in the brain.

- 5-HT$_{1A}$ receptor **binding potential** (BP), a measure of the receptors’ availability, can be mapped at each point of the brain via PET imaging.

- Question: How can we relate these BP maps to a depression-related outcome such as Hamilton Depression Score (HAM-D)?
Idea: regress HAM-D (scalar) on BP (image)

\[ y_i = \alpha + s_i^T f + \varepsilon_i \]
Functional principal component regression

We model a vector $y$ of $n$ scalar responses as

$$y = X\alpha + Sf + \varepsilon$$

where

- $X$ is an $n \times p$ covariate matrix
- $S$ is an $n \times N$ matrix:
  - each row represents a signal predictor defined at points $v_1, \ldots, v_N$, and each column has mean zero
- $\varepsilon$ denotes iid errors

Since $n \ll N$, Reiss and Ogden (2007) propose a double restriction of the coefficient function: assume $f = BV_q \beta$ where

- $B$ is an $N \times K$ matrix whose columns form a radial $B$-spline basis
- $V_q$ is a $K \times q$ matrix whose columns are the leading columns of $V$, given the singular value decomposition $SB = UDV^T$.

In other words, the (second) design matrix in the restricted model

$$y = X\alpha + SBV_q \beta + \varepsilon$$

comprises the first $q$ principal components of $SB$. 
The coefficient function estimate \( \hat{f} = BV_q \hat{\beta} \) is obtained by taking \((\hat{\alpha}, \hat{\beta})\) which minimize the penalized least squares criterion

\[
\|y - X\alpha - SBV_q \beta\|^2 + \lambda \beta^T V_q^T P V_q \beta,
\]

where \( P \) is chosen so that

\[
\beta^T V_q^T P V_q \beta \approx \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left[ \left( \frac{\partial^2 f}{\partial x_1^2} \right)^2 + 2 \left( \frac{\partial^2 f}{\partial x_1 \partial x_2} \right)^2 + \left( \frac{\partial^2 f}{\partial x_2^2} \right)^2 \right] dx_1 dx_2,
\]

a measure of the “roughness” of \( f \). The choice of \( \lambda \) governs the tradeoff between “fidelity to the data” and smoothness of the coefficient function (higher \( \lambda \longrightarrow \) smoother). Automatic criteria for choosing \( \lambda \) are generalized cross-validation (Craven and Wahba, 1979) and restricted maximum likelihood (Ruppert, Wand, and Carroll, 2003).
Why not just use SPM- or FSL-type modeling?

Standard mass-univariate approach regresses $s$ on $y$, separately at each voxel.

1. If $s$ is believed to cause $y$ (e.g., brain chemistry $\rightarrow$ psychiatric state), our model may be more appropriate.
   - But if not (e.g., $y=$gender), mass-univariate modeling makes more sense.

2. Regressing $y$ on $s$ produces subject-specific predictions.
Extension to generalized linear models

Functional principal component regression can be extended to generalized linear models

\[ g(\mu) \equiv (g(\mu_1) \ldots g(\mu_n))^T = X\alpha + Sf \]

where \( g \) is an appropriate link function. Again we apply the restriction \( f = BV_q\beta \) to obtain, e.g., the logistic model

\[ y_i \sim \text{Bernoulli} \left( \frac{\exp[(X\alpha + SBV_q\beta)_i]}{1 + \exp[(X\alpha + SBV_q\beta)_i]} \right). \]
In the linear case we have a closed-form solution
\[
\begin{pmatrix}
\hat{\alpha} \\
\hat{\beta}
\end{pmatrix} = \left[ \begin{array}{cc} X^T X & X^T Z \\ Z^T X & Z^T Z + \lambda V_q^T P V_q \end{array} \right]^{-1} \begin{pmatrix} X^T y \\ Z^T y \end{pmatrix},
\]
where \( Z = SBV_q \).

For GLMs we use penalized iteratively reweighted least squares:
\[
\begin{pmatrix}
\hat{\alpha}^{(k)} \\
\hat{\beta}^{(k)}
\end{pmatrix} = \left[ \begin{array}{cc} X^T W X & X^T W Z \\ Z^T W X & Z^T W Z + \lambda V_q^T P V_q \end{array} \right]^{-1} \begin{pmatrix} X^T W z \\ Z^T W z \end{pmatrix},
\]
where \( W \) is the weight matrix and \( z \) is the “working” dependent variable vector, both defined as for unpenalized GLMs.

The penalized IRLS procedure can often fail to converge, and may be numerically unstable even when it does converge due to numerical rank deficiency of the design matrix.

Fortunately, the same problems have been largely solved for fitting of generalized additive models (GAMs) by penalized IRLS (Wood, 2004).

The R package \texttt{mgcv} (Wood, 2006), used mainly for GAMs, can also fit functional regression models such as FPCR (as of version 1.4.0).
Nonparametric bootstrap-based simultaneous confidence bands for $f$

Basic idea:

1. Draw $B$ random samples with replacement of $n$ data points $(y^*, x^*, s^*)$ from the $n$ observations.

2. Obtain estimates $\hat{f}_1^*, \ldots, \hat{f}_B^*$ of the coefficient function.

3. Use these to form (say) 95% simultaneous confidence bands $(\hat{f}_{(L)}, \hat{f}_{(U)})$ for $f$.

4. For all $v$ such that $\hat{f}_{(L)}(v) > 0$ the coefficient function is declared significantly positive at $v$.  

Oversimplified example
Is this really a form of hypothesis testing?

- A more standard form of simultaneous inference for function estimates is to form “null bands” (simulated based on the null hypothesis $f = 0$) and see whether an observed function estimate stays within them.

- But Buja and Rolke (2007) argue for extending the concept of “adjusted $p$-values” to a general class of resampling-based bands, including simultaneous confidence bands of the kind considered here.
• In reality, the function estimates cross over each other, so choosing the 95% most central ones is more complicated.

• Mandel and Betensky (2008): Order the $B$ bootstrap estimates $\hat{f}^*_1(v) \leq \ldots \leq \hat{f}^*_B(v)$ at each point $v$, then define (for $k \geq 1$)

\[
E(k) = \prod_v [\hat{f}^*_k(v), \hat{f}^*_{(B+1-k)}(v)],
\]

the Cartesian product of symmetric 100(1 − $\frac{2k}{B+1}$)% pointwise CIs at each point. E.g., for $k = 2$: 
• Letting $E_{-b}(k)$ be the envelope formed by all except the $b$th resample, the non-coverage rate of $E(k)$ (i.e., 1 minus the simultaneous coverage rate) is estimated as the proportion of the $\hat{f}^*_b$ which exit $E_{-b}(k)$ at some point.

• This can be calculated as the proportion of $b \in \{1, \ldots, B\}$ such that, for some $v$, the rank of $\hat{f}^*_b(v)$ among $\hat{f}^*_1(v), \ldots, \hat{f}^*_B(v)$ is $\leq k$ or $\geq B + 1 - k$.

• Problem: For high-dimensional, noisy function estimates, even $E(1)$ may have less than 95% simultaneous coverage!
To surmount this difficulty we propose to stretch the bands by a constant factor. Define

\[ E^c = \prod_v [\hat{f}(v) + c\{\hat{f}^*_1(v) - \hat{f}(v)\}, \hat{f}(v) + c\{\hat{f}^*_B(v) - \hat{f}(v)\}]. \]

The modified confidence band is then \( E^{c\alpha} \), where \( c_\alpha > 1 \) is the smallest constant such that at most 100\( \alpha \)% of the \( \hat{f}^*_b \) exit \( E^{c\alpha} \).
Another concern: smoothing parameter selection

- Automatic smoothing parameter selection methods tend to undersmooth when applied to case-resampled data sets. This can result in confidence bands that are too wide.
- To correct for this, we have developed a modified generalized cross-validation criterion accounting for replicate observations, and an iterative algorithm to maximize this criterion using standard software by multiplying the degrees of freedom by a data-dependent factor.
- For simulations such as those reported here, it seems best just to multiply the df by $\approx 1.3$. 
Simulation study

To test the performance of our simultaneous inference procedure for logistic functional principal component regression, we used maps of binding potential of 5-HT$_{1A}$ receptors obtained by Parsey et al. (2006) from 27 subjects with major depressive disorder and 41 controls. We simulated binary outcomes

$$y_i \sim \text{Bernoulli} \left[ \frac{\exp(s_i^T f)}{1 + \exp(s_i^T f)} \right]$$

for $i = 1, \ldots, 68$, where

- $s_i$ denotes subject $i$'s binding potential map (1 slice, 5778 voxels), and
- $f = kf_0$.

Here $f_0$ is the artificial coefficient function on the next slide, and $k$ is chosen to attain specified values of the coefficient of determination—defined for logistic regression as

$$R^2_L = \frac{\log(L_0) - \log(L_M)}{\log(L_0)},$$

where $L_M$ is the likelihood of the given model while $L_0$ is the likelihood of the model containing only an intercept (Menard, 2000).
Simulation study, continued

- For $R^2_L = .5, .6, .7, .8, .9$, we simulated 20 sets of 68 binary outcomes

  $$y_i \sim \text{Bernoulli} \left( \frac{\exp(s_i^T f)}{1 + \exp(s_i^T f)} \right)$$

  with the true coefficient function $f = k f_0$ attaining the given $R^2_L$ value.

- For each set of outcomes, we estimated $f$ by logistic functional principal component regression with 35 PCs, and formed simultaneous confidence bands from 1999 bootstrap resamples.

- Smoothing parameter chosen by approximate corrected AIC (Hurvich, Simonoff, and Tsai, 1998), with additional correction for bootstrapping.

- Using these bands, we determined how many times (out of 20) each voxel was significantly positive (or negative).
Discussion

• Regressing scalar outcomes on entire images is a very challenging problem, but our method seems to do well at detecting salient \((f \neq 0)\) regions.

• Extension to 3D images will require tackling several practical issues.

• Another extension: locally adaptive smoothing parameter.

• Ongoing work (Todd Ogden and Yihong Zhao) uses wavelets rather than splines, and seeks a sparse representation of the coefficient function via thresholding.
Thank you!
References


