Ordinal probit wavelet-based functional models for eQTL analysis

Mark J. Meyer, Bucknell University
Jeffrey S. Morris
Craig P. Hersh
Jarret D. Morrow
Ordinal probit wavelet-based functional models for eQTL analysis

MARK J. MEYER∗

Department of Mathematics, Bucknell University, Lewisburg, PA 17837, USA
mark.meyer@bucknell.edu

JEFFREY S. MORRIS

Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX 77230-1402, USA

CRAIG P. HERSH AND JARRETT D. MORROW
Channing Division of Network Medicine, Brigham and Women’s Hospital, Harvard Medical School, Boston, MA 02115, USA

CHRISTOPH LANGE AND BRENT A. COULL

Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA

SUMMARY
Current methods for conducting expression Quantitative Trait Loci (eQTL) analysis are limited in scope to a pairwise association testing between a single nucleotide polymorphism (SNPs) and expression probe set in a region around a gene of interest, thus ignoring the inherent between-SNP correlation. To determine association, p-values are then typically adjusted using Plug-in

∗To whom correspondence should be addressed.
False Discovery Rate. As many SNPs are interrogated in the region and multiple probe-sets taken, the current approach requires the fitting of a large number of models. We propose to remedy this by introducing a flexible function-on-scalar regression that models the genome as a functional outcome. The model is formulated for a three-level ordinal categorical outcome in the Bayesian context and allows for the inclusion of a potentially large set of covariates. We examine the properties of the model in both simulation and in application to a chronic obstructive pulmonary disease genetic data set where eQTL analysis is of interest alongside a comparison of the standard approach.

Key words: Generalized function-on-scalar regression; Functional data analysis; Bayesian inference; Wavelet regression; eQTL Analysis.

1. Introduction

Expression Quantitative Trait Loci (eQTL) Analysis examines the association between expression levels of a gene measured using microarray and a fine mapping of single nucleotide polymorphisms (SNPs) in the same region as the gene. One goal of an eQTL analysis is to find the functional variant of the gene and to identify markers in the region associated with its expression. To obtain the fine mapping, researchers commonly interrogate the genome within a specified distance of the start and end of a gene.

The resulting data is high dimensional and complex in nature yet the standard analytical practice is rather straightforward and ignores the data’s potential complexity. Currently, the standard approach used for eQTL analysis involves pair-wise regressions comparing a single expression probe set to a single SNP at a time and determining a p-value. This procedure is repeated for each SNP sampled in the region with adjustments, such as the Plug-in False Discovery Rate (FDR) or Benjamini-Hochberg FDR, implemented to account for multiple testing. Such an analysis can be seen, for example, in both Qiu and others (2011) and Castaldi and others (2015) where
general linear models are used to assess association between the probe set, which has been normalized and is thus assumed to be a Gaussian outcome, and SNP while adjusting for phenotypes and population stratification. This approach, however, ostensibly assumes independence amongst the SNP-probe set relationships and thus ignores the spatial relationships among interrogated genotypes. Further, a separate model is fit for each relationship resulting in a potentially large number of models. In fact, very few procedures exist that attempt to model all probe sets and all SNPs concurrently.

Some existing statistical literature does propose a modeling procedure that treats genotype as a set of covariates and used a modified BIC approach to eliminate non-significant SNPs (Zak-Szatkowska and Bogdan, 2011; Frommlet and others, 2012). However, the modified BIC approach does not allow for multiple probe sets. Further, inclusion of all SNPs of interest as covariates may induce issues with dimensionality. Flutre and others (2013) proposed a Bayesian model averaging (BMA) framework for an eQTL analysis that models multiple probe sets from different tissues. This BMA method, however, only examines one candidate SNP at a time. Thus the issue of running a model for every SNP of interest remains and the need exists for a method that simultaneously accounts for all SNPs and all probe sets.

In this paper, we propose a novel eQTL analysis that treats SNP as a function of location on the genome and then models it as the outcome using functional regression—a popular technique for modeling functional data where the unit observation is a curve or potentially a set of curves measured over a finely sampled grid. While several types of functional regression exist, we are interested in regressing a functional response on to a scalar covariate or a set of scalar covariates, otherwise known as function-on-scalar regression.

Beginning with Faraway (1997) and Ramsay and Silverman (1997), the literature for function-on-scalar regression has grown considerably. Contributions by Wu and Chiang (2000), Chiang and others (2001), Shi and others (2007), and Reiss, Huang, and Mennes (2010) considered
kernel smoothing and spline-based approaches to modeling a functional outcome while Krafty and others (2008) explore the use of principle components. Morris and Carroll (2006) introduced Wavelet-based Functional Mixed Models (WFMM) for function-on-scalar regression via Bayesian estimation. The WFMM is a flexible framework for modeling functional outcomes in a number of settings and indeed several authors have extended this methodology. Zhu, Brown, and Morris (2011, 2012) discussed robust adaptive regression and robust classification respectively. Meyer and others (2015) introduced the function-on-function extension of the WFMM. These approaches, however, all make the assumption that the functional response is Gaussian. Since non-imputed SNPs take on the integer values 0, 1 and 2—which counts the occurrence of the least common allele at that SNP—we cannot assume it to be Gaussian. Thus we require an approach for generalized outcomes.

Existing literature for generalized outcomes primarily focuses on scalar outcomes, with very little dealing with generalized functional responses, see Morris (2015) for an overview of this and functional regression in general. The first paper to directly deal with generalized functional outcomes is Goldsmith, Zipunnikov, and Schrack (2015) who developed a model for multilevel generalized function-on-scalar regression using functional principal components (fPC) and penalized splines. Their approach used logit and log link functions to model binary and count functional response data, respectively, for analyzing physical activity curves. For inference, the authors used point-wise posterior credible intervals. This work did not consider functional ordinal outcomes, and all inference was based on tests that did not account for multiple testing.

Thus we introduce the Ordinal Probit Wavelet-based Functional Model (OPWAVFM) for regressing an ordinal functional outcome and on scalar covariates in a Bayesian context. The model we propose allows for adaptive regularization of coefficients and thus smoothing across the functional regression coefficients. The OPWAVFM constitutes an extension of the WFMM framework to the non-Gaussian setting and allows for the inclusion of many covariates of interest.
as well as a large number of outcome measurements. We propose an MCMC algorithm for generating posterior estimates of model parameters combining a modified Bayesian Probit regression with the WFMM framework. Estimation is achieved using a latent variable representation of the Probit model and the flexibility of the WFMM.

Using this approach, we achieve a model that incorporates information from all interrogated SNPs in the region while jointly modeling all probe sets of interest. The benefit of modeling the genome as a function, as shown by Lee and Morris (2015) in the context of whole-genome methylation arrays, is that by using basis functions we can adaptively integrate information among nearby genomic locations thus providing better efficiency than the typical independent marker-by-marker analysis. By using wavelet basis functions, we can better detect and model spikes and localities in the signals which is more likely to occur in genetic data. As demonstrated by Sardy and others (1999) and Hsu and others (2005), the use of wavelets over other basis functions also alleviates the need to adjust for unequally spaced measurements which are common in SNP data. The advantages of a wavelet analysis over existing approaches are demonstrated in the analysis of a chronic obstructive pulmonary disease genetic data presented in Section 4.

We also propose the use of several procedures for performing posterior functional inference while accounting for multiple testing using FDR or experiment-wise error rate criteria. Previous work in the WFMM context implements both the Bayesian False Discovery Rate (BFDR), Morris and others (2008) and Malloy and others (2010), and Simultaneous Band Scores (SimBaS), Meyer and others (2015). We formulate the OPWAVFM version of these posterior inference procedures while noting that any statistic of interest can be calculated from our posterior samples.

The remainder of the paper is organized as follows: Section 2 presents the model formulation and inference procedures. In Section 3, we describe a simulation study demonstrating the abilities of our method overall and relative the standard analysis. In Section 4, we apply the OPWAVFM to an example genomic data set and compare our approach to the standard analysis and in Section
we give a discussion of the methodology.

2. Ordinal Probit Functional Model

Here we detail the modeling framework for the OPWAVFM. Let $Y_i(t)$ be the observed genotype for subject $i$ at location $t$ along the chromosome, $i = 1, \ldots, N$ and $t = t_1, \ldots, t_T$. Thus $Y_i(t)$ takes on the values $g = \{0, 1, 2\}$ which counts the occurrence of the least common allele at SNP $t$. Note that $Y_i(t)$ need not necessarily be sampled on an equally spaced grid as $t$ only indexes measurement locations on the chromosome. Further, let $X_i$ represent subject $i$’s scalar covariates of interest. Common covariates in eQTL analysis include expression probe sets, age, sex, array lot number, and genetic ancestry. For model formulation and without-loss of generality, we assume $X_i$ is a single scalar covariate, however it may also be a matrix of scalar covariates.

Suppose $Y_i(t)$ is actually the observable value of some latent process $Y_i^*(t)$ for subject $i$ at measurement occurrence $t$. Then the behavior of $Y_i(t)$ is dictated by the relationship

$$Y_i(t) = \begin{cases} 
0 & \text{if } Y_i^*(t) < c_1 \\
1 & \text{if } c_1 \leq Y_i^*(t) < c_2 \\
2 & \text{if } Y_i^*(t) \geq c_2 
\end{cases} \quad (2.1)$$

for cut points $c_1$ and $c_2$, satisfying $c_1 < c_2$. This formulation can be extended to more levels; however we restrict it to the three level case. Using the mapping in Model (2.1), the probability that $Y_i(t)$ equals the $g$th level can be expressed as

$$P[Y_i(t) = g] = P[Y_i^*(t) \in (c_g, c_{g+1})], \; g = 0, 1, 2, \quad (2.2)$$

where $c_0$ and $c_3$ will vary depending on the support of $Y_i^*(t)$. Now let the form of $Y_i^*(t)$ be the function-on-scalar regression model

$$Y_i^*(t) = X_i \beta(t) + E_i(t), \quad (2.3)$$

where the model errors, $E_i(t)$, could come from a variety of distributions. Note that choice of this distribution will dictate the remainder of the modeling framework.
If we assume Gaussian Process errors, then $Y_i^*(t)$ is also Gaussian. Normalizing $Y_i^*(t)$, we can re-express Model (2.2) in terms of the CDF of the standard Gaussian, denoted $\Phi(\cdot)$. The probabilities at a fixed $t$ for subject $i$ are then

$$P[Y_i(t) = g] = \Phi[c_{g+1} - X_i\beta(t)] - \Phi[c_g - X_i\beta(t)]$$ (2.4)

where $g = 0, 1, 2$. This assumption results in the Probit formulation of the model.

Models (2.3) and (2.4) are formulated for continuous functions, however we only observe discretized realizations. Assuming all functions are sampled on the same—not necessarily equally spaced—grid of size $T$, the discrete version of Model (2.3) is

$$Y^* = X\beta + E, \ E \sim \mathcal{GP}(0, \Sigma_E),$$ (2.5)

where $Y^*$ and $E$ are $N \times T$, $X$ is $N \times 1$, and $\beta$ is $1 \times T$. The covariate matrix, $X$, can be of size $N \times P$ depending on the desired number of covariates, $P$. If $P > 1$, then $\beta$ becomes $P \times T$ with one function per covariate. Now let $y_{it}$ represent subject $i$'s $t$th outcome, $\beta_t$ corresponds to the $t$th element of $\beta$, and $x_i$ denotes subject $i$'s covariate pattern. Model (2.4) can then be written as

$$P(y_{it} = g) = \Phi(c_{g+1} - x_i\beta_t) - \Phi(c_g - x_i\beta_t)$$ for $g = 0, 1, 2$.

Non-functional Bayesian Probit regression utilizes a similar latent variable formulation to produce model estimates by sampling from the latent outcome (Albert and Chib, 1993). In this approach, the latent outcome is assumed independent which does not hold in the functional setting. However, we can assume independence after a wavelet transformation of $Y^*$. If we allow the variance components to vary across the wavelets coefficients, then assuming independence in the wavelet-space does not imply independence in the data-space and actually induces local dependencies in the data-space thereby integrating information across SNPs into the model. This was the strategy employed by Morris and Carroll (2006). Thus, after transforming, we estimate associations between this latent response and covariates in this transformed space.
2.1 Wavelet-based Modeling of the Latent Outcome

Morris and Carroll (2006) develop a function-on-scalar regression model for hierarchical data. Thus their formulation allows for the observation and modeling of multiple curves measured on each subject. When only one curve is observed per subject, their model reduces to the latent variable model found in Model (2.3). Given this relationship, we can extend this modeling framework within a Bayesian probit model to estimate $\beta(t)$.

2.1.1 Model Formulation Working from the discretized model (2.5) and applying a Discrete Wavelet Transformation (DWT) to the latent outcome gives the decomposition $Y^* = Y^{*W}W$ where $Y^{*W}$ is the resulting wavelet-space coefficients and $W$ is a matrix of wavelet basis functions. Use of the DWT requires selecting a type of wavelet as well as a boundary padding procedure. Common choices include Daubechies with varying vanishing moments for the the mother wavelets and symmetric half-point or zero-padding for the boundary padding. Decomposing $\beta$ gives $\beta = \beta^W W$ and the DWT applied to $E$ results in $E = E^W W$. Given these decompositions, Model (2.5) can be expressed as

$$Y^{*W} = X\beta^W W + E^W W.$$  \hspace{1cm} (2.6)

Note that the matrix representations of the wavelet basis are orthogonal, thus $WW' = I_{T^*}$ a $T^* \times T^*$ identity matrix where $T^*$ is the number of wavelet basis functions. Post-multiplying Model (2.6) by $W'$ gives

$$Y^* = X\beta^W + E^W,$$  \hspace{1cm} (2.7)

where $E^W \sim GP(0, \Sigma_{E^W})$ since $\Sigma_{E} = \Sigma_{E^W} W$. We then use an MCMC procedure to obtain posterior estimates of $\beta^W$, transforming them back into the data space through $\beta = \beta^W W$ before performing inference.

The fast DWT assumes equally spaced data, while here the SNPs are not equally spaced.
The lifting scheme of Sweldens (1996) can be used to account for the spacing. Sardy and others (1999) showed that wavelets for equally spaced data can be used, for which the domain of the wavelet basis functions are effectively $1, \ldots, T$ instead of $t_1, \ldots, t_T$. This approach gives similar results as the lifting scheme but runs much faster. Further, Hsu and others (2005) implemented this approach in the genomic context, so this is the strategy we use here.

2.1.2 Prior Specification and Identifiability  
First, we assume independence in the wavelet space. Then, Model (2.7) can be split up into a series of $T^*$ separate models for each coefficient in the wavelet space, double-indexed by $(j, k)$, resulting in

$$y_{(j,k)}^W = X\beta_{(j,k)}^W + \mathbf{e}_{(j,k)}^W. \quad (2.8)$$

Next, we place spike and slab priors on the coefficients $\beta_{(j,k)}^W = \{\beta_{(p,j,k)}^W\}$ where $p$ indexes the number of columns of $X$. Thus the prior on the coefficients from Model (2.8) is

$$\beta_{(j,k)}^W \sim \gamma_{(p,j,k)}N(0, \tau_{pj}) + (1 - \gamma_{(p,j,k)})d_0, \quad \gamma_{(p,j,k)} \sim \mathcal{B} (\pi_{pj}), \quad (2.9)$$

where $\mathcal{B}$ denotes the Bernoulli distribution and $d_0$ is a point-mass distribution at zero. This adaptive regularization performs smoothing in the wavelet space. Wavelet-space independence and spike and slab priors are consistent with the previous literature on wavelet-based regression (Morris and Carroll, 2006; Malloy and others, 2010; Meyer and others, 2015).

Regularization parameters can either be sampled or fixed. Both Morris and Carroll (2006) and Malloy and others (2010) rely on an empirical Bayes approach to estimate and then fix $\tau_{pj}$ and $\pi_{pj}$. However Zhu, Brown, and Morris (2011) and Meyer and others (2015) propose using an inverse-gamma distribution for $\tau_{pj}$ and a beta distribution for $\pi_{pj}$. Thus we place priors on both of the form $\tau_{pj} \sim IG(a_\tau, b_\tau)$ and $\pi_{pj} \sim Beta(a_\pi, b_\pi)$, where hyper-parameters $a_\tau, b_{\tau_0}, a_\pi$, and $b_\pi$ are fixed and based on empirical Bayes estimates as described in Morris and Carroll (2006), calculated using the initial estimated latent outcome which is taken to be a vector of zeros.
An assumption of standard Bayesian Probit regression is that the mapping from \( Y_i(t) \) to \( Y_i^*(t) \) captures the location and scale of \( Y_i(t) \). For identifiability purposes in a Bayesian Probit regression, we assume the intercept is fixed at zero and, for \( Y^*|X, \beta \sim \mathcal{N}(X\beta, \Sigma_E) \), we assume \( \Sigma_E \) is fixed to be \( I_N \), the \( N \times N \) identity Matrix. The other identifiability issue involves the cut points dictating the mapping of \( Y_i(t) \) to \( Y_i^*(t) \), namely \( c_1 \) and \( c_2 \). We implement the approach used by Albert and Chib (1993) which fixes \( c_1 \) at 0 and samples \( c_2 \) from a uniform distribution, \( \mathcal{U}(a_{c_2}, b_{c_2}) \), where \( a_{c_2} \) and \( b_{c_2} \) are determined using \( c_1 \) and information from the data.

The MCMC algorithm involves sampling from truncated normal distributions to update \( Y^* \), then a uniform distribution to update \( c_2 \), followed by projecting \( Y^* \) into the wavelet space, updating \( \beta^W \) and the regularization parameters, and finally projecting \( \beta^W \) into the data space. A full description of the sampler can be found in Appendix A of the Supplementary Materials.

### 2.2 Posterior Functional Inference

Our key inferential goal is to identify regions of \( \beta(t) \) different from zero, indicating regions where genotype is associated with the predictor. Previous authors have detailed two procedures for inference: the BFDR and SimBaS (Müller, Parmigiani, and Rice (2006); Morris and others (2008); Malloy and others (2010); Meyer and others (2015)). Both procedures have advantages over the use of Plug-in FDR and Benjamini-Hochberg FDR that are commonplace in standard eQTL analyses. While existing methods do control for the false discovery rate, they assume independence between markers. Both BFDR and SimBaS account for possible marker-marker correlations while adjusting for multiplicity in their formulation, which in principle can lead to greater efficiency and power to detect significantly associated regions. For completeness, we now discuss these procedures in the OPWAVFM setting.

Let \( m = 1, \ldots, M \) MCMC samples. Then \( \beta^{(m)}(t) \) is the \( m \)th draw from the posterior and for
OPWAVFMs for eQTL analysis

a specific $t$, $t = 1, \ldots, T$, we calculate

$$P_{BFDR}(t) = Pr \left[ |β(t)| > δ | y \right] \approx \frac{1}{M} \sum_{m=1}^{M} 1 \left[ |β^{(m)}(t)| > δ \right].$$

For a pre-specified global FDR-bound $α$, we then flag a set of locations satisfying, $ψ$, defined as

$$ψ = \{ (t) : P_{BFDR}(t) ≥ ν_α \}$$

where $ν_α = P(λ)$ and, given the ordered set $\{ P_r : r = 1, \ldots, R \}$ for $R = T$, the cutoff value $λ = \text{max} \left[ r^* : \frac{1}{r^*} \sum_{r=1}^{r^*} \{ 1 - P_r \} ≤ α \right]$. Choice of $δ$ can be based on “expert knowledge,” but is still a somewhat arbitrary choice, as discussed in Meyer and others (2015). The SimBaS procedure does not require the selection of an intensity change.

To formulate the SimBaS for OPWAVFM, we begin by working from the joint credible bands described by Ruppert, Wand, and Carroll (2003), and construct a $100(1 - α)$% credible band about $β(t)$. For the mean and standard deviation of $β(t)$ taken over the $M$ MCMC samples, we construct the interval $I_α(t) = \hat{β}(t) ± q(1 - α) \left[ \text{St.Dev} \{ \hat{β}(t) \} \right]$. The value $q(1 - α)$ is then the $(1 - α)$ quantile taken over the $M$ MCMC samples of the quantity

$$Z^{(m)} = \max \{ v \in V, t \in T \} \frac{β^{(m)}(t) - \hat{β}(t)}{\text{St.Dev} \{ \hat{β}(t) \}}.$$

Next we vary the values of $α$, noting the minimum $α$ at which $I_α(t)$ excludes 0. These values, more formally defined by $P_{SimBaS}(t) = \min \{ α : 0 /∈ I_α(v, t) \}$, constitute the scores for SimBaS.

More directly, we can calculate SimBa Scores using

$$P_{SimBaS}(t) = \frac{1}{M} \sum_{m=1}^{M} 1 \left\{ \left| \frac{\hat{β}(t)}{\text{St.Dev} \{ \hat{β}(t) \}} \right| ≤ Z^{(m)} \right\}.$$

Each score essentially provides a test of the null $H_0 : β(t) = 0$ for a specific location $t$ while retaining joint coverage probabilities. We can also calculate a global Bayesian p-value (GBP) using SimBaS. Similar to Meyer and others (2015), the form of the GBP is $P_{GBP} = \min \{ P_{SimBaS}(t) \}$ which we can use to test the global null hypothesis $H_0 : β(t) = 0 \forall t = 1, \ldots, T$. Matlab code to implement both the sampler as well as the inference procedures can be obtained online at http://works.bepress.com/markjmeyer/6/.
3. Simulation

Simulations are modeled after SNP and gene expression data. To simulate SNPs, we start by generating probabilities for subject $i$ at measurement $t$ using $P_0 = \Phi [c_1 - X_i \beta(t)]$, $P_1 = \Phi [c_2 - X_i \beta(t)] - \Phi [c_1 - X_i \beta(t)]$, and $P_2 = 1 - \Phi [c_2 - X_i \beta(t)]$, where $X_i$ is drawn from a standard normal distribution to mimic a single expression probe set. We select $c_1 = 0$ and $c_2 = 0.5$.

Simulated SNP values, $Y_i(t)$, were generated by first sampling a standard uniform random variable for each subject $i$ and measurement $t$, call that value $u \sim U(0, 1)$, and then assigning 0, 1, or 2 using $Y_i(t) = 0 \iff u \in (0, P_0)$, $Y_i(t) = 1 \iff u \in (P_0, P_0 + P_1)$ and $Y_i(t) = 2 \iff u \in (P_0 + P_1, 1)$. The total number of measurements generated was $T = 256$ for $N = 200$ subjects.

For the DWT on the latent outcome, a choice of padding is needed. Consistent with Malloy and others (2010) and Meyer and others (2015), we use zero padding which in preliminary simulations demonstrated the least amount of edge effects. Other choices of padding, such as symmetric half point, can be implemented however they tend to amplify edge effects.

Two hundred datasets were generated for four scenarios of $\beta(t)$ for a single simulated probe set. Two of these scenarios have different shaped peaks centered in a region of interest:

\[
\beta(t) = \frac{10}{\sqrt{2\pi(10)}} \exp \left[ -\frac{1}{2} \frac{(t - 128)^2}{10} \right], \text{ single normal}
\]
\[
\beta(t) = \frac{2}{5} \max \left[ 1 - \frac{|t - 128|}{32}, 0 \right], \text{ single triangle.}
\]

The remaining scenarios have a central peak of the same magnitude as the single normal and the single triangle centered at $t = 128$ along with an additional attenuated peak centered at $t = 44$:

\[
\beta(t) = \frac{10}{\sqrt{2\pi(10)}} \exp \left[ -\frac{1}{2} \frac{(t - 128)^2}{10} \right] + \frac{10}{3\sqrt{2\pi(10)}} \exp \left[ -\frac{1}{2} \frac{(t - 44)^2}{10} \right], \text{ double normal}
\]
\[
\beta(t) = \frac{2}{5} \max \left[ 1 - \frac{|t - 128|}{32}, 0 \right] + \frac{1}{7} \max \left[ 1 - \frac{|t - 44|}{32}, 0 \right], \text{ double triangle.}
\]

Each of these true scenarios is included as the dark gray solid curve in Figure 1. These scenarios represent settings where regions of the genome are associated with varying effect sizes with
a single probe set of interest. After a burn-in of 5000 iterations, we sampled 1000 iterations for posterior estimation and inference. For each model, we use Daubechies wavelets with four vanishing moments.

Because interest focuses on including multiple probe sets as covariates in our application, we also examine the abilities of the OPWAVFM to detect signals from multiple covariates. One potential issue with the inclusion of multiple probe sets is that it may induce collinearity. Thus our final simulation setting involves a combination of two scenarios: the single normal, denoted as \( \beta_1(t) \), and the double normal, denoted as \( \beta_2(t) \). To examine the effects of collinearity, we compare three levels of correlation between the two covariates. The two covariates, \( x_1 \) and \( x_2 \), for this setting come from a mean zero bivariate normal distribution with \( \text{var}(x_1) = \text{var}(x_2) = 1 \) and varying correlation, \( \text{corr}(x_1, x_2) = 0, 0.5, \) and 0.9. To evaluate this scenario and the previous four single covariate settings, we examine the average estimation and point-wise bias calculated as the average of the difference between the estimate and the truth at each time point as well as root-Mean Square Error (rMSE) calculated for each simulated data set.

To evaluate inference, we implement three measures used by Meyer and others (2015) to assess both BFDR and SimBaS. First define a flagged coefficient as a coefficient belonging to the set \( \psi \) when using BFDR with global \( \alpha = 0.05 \) and score less than 0.05 when using SimBaS. Define the measure false discovery rate, denoted \( \text{FDR}_\epsilon \), as the number of flagged coefficients with true value \( \leq \epsilon \) divided by the total number of flagged locations. The second measure is sensitivity, denoted \( \text{SEN}_\Upsilon \), which we define as the number of flagged coefficients with true magnitude \( > \Upsilon \) divided by the number of true coefficients with magnitude \( > \Upsilon \). The final measure is experiment-wise error rate or \( \text{EWER}_\epsilon \), which is calculated as the proportion of simulated datasets with at least one falsely discovered location, i.e. a selected location with true value \( \leq \epsilon \). For our evaluation, we consider a range of both \( \epsilon \) and \( \Upsilon \) with \( \epsilon = 0.0675 \) to 0.2 and \( \Upsilon = 0.1 \) to 0.365. We compare these values for SimBaS and the BFDR for a \( \delta \) equal to half the maximum simulated signal, \( \delta = 0.2 \).
For comparison, we also apply the standard approach to each simulation scenario, implementing both the Benjamini-Hochberg FDR and the Plug-in FDR to adjust for multiplicity.

Figure 1 shows posterior means from a single simulated data set for each $\beta(t)$ along with 95% joint credible bands calculated as described by Ruppert, Wand, and Carroll (2003). The true $\beta(t)$ is plotted for each scenario as well for comparison. In each scenario, the joint credible bands consistently include the truth. Additionally, the each posterior mean effectively reproduces the true function and consistently detects the peaks in each scenario. Estimation in the two scalar covariates scenarios behaves similarly to single covariate scenarios, regardless of correlation. Bias and rMSE are minimal for all scenarios. The results for the multiple covariate scenarios as well as bias and rMSE plots can be found in Appendix C of the Supplementary Material.

To compare all inference approaches, we plot curves of SEN$_{\Upsilon}$, FDR$_{\epsilon}$, and EWER$_{\epsilon}$ for each, varying $\Upsilon$ and $\epsilon$. The curves resulting from the single covariate, double normal scenario are in Figure 2. Both SimBaS and the BFDR with $\delta = 0.2$ perform similarly well on all measures. Further, they outperform the standard approach regardless of the adjustment selected. This is particularly noticeable when examining FDR$_{\epsilon}$ and EWER$_{\epsilon}$ as the standard approaches have consistently higher values than either SimBaS or the BFDR. In fact, in most scenarios, the BFDR and SimBaS have consistently higher SEN$_{\Upsilon}$ and lower FDR$_{\epsilon}$ and EWER$_{\epsilon}$ than the standard approach. Similar plots to those found in Figure 2 can be found in Appendix C of the Supplementary Material for the remaining scenarios. As the BFDR requires a choice of $\delta$, we can consider calculating these measures for a range of $\delta$ values. Heat-maps depicting SEN$_{\Upsilon}$ and FDR$_{\epsilon}$ for varying $\delta$, $\Upsilon$, and $\epsilon$ can be found in Appendix C of the Supplementary Material. But regardless of $\delta$, the BFDR controls false discovery rate well while maintaining high sensitivity.

4. Application

Motivating data comes from the TESRA study of chronic obstructive pulmonary disease (COPD) which collected genotypes and expression profiles for 202 subjects (Cheng and others, 2013;
Of interest are genotyped SNPs from chromosome 15 near the IREB2 gene, which has been previously identified as a disease susceptibility locus (Pillai and others, 2009; Wilk and others, 2009; Cho and others, 2010). The gene expression array contained four different expression probe sets for IREB2 with probes starting at 78730518 bp and ending at 78793798 bp (human genome version 19). We consider five models: one for each probe set separately and one jointly modeling the three probe sets that exhibited a GBPV less than 0.05 in the separate models. All models are adjusted for age, sex, array lot number, and genetic ancestry. The four expression probe sets of interest are 1555476_at, 214666_s_at, 225892_at, and 242261_s_at.

For comparison, we conducted a standard analysis using two modeling approaches. As the standard analysis treats probe set as the outcome and SNP as the predictor we must consider how we treat SNP as a covariate as it can be either a linear or a categorical effect. We will examine both and determine the appropriate adjusted p-values using the Plug-in FDR as well as the Benjamini-Hochberg FDR and the Bonferroni correction for comparison. All of these models are also adjusted for age, sex, array lot number, and genetic ancestry.

Commonly, a fine mapping of SNPs around a candidate gene is sampled and all SNPs falling within a specified distance of the start and end of the gene, such as 250 kilobases (kb) or 2 megabases (Mb), are interrogated. Taking 250kb to either side of IREB2 results in 190 total SNPs. Examining 2Mb to either side of the gene results in 1135 SNPs. As with the simulation, all models were run using Daubechies wavelets with 4 vanishing moments and zero padding. Each model was run for 6000 total samples with the first 5000 discarded. Models taking 250kb to either side took just over an hour to finish while taking 2Mb to either side took just under three hours running on a high performance computing cluster. For brevity, only results from the 250kb models are presented here. Results from the 2Mb models can be found in Appendix D of the Supplementary Material.

Since we are interested in identifying SNPs that are significantly associated with probe sets,
we further focus our analysis on first detecting SNPs that are significantly different from zero based on SimBaS. Next we examine those SNPs at varying intensity cutoffs using probabilities from the BFDR. Finally, we compare the association of these SNPs with different expression probe sets in both the single probe set models and the joint probe set model. Manhattan-style plots of SimBa Scores from the single probe set models can be found in Figure 3 while Figure 4 contains Scores from the joint probe set model. In each figure, the dotted-dashed horizontal gray line represents a global $\alpha$ of 0.05.

Ostensibly, the SimBa Score is a multiplicity adjusted probability testing the null hypothesis of no association between a specific SNP and the expression probe set. Thus as a first step, we can determine which SNPs are significantly associated with each probe set separately. In Figure 3, we exclude the results related to probe sets 225892 at as no significant SNPs were detected. Thus for our joint model, we exclude probe set 225892 at. To identify candidate SNPs, we use the joint model for the expression probe sets 1555476 at, 214666 x at, and 242261 at considering only SNPs with SimBa Scores at or above the global $\alpha$-level of 0.05. This results in seven SNPs associated with probe set 1555476 at, four with 214666 x at, and six with 242261 at.

Table 1 contains the SimBa Scores and BFDR probabilities of SNPs detected in the joint model. For the joint model, both SimBaS and BFDR were performed across all expression probe sets simultaneously. Interestingly, for 1555476 at, only one of the SNPs identified was significant in the single probe set model while for 214666 x at two of the four were significant in both models. This is in contrast to 242261 at where all but one SNP was found significant in both models.

We can use the BFDR to both flag significant coefficients and to determine posterior probabilities that a given coefficient falls above a pre-determined threshold. For Table 1, we use three different thresholds on the probability scale: $\delta = 0.10$, $0.15$, and $0.20$. Examining all three allows us to see the magnitude at which these coefficients are still significant. This is useful as significance via the SimBa Score does not necessarily translate to a large effect size. The SimBa Score
only tests if the association is different from zero it does not give us an idea of the magnitude of effect. Of potential interest is determining SNPs that not only are significantly associated but also demonstrate meaningful effect sizes.

For instance, four of the SNPs that are significantly associated with probe set 1555476_at have BFDR probabilities above 90% for all values of \( \delta \). This suggests that not only are these SNPs significantly associated, but they also have a large magnitude of effect with over 0.90 probability of being larger than 0.2. Conversely, most of the SNPs that are significantly associated with probe set 214666_x_at do not have large effect sizes. One could consider using SimBaS to determine significantly associated SNPs and then selecting a sufficiently large cutoff \( \delta \) to narrow in on SNPs that are not only significant but also have large effect sizes.

One final observation is that each probe set is associated with different sets of SNPs. In fact there is no overlap between SNPs associated with 214666_x_at and the other two probe sets. This is not surprising as probe sets designated x may hybridize with other regions in the genome, and therefore measure other genes. Thus results from from 214666_x_at should be viewed with a degree of skepticism. However the remaining probe sets do have considerable overlap with three SNPs commonly associated with 1555476_at and 242261_at. All three of those SNPs, rs8034191, rs2036527, and rs16969968, maintain large BFDR probabilities as \( \delta \) increases. SNPs rs2036527 and rs16969968 both have rather small SimBa Scores from the joint model while rs8034191’s score falls right on the cutoff value.

In contrast, the standard analysis, when using the popular Plug-in FDR based on 1000 permutations, failed to detect any significant SNPs. Only when using Benjamini-Hochberg and Bonferroni did the standard analysis detect any SNPs. Table 2 contains the minimum adjusted p-values for both modeling approaches and the three corrections. When SNP is treated linearly, no pairwise associations are considered significant at the nominal level via any correction. Only when SNP is treated categorically are potentially significant SNPs detected. In the case of probe set
the Benjamini-Hochberg correction detects four significant SNPs while the Bonferroni correction only detects one. However, the Plug-in FDR does not detect any significant SNPs.

5. DISCUSSION

Increasingly complex genetic data requires the development of advanced analytical procedures that can better capture features of interest. In the eQTL setting, several attempts have been made to accommodate either all SNPs or all probe sets, yet no model has attempted to address both. Further, functional data continues to be collected in vast quantities across many scientific disciplines. To keep up with the demand for inferential methods for such data we must continue to expand the field of functional data analysis. While methods for Gaussian functional outcomes and scalar categorical outcomes with functional predictors are abundant in the literature, functional categorical outcomes have received little attention.

Here we have presented a novel method for conducting eQTL analysis using a function-on-scalar regression technique where the outcome comes from finely sampled categorical process. To implement this methodology, we have developed an MCMC procedure which builds on the framework of the WFMM while utilizing the latent variable representation of the probit model. As our outcome of interest in application had three levels, we proposed an ordinal model which can easily be refined down to the binary case or extended to allow for more than three categories. Further, we apply the BFDR and SimBaS procedures in the OPWAVFM framework.

To evaluate the operating characteristics of the OPWAVFM, we presented seven simulation scenarios which display the ability of our method to detect various true signals at a sample size that is modest in the context of genomic studies. Additionally, we examined the behavior of the model for multiple predictors of varying levels of correlation. In all scenarios, the OPWAVFM yields estimates with low bias and rMSE. Even when correlation is induced between multiple covariates, bias is minimal. Also in simulation, we show that both the BFDR and SimBaS have good properties, displaying low false discovery rates and experiment-wise error rate and high
sensitivity. In applying the method to data interrogating the region around IREB2 in COPD patients, we illustrate the use of the OPWAVFM to perform an eQTL analysis both with a single expression probe set of interest and with multiple probe sets. Further we note the ability of our inference procedures: the GBPV indicated which probe sets lacked significant associations, SimBaS detected several significant SNPs for the remaining probe sets, and the BFDR quantified the strength of association between multiple SNPs and a given probe set. Finally, the OPWAVFM is able to scale up to a large number of SNPs at only a minimal computational cost.

In comparison to the standard approach, several challenges associated with the standard approach are addressed by the OPVAFM. First, we no longer have to fit all pair-wise models and thus are no longer essentially modeling the relationships between SNPs and probe sets as independent, but can integrate information across correlated SNPs through the basis space modeling inherent to our functional regression method. Second, in simulation we demonstrate the short comings of the standard approach. Third, by maximizing the correlation across the genetic region, we were able to detect more candidate SNPs than either approach to the standard analysis.

6. Supplementary Material

The reader is referred to the online Supplementary Materials for technical appendices regarding the MCMC sampler and an extension to generalized function-on-function regression, as well as additional simulations and additional application results.

Acknowledgments

This work was supported by grants from the National Institutes of Health (ES-007142, ES-000002, CA-134294, CA-178744, R01MH081862) and the Cure Alzheimer’s Fund. The TESRA study was supported by Roche.


REFERENCES


[Received October, 2015; revised ?; accepted for publication ?]
Fig. 1. Posterior estimates of $\beta(t)$ as a function of $t$ from a single simulated dataset with near average rMSE. Light gray bands depict the 95% joint credible bands. Dashed lines are posterior means while true functions are in solid dark gray. The top row contains estimates for the single peak scenarios, bottom row contains double peak.
Fig. 2. Plotted SEN, FDR, and EWER, for the Plug-in FDR and Benjamini-Hochberg FDR using the standard approach and SimBaS and the Bayesian FDR with \( \delta = 0.2 \) using the OPWAVFM. The single covariate setting with the double normal effect is presented here.
Fig. 3. SimBa Scores for single probe models plotted as functions of position on the chromosome. The location of IREB2 is noted as a horizontal bar below the probabilities. For convenience, a dotted-dashed gray line depicts a global α-level of 0.05 plotted on the $-\log_{10}$ scale.
Fig. 4. Joint model SimBa Scores plotted as functions of position on the chromosome. The location of IREB2 is noted as a horizontal bar below the probabilities. For convenience, a dotted-dashed gray line depicts a global $\alpha$-level of 0.05 plotted on the $-\log_{10}$ scale.
Table 1. Significant SNPs from the joint model by expression probe set. Significance is based on the joint model SimBa Score exceeding the global alpha of 0.05. Joint model scores are also compared to single probe scores for the same SNP as well as BFDR probabilities from the joint model for three different $\delta$ values. $^1$ denotes SNPs significantly associated with both 1555476_at and 242261_at.

<table>
<thead>
<tr>
<th>Probe Set</th>
<th>SNP</th>
<th>Position</th>
<th>SimBaS</th>
<th>$P(|\beta(t)| &gt; \delta)$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Joint</td>
<td>Single</td>
</tr>
<tr>
<td>1555476_at</td>
<td>rs7163013</td>
<td>78698759</td>
<td>3.00%</td>
<td>36.0%</td>
</tr>
<tr>
<td></td>
<td>rs8034191$^1$</td>
<td>78806023</td>
<td>0.01%</td>
<td>33.0%</td>
</tr>
<tr>
<td></td>
<td>rs2036527$^1$</td>
<td>78851615</td>
<td>0.01%</td>
<td>25.0%</td>
</tr>
<tr>
<td></td>
<td>rs951266</td>
<td>78878541</td>
<td>0.01%</td>
<td>39.0%</td>
</tr>
<tr>
<td></td>
<td>rs16969968$^1$</td>
<td>78882925</td>
<td>0.01%</td>
<td>50.0%</td>
</tr>
<tr>
<td></td>
<td>rs6495308</td>
<td>78907656</td>
<td>0.01%</td>
<td>2.00%</td>
</tr>
<tr>
<td></td>
<td>rs11639372</td>
<td>78966655</td>
<td>0.30%</td>
<td>50.0%</td>
</tr>
<tr>
<td>214666_x_at</td>
<td>rs1040262</td>
<td>78587910</td>
<td>0.80%</td>
<td>0.01%</td>
</tr>
<tr>
<td></td>
<td>rs2656057</td>
<td>78723082</td>
<td>4.00%</td>
<td>50.0%</td>
</tr>
<tr>
<td></td>
<td>rs2869550</td>
<td>78981001</td>
<td>3.00%</td>
<td>0.01%</td>
</tr>
<tr>
<td></td>
<td>rs3825806</td>
<td>78985342</td>
<td>0.30%</td>
<td>10.0%</td>
</tr>
<tr>
<td>242261_at</td>
<td>rs4283201</td>
<td>78681365</td>
<td>0.01%</td>
<td>4.00%</td>
</tr>
<tr>
<td></td>
<td>rs8034191$^1$</td>
<td>78806023</td>
<td>5.00%</td>
<td>50.0%</td>
</tr>
<tr>
<td></td>
<td>rs2036527$^1$</td>
<td>78851615</td>
<td>0.30%</td>
<td>0.60%</td>
</tr>
<tr>
<td></td>
<td>rs6495306</td>
<td>78865893</td>
<td>4.00%</td>
<td>0.20%</td>
</tr>
<tr>
<td></td>
<td>rs951266</td>
<td>78878541</td>
<td>0.30%</td>
<td>5.00%</td>
</tr>
<tr>
<td></td>
<td>rs16969968$^1$</td>
<td>78882925</td>
<td>0.70%</td>
<td>5.00%</td>
</tr>
</tbody>
</table>

Table 2. Minimum p-values resulting from the standard analysis. The top portion of the table contains p-values corresponding to the model that treats SNP linearly while the bottom portion contains p-values for models that treat SNP categorically. The three adjustments used were the Plug-in FDR, Benjamini-Hochberg FDR, and the Bonferroni. P-values can thus be compared to the nominal $\alpha = 0.05$.

<table>
<thead>
<tr>
<th>Probe Set</th>
<th>Model</th>
<th>Correction</th>
<th>555476_at</th>
<th>214666_x_at</th>
<th>225892_at</th>
<th>242261_at</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Linear</td>
<td>Plug-in FDR</td>
<td>0.219</td>
<td>0.847</td>
<td>0.543</td>
<td>0.085</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Benjamini-Hochberg</td>
<td>0.155</td>
<td>0.930</td>
<td>0.196</td>
<td>0.083</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bonferroni</td>
<td>0.923</td>
<td>1.000</td>
<td>1.000</td>
<td>0.249</td>
</tr>
<tr>
<td></td>
<td>Categorical</td>
<td>Plug-in FDR</td>
<td>0.106</td>
<td>0.396</td>
<td>0.116</td>
<td>0.146</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Benjamini-Hochberg</td>
<td>0.323</td>
<td>0.999</td>
<td>0.385</td>
<td>0.006</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bonferroni</td>
<td>0.679</td>
<td>1.000</td>
<td>1.000</td>
<td>0.006</td>
</tr>
</tbody>
</table>
Appendix A: MCMC Algorithm

Given the prior specifications in Section 2.1.2, we now describe the MCMC algorithm for obtaining estimates of $\beta(t)$. The procedure is a modification of the standard Bayesian Probit regression to accommodate an ordinal and functional outcome and scalar predictor or predictors. The standard
algorithm can vary slightly depending on the approach to estimating the cut points. We begin our procedure by first sampling the latent variable, then updating the cut points, and finally sampling the parameters from the latent variable model. Two additional steps are needed to project the latent variable into the wavelet space and to bring the coefficients back into the data space.

Define $y$ as the vectorized form of $Y$, the matrix of ordinal outcomes and likewise define $y^*$ as the vectorized form of $Y^*$, the matrix form of the latent variable. Further let $\mu^* = \mathbb{E}(y^*)$, the expected value of the vectorized latent variable. Our MCMC algorithm is then

**Step 1:** Update the latent variable $y^*|y, \mu^*, c_1, c_2$ using truncated normals of the form

$$y^*|y = 0, \mu^*, c_1 \sim \mathcal{N}(\mu^*, I)1(y^* \in (-\infty, c_1))$$

$$y^*|y = 1, \mu^*, c_1, c_2 \sim \mathcal{N}(\mu^*, I)1(y^* \in (c_1, c_2))$$

$$y^*|y = 2, \mu^*, c_2 \sim \mathcal{N}(\mu^*, I)1(y^* \in (c_2, \infty))$$

where $I$ denotes the identity matrix and $1(\cdot)$ is the indicator function. Further, $\mu^*$ is taken to be $X\beta_W^{(m)}$ where $\beta_W^{(m)}$ is the $m$th posterior draw of $\beta_W$.

**Step 2:** Update the cut point $c_2|Y^*, Y, c_1$ using the uniform distribution

$$c_2|y^*, y, c_1 \sim \mathcal{U}(a, b)$$

for $a = \max [\max(y^*|y = 1), c_1]$ and $b = \min(y^*|y = 2)$.

**Step 3:** Perform DWT on the rows of the updated latent variable matrix: $Y^* \overset{\text{DWT}}{\rightarrow} Y^*_W$

**Step 4:** Update $\beta_W$ using

$$\beta_{(p,jk)}^W|y_{(j,k)}^W, \beta_{(-p),jk}^W, \Sigma^* \sim \gamma_{p,jk}\mathcal{N}(\mu_{p,jk}, \epsilon_{p,jk}) + (1 - \gamma_{p,jk})d_0$$

Where the mixture probability $\alpha_{p,jk}$ is given by

$$\alpha_{p,jk} = \Pr \left( \gamma_{p,jk} = 1 | y_{(j,k)}^*, \beta_{(-p),jk}^W, \Sigma^* \right) = O_{p,jk} / (O_{p,jk} + 1)$$

for $O_{p,jk} = \pi_{p,j}/(1 - \pi_{p,j})BF_{p,jk}$ and $BF_{p,jk} = (1 + \tau_{p,jk}/V_{p,jk})^{-1/2} \exp \left\{ \frac{1}{2}\epsilon_{p,jk}^2 \right\}$

and $\mu_{p,jk} = \hat{\beta}_{(p,jk),\text{MLE}}^W(1 + V_{p,jk}/\tau_{p,jk})^{-1}$ and $\epsilon_{p,jk} = V_{p,jk}(1 + V_{p,jk}/\tau_{p,jk})^{-1}$. Both $\hat{\beta}_{(p,jk),\text{MLE}}^W$
and \( V_{p,jk} \) are initial values taken from a maximum likelihood estimation of the latent variable model.

**Step 5:** Update \( \tau_{pj} \) and \( \pi_{pj} \) using

\[
\tau_{pj} | a_\tau, b_\tau, \gamma_{p,jk}, \beta^W_{(p,jk)} \sim IG \left( a_\tau + \frac{1}{2} \gamma_{p,jk}, b_\tau + \frac{1}{2} \gamma_{p,jk} \{ \beta^W_{(p,jk)} \}^2 \right)
\]

\[
\pi_{pj} | a_\pi, b_\pi, \gamma_{p,jk} \sim Beta \left( a_\pi + \gamma_{p,jk}, b_\pi + \gamma_{p,jk} \right)
\]

for \( a_\tau, b_\tau, a_\pi, \) and \( b_\pi \) fixed and estimated via an Empirical Bayes approach described in Morris and Carroll (2006) based on the initial latent outcome.

**Step 6:** Project \( \beta^W \) into the data space using the inverse DWT, \( \beta = \beta^W W \). Because the algorithm involves the projection of \( \beta^W \) back into the data space, post processing only involves the calculation of summary measures based on the posterior samples and statistics to perform inference.

**Appendix B: Extension to Generalized Function-on-Function Regression**

Model (2.5) in the Manuscript allows \( X \) be of size \( N \times P \) which, in function-on-scalar regression, allows for any number of scalar covariates. In fact, we can let \( P \) get relatively large with respect to \( T \), the total number of measurements observed for \( Y_i(t) \). In other words, we can let \( X \) become \( X(v) \) for the grid \( v = 1, \ldots, V \) which is to say the formulation allows for the inclusion of a functional covariate where the values of \( T \) and \( V \) need not be equal. Further, \( t \) and \( v \) only index measurement occurrence and thus do not necessarily have to represent the same grid. In other words, \( Y_i(t) \) may be sampled more finely than \( X_i(v) \) or vice versa. We now present the Ordinal Probit Wavelet-Based Function-on-Function Regression (OPWAVFR) as an extension of the OPWAVFM.

Formulating the OPWAVFR only requires minor alterations the model describe above. The latency assumption presented in Model (2.1) of the Manuscript holds as does the probability
described in Model (2.2). To incorporate subject $i$’s functional covariate $X_i(v)$, we re-express Model (2.3) of the Manuscript as

$$Y_i^*(t) = \int_V X_i(v) \eta(v, t) dv + E_i(t)$$

(0.1)

where $V$ is the support of $X_i(v)$. We are now interested in the estimation of the surface $\eta(v, t)$. Assuming the errors are Gaussian once again gives us the Probit model with probabilities for subject $i$ at a fixed $t$ given by

$$P(Y_i(t) = g) = \Phi \left( c_{g+1} - \int_V X_i(v) \eta(v, t) dv \right) - \Phi \left( c_g - \int_V X_i(v) \eta(v, t) dv \right)$$

for $g = 0, 1, 2$. The discretized form of Model (0.1) is $Y^* = X\eta + E$ where $Y^*$ and $E$ are $N \times T$, $X$ is $N \times V$, and $\eta$ is $V \times T$.

The MCMC algorithm described in Appendix A can accommodate an $X$ design matrix of reasonable size, however computational burden increases as $P$ increases. Thus in the function-on-function model we must either limit the size of $V$ or perform data reduction. Meyer and others (2015) explores a function-on-function regression for hierarchical data with a Gaussian Process outcome. They suggest the use of Wavelet-Principal Components (wPC) for decomposing a functional covariate when data reduction is needed. The formulation of the latent variable in the OPWAVFFR then follows the procedure described in Meyer and others (2015) for the case where only a single set of curves, $\{Y_i(t), X_i(t)\}$ is observed on each subject as opposed to multiple. Modeling of the OPWAVFFR can easily take place in the OPWAVFM context given a design matrix $X$ containing not a set of scalar covariates, but measurements from a functional covariate projected into a desired space.

Denote the functional covariate as $X$ and decompose it using a DWT, $X = X^W W X$. Next decompose $X^W$ using a singular value decomposition, $X^W = X^W P_{svd}$ where $P_{svd}$ is the matrix of right singular vectors. The function-on-function representation of the wavelet-space model in
Model (2.6) in the Manuscript is

\[ Y^W = X^W W X P_{svd} P_{svd}' W X' \eta^P + E^W W \]  

(0.2)
given the decomposition \( \eta = P_{svd}' W X' \eta^P W \). Note that since \( W X \) and \( P_S \) are orthogonal,

\[ W X P_{svd} P_{svd}' W X' = I_P \]

where \( P \) is number of columns of \( X \). Thus after post-multiplying by \( W' \), Model (0.2) reduces to

\[ Y^W = X^W \eta^P + E^W \]
or essentially Model (2.7) with \( X^W \) replacing \( X \) and \( \eta^P \) replacing \( \beta \).

Alterations to Appendix A and Section 2.2 of the Manuscript are minor in order to implement the OPWAVFFR. For instance, the MCMC procedure remains the same with the exception of Step 3 which now involves not only the inverse DWT for the \( Y \) wavelet space but also the inverse of the transformation used for \( X_i(v) \). Additionally, instead of performing inference on \( \beta(t) \), Section 2.2 of the Manuscript can be modified to accommodate \( \eta(v, t) \). For a detailed description of the formulation of the BFDR and SimBaS for a surface of coefficients, see Meyer and others (2015).
**APPENDIX C: ADDITIONAL SIMULATION RESULTS**

*Estimation and Bias*

Fig. 1. Posterior estimates as a function of $t$ averaged over 200 simulated data sets for all simulation settings. Light gray bands depict the 2.5th to 97.5th percentiles across the simulated data sets while true functions are in solid dark gray. The top two rows contain averaged estimates from the single covariate scenarios. The bottom row contains averaged estimates from the two covariate scenarios.
Fig. 2. The figure on the left contains the average point-wise bias of the estimated \( \beta(t) \) as a function of \( t \) taken across 200 simulated data sets. The figure on the right compares box plots across scenarios of rMSE calculated for each simulated data set.

Fig. 3. Average point-wise bias of \( \beta_1(t) \) and \( \beta_2(t) \) as a function of \( t \) averaged over 200 simulated data sets for two covariates of increasing correlation. The left figure contains point-wise bias for \( \beta_1(t) \) while the right column contains point-wise bias for \( \beta_2(t) \).
Inference

Metrics Key  Define false discovery rate, \( \text{FDR}_\epsilon \), as the number of selected locations \((v, t)\) with true value \( \leq \epsilon \) divided by the total number of selected locations. Next define the sensitivity, \( \text{SEN}_\Upsilon \), as the number of selected locations \((v, t)\) with true magnitude \( > \Upsilon \) divided by the total number of locations with true magnitude \( > \Upsilon \). \( \text{EWER}_\epsilon \) is calculated as the proportion of simulated datasets with at least one falsely discovered location, i.e. a selected location with true value \( \leq \epsilon \).

Fig. 4. Plotted \( \text{SEN}_\Upsilon \), \( \text{FDR}_\epsilon \), and \( \text{EWER}_\epsilon \) for the Plug-in FDR and Benjamini-Hochberg FDR using the standard approach and SimBaS and the Bayesian FDR with \( \delta = 0.2 \) using the OPWAVFM. The single covariate setting with the single normal effect is presented here.
Fig. 5. Plotted SEN$_\tau$, FDR$_\epsilon$, and EWER$_\epsilon$ for the Plug-in FDR and Benjamini-Hochberg FDR using the standard approach and SimBaS and the Bayesian FDR with $\delta = 0.2$ using the OPWAVFM. The single covariate setting with the single triangle effect is presented here.
Fig. 6. Plotted SEN, FDR, and EWER for the Plug-in FDR and Benjamini-Hochberg FDR using the standard approach and SimBaS and the Bayesian FDR with \( \delta = 0.2 \) using the OPWAIFM. The single covariate setting with the double triangle effect is presented here.
Fig. 7. Plotted $\text{SEN}_\gamma$, $\text{FDR}_\epsilon$, and $\text{EWER}_\epsilon$ for the Plug-in FDR and Benjamini-Hochberg FDR using the standard approach and SimBaS and the Bayesian FDR with $\delta = 0.2$ using the OPWAVFM. The two covariate setting with no correlation is presented here. The right column contains figures for $\beta_1$ while the left column contains figures for $\beta_2$. 
Fig. 8. Plotted SEN, FDR, and EWER, for the Plug-in FDR and Benjamini-Hochberg FDR using the standard approach and SimBaS and the Bayesian FDR with \( \delta = 0.2 \) using the OPWAVFM. The two covariate setting with moderate correlation is presented here. The right column contains figures for \( \beta_1 \) while the left column contains figures for \( \beta_2 \).
Fig. 9. Test attributes of the single covariate simulations for the single peak scenarios. Left column: heat maps of the false discovery rate for a given cut-point, FDR_\epsilon, as functions of \epsilon for the BFDR with varying levels of \pi_\delta. Right column: heat maps of the sensitivity to detect a magnitude of \Upsilon, SEN_\Upsilon, as functions of \Upsilon for the BFDR with varying levels of \pi_\delta. For both, darker indicates lower false discovery and sensitivity while lighter indicates higher.
Fig. 10. Test attributes of the single covariate simulations for the double peak scenarios. Left column: heat maps of the false discovery rate for a given cut-point, \( FDR_\varepsilon \), as functions of \( \varepsilon \) for the BFDR with varying levels of \( \pi_3 \). Right column: heat maps of the sensitivity to detect a magnitude of \( \Upsilon \), \( SEN_\Upsilon \), as functions of \( \Upsilon \) for the BFDR with varying levels of \( \pi_3 \). For both, darker indicates lower false discovery and sensitivity while lighter indicates higher.
In Section 4 of the Manuscript, we mention the ability of the OPWAVFM to scale up to a large number of SNPs. Here we present the results from interrogating SNPs within 2Mb of IREB2. Figure 11 contains the SimBa Scores from examining a single probe set at a time. Similar to the smaller interrogation describe in Section 4 of the Manuscript, probe set 225892_at exhibited no significant SNPs so we’ve excluded these results. Thus the joint model for this analysis also excludes that probe set.

Fig. 11. Single probe models SimBa Scores plotted as functions of position on the chromosome. The location of IREB2 is noted as a horizontal bar below the probabilities. For convenience, a dotted-dashed gray line depicts a global $\alpha$-level of 0.05 plotted on the $−\log_{10}$ scale. Scores are from the models taking 2Mb to either side of IREB2.
Fig. 12. Joint model SimBa Scores for the probe sets 1555476_at, 214666_x_at, and 242261_at plotted as functions of position on the chromosome. The location of IREB2 is noted as a horizontal bar below the probabilities. For convenience, a dotted-dashed gray line depicts a global $\alpha$-level of 0.05 plotted on the $-\log_{10}$ scale. Scores are from the models taking 2Mb to either side of IREB2.

REFERENCES


[Received October, 2015; revised ?; accepted for publication ?]