Supplementary Materials for “Adaptive, Robust Functional Regression in Functional Mixed Model Framework”

Hongxiao Zhu, Philip J. Brown and Jeffrey S. Morris

February 2, 2011

Contents

1 MCMC Algorithm **2**

1.1 Details of MCMC Algorithm 2

1.2 Some Derivations for the MCMC Algorithm 4

2 Initial Values and Empirical Bayes Parameters **6**

2.1 Initial Values for MCMC 6

2.2 Empirical Bayes Parameters for Gamma 6

2.3 A Sensitivity Study 7

3 Further Explanation of Bayesian FDR **7**

4 Extra Simulation Results **12**

4.1 Extra Results for the Main Simulation 12

4.2 Details for the Extra Simulation 17

5 Extra Real Data Application Results **20**

5.1 More Results Using Bayesian FDR 20

5.2 Results for Outlier Detection 20

6 The Gamma Mixture of DE prior **21**

7 The Exponential-Gamma Mixing Distribution **21**

8 Proof of Robustness as One Curve Approaches Infinity **22**

1. MCMC ALGORITHM

1.1 Details of MCMC Algorithm

We perform a Markov Chain Monte Carlo algorithm to draw samples from the posterior of the wavelet-space model parameters, to which the IDWT can be applied to obtain estimates of the corresponding parameters in the data-space model. The following are the details of the MCMC.

Step 0. Initialize $\{\nu_{jk}^E\}, \{\nu_{jk}^U\}, \{\nu_{jk}^B\}$ and $\{\lambda_{ijk}\}, \{\phi_{bjk}\}, \{\psi_{ajk}\}$ based on automatic MLE estimation and set up prior parameters.

Step 1. For each j, k, rescale the (j,k)th column of model (3) in the paper by premultipling by $\mathbf{\Lambda}_{jk}^{-1/2} = \text{diag}\{\lambda_{ijk}\}_{i=1}^n$, to obtain

$$\mathbf{d}_{jk}^+ = \mathbf{X}_{jk}^+ \mathbf{b}_{jk}^* + \mathbf{Z}_{jk}^+ \mathbf{u}_{jk}^* + \mathbf{e}_{jk}^+,$$

where $\mathbf{d}_{jk}^+ = \mathbf{\Lambda}_{jk}^{-1/2}\mathbf{d}_{jk}$, $\mathbf{X}_{jk}^+ = \mathbf{\Lambda}_{jk}^{-1/2}\mathbf{X}$, and $\mathbf{Z}_{jk}^+ = \mathbf{\Lambda}_{jk}^{-1/2}\mathbf{Z}$, and $\mathbf{E}_{jk}^+ = \mathbf{\Lambda}_{jk}^{-1/2}\mathbf{e}_{jk}^*$ are weighted versions of the data and design matrices for wavelet coefficient (j, k). Performance of this rescaling up front simplifies and speeds calculations in the later steps. We can see that $\mathbf{d}_{jk}^+|\mathbf{b}_{jk}^*, \mathbf{\Sigma}_{jk}^+ \sim N(\mathbf{X}_{jk}^+\mathbf{b}_{jk}^*, \mathbf{\Sigma}_{jk}^+)$, with $\mathbf{\Sigma}_{jk}^+ = \mathbf{Z}_{jk}^+\mathbf{\Phi}_{jk}(\mathbf{Z}_{jk}^+)^T + \mathbf{I}_n$, where $\mathbf{\Phi}_{jk}$=diag$(\phi_{bjk})_b$.

Step 2. For each a, j, k, update B_{ajk}^* from $f(B_{ajk}^*|B_{(-a)jk}^*, \boldsymbol{\lambda}_{jk}, \boldsymbol{\phi}_{jk}, \psi_{ajk}, \mathbf{d}_{jk})$, where $\boldsymbol{\lambda}_{jk} = \{\lambda_{ijk}\}_{i=1}^n$, $\boldsymbol{\phi}_{jk} = \{\phi_{bjk}\}_{b=1}^m$, and $\mathbf{d}_{jk} = \{d_{ijk}\}_{i=1}^n$. This distribution is a mixture of a point mass at zero and a Gaussian distribution, with the Gaussian probability given by $\alpha_{ajk} = \Pr\{\gamma_{ajk} = 1|\mathbf{B}_{(-a)jk}^*, \boldsymbol{\lambda}_{jk}, \boldsymbol{\phi}_{jk}, \psi_{ajk}, \pi_{aj}, \mathbf{d}_{jk}\} = \Pr\{\gamma_{ajk} = 1|\mathbf{d}_{jk}^+, \mathbf{B}_{(-a)jk}^*, \mathbf{\Sigma}_{jk}^+, \psi_{ajk}, \pi_{aj}\}$, which can be obtained from the conditional odds ratio:

$$\frac{\Pr\{\gamma_{ajk} = 1|\mathbf{d}_{jk}^+, \mathbf{B}_{(-a)jk}^*, \mathbf{\Sigma}_{jk}^+, \psi_{ajk}, \pi_{aj}\}}{\Pr\{\gamma_{ajk} = 0|\mathbf{d}_{jk}^+, \mathbf{B}_{(-a)jk}^*, \mathbf{\Sigma}_{jk}^+ \psi_{ajk}, \pi_{aj}\}} = \text{Conditional Bayes Factor} \times \text{Prior Odds}.$$

The prior odds is given by $\pi_{aj}/(1-\pi_{aj})$, and the conditional Bayes factor is $(1+\psi_{ajk}/V_{ajk})^{-1/2}\exp\{\zeta_{ajk}^2(1+V_{ajk}/\psi_{ajk})^{-1}/2\}$, with $V_{ajk} = [\{\mathbf{X}_{ajk}^+\}^T(\mathbf{\Sigma}_{jk}^+)^{-1}\mathbf{X}_{ajk}^+]^{-1}$,

$\zeta_{ajk} = \hat{B}^*_{ajk}/\sqrt{V_{ajk}}$, and $\hat{B}^*_{ajk} = V_{ajk}\{\mathbf{X}^+_{ajk}\}^T(\mathbf{\Sigma}^+_{jk})^{-1}\{\mathbf{d}^+_{jk} - \mathbf{X}^+_{(-a)jk}\mathbf{B}^*_{(-a)jk}\}$. $\mathbf{X}^+_{ajk}$ represents the a^{th} column of $\mathbf{X}^+_{jk}$ and $\mathbf{X}^+_{(-a)jk}$ is $\mathbf{X}^+_{jk}$ with the a^{th} column removed. Drawing $\gamma_{ajk} \sim$ Bernoulli(α_{ajk}), if $\gamma_{ajk} = 0$ we set $B^*_{ajk} = 0$. Otherwise, if $\gamma_{ajk} = 1$, we draw B^*_{ajk} from $N(\mu_{B^*_{ajk}}, V_{B^*_{ajk}})$, where $\mu_{B^*_{ajk}} = \hat{B}^*_{ajk}(1 + V_{ajk}/\psi_{aj})^{-1}$ and $V_{B^*_{ajk}} = V_{ajk}(1 + V_{ajk}/\psi_{aj})^{-1}$.

Take note of the form of $\hat{B}^*_{ajk}$, which is involved in $\mu_{B^*_{ajk}}$, the conditional mean when $\gamma_{ajk} = 1$. We see from the X^+_{jk} (involving λ_{ijk}) that observations with outlying residuals are down-weighted. From the expression of Σ^+_{jk} (involving ϕ_{bjk}), we see that observations linked to outlying random effect units are also down-weighted, since the b with large ϕ_{bjk} have larger contributions to the variance $\mathbf{\Sigma}^+_{jk}$, and thus are down-weighted by the term $(\mathbf{\Sigma}^+_{jk})^{-1}$ in $\hat{B}^*_{ajk}$.

Also, note that this update step was done while integrating out the random effects, which we have found leads to an improved sampler.

Step 3. Update $\mathbf{u}^*_{jk}$ from $f(\mathbf{u}^*_{jk}|\mathbf{b}^*_{jk}, \boldsymbol{\lambda}_{jk}, \boldsymbol{\phi}_{jk}, \mathbf{d}_{jk})$, which is given by $N(\boldsymbol{\mu}_{u^*_{jk}}, \mathbf{V}_{u^*_{jk}})$, where $\boldsymbol{\mu}_{u^*_{jk}} = \{(\mathbf{Z}^+_{jk})^T\mathbf{Z}^+_{jk} + \mathbf{\Phi}^{-1}_{jk}\}^{-1}(\mathbf{Z}^+_{jk})^T(\mathbf{d}^+_{jk} - \mathbf{X}^+_{jk}\mathbf{B}^*_{jk})$ and $\mathbf{V}_{u^*_{jk}} = \{(\mathbf{Z}^+_{jk})^T\mathbf{Z}^+_{jk} + \mathbf{\Phi}^{-1}_{jk}\}^{-1}$, with $\mathbf{\Phi}_{jk} = \text{diag}\{\phi_{bjk}\}^m_{b=1}$. Note from the conditional mean $\boldsymbol{\mu}_{u^*_{jk}}$ that the λ_{ijk} implicit in Z^+_{jk} act as weights on the observations, down-weighting the influence of outliers, and ϕ_{bjk} act as prior variances leading to nonlinear shrinkage of $\mathbf{u}^*_{jk}$, with wavelet coefficients with larger random effect magnitudes tending to have larger prior variances, and thus less shrinkage.

Step 4. Conditional on $\mathbf{b}^*_{jk}$, $\mathbf{u}^*_{jk}$ and the lasso parameters $\{\nu^E_{jk}\}, \{\nu^U_{jk}\}, \{\nu^B_{jk}\}$, update the scaling parameters $\{\lambda_{ijk}\}_i$, $\{\phi_{bjk}\}_b$ and $\{\psi_{ajk}\}_a$ from their complete conditional distributions. We credit Park & Casella (2008) with demonstrating that the complete conditional of the inverse of a scaling parameter in the Bayesian lasso model has a closed form expression as an inverse Gaussian distribution. Based on those results, similar calculations in our setting reveal that the complete conditional distribution of the inverse of all scaling parameters in the R-FMM are also inverse Gaussians, specified as follows.

$$\begin{aligned}
\lambda_{ijk}^{-1}|d_{ijk}, \mathbf{b}_{jk}^*, \mathbf{u}_{jk}^*, \nu_{jk}^E &\sim \text{Inv-Gauss}\{\sqrt{(\nu_{jk}^E)^2/(d_{ijk} - \mathbf{X}_i^T\mathbf{b}_{jk}^* - \mathbf{Z}_i^T\mathbf{u}_{jk}^*)^2}, (\nu_{jk}^E)^2\}, \\
\phi_{bjk}^{-1}|U_{bjk}^*, \nu_{jk}^U &\sim \text{Inv-Gauss}\{\sqrt{(\nu_{jk}^U)^2/(U_{bjk}^*)^2}, (\nu_{jk}^U)^2\}, \\
(\psi_{ajk}^{-1}|B_{ajk}^*, \nu_{jk}^B, \gamma_{ajk} = 1) &\sim \text{Inv-Gauss}\{\sqrt{(\nu_{jk}^B)^2/(B_{ajk}^*)^2}, (\nu_{jk}^B)^2\}, \\
(\psi_{ajk}|B_{ajk}^*, \nu_{jk}^B, \gamma_{ajk} = 0) &\sim \text{Exp}((\nu_{jk}^B)^2/2).
\end{aligned}$$

Here $\mathbf{X}_i^T$ and $\mathbf{Z}_i^T$ are the i^{th} rows of the design matrices $\mathbf{X}$ and $\mathbf{Z}$, respectively. Note that in the final row above, when $\gamma_{ajk} = 0$, the Gibbs update step for ψ_{ajk} amounts to sampling from the mixing distribution, since in that state of the model the distribution is independent of the data conditional on ν_{aj}^{ψ}.

Step 5. Update the lasso parameters $\{\nu_{jk}^E\}$, $\{\nu_{jk}^U\}$, $\{\nu_{jk}^B\}$ from their complete conditional distributions. Their squared values are conjugate gammas, i.e., $(\nu_{jk}^E)^2|\{\lambda_{ijk}\}_i \sim \text{Gamma}(n+a^E, \sum_{i=1}^n \lambda_{ijk}/2+b^E)$, $(\nu_{jk}^U)^2|\{\phi_{bjk}\}_b \sim \text{Gamma}(m+a^U, \sum_{b=1}^m \phi_{bjk}/2+b^U)$, and$(\nu_{jk}^B)^2|\{\psi_{ajk}\}_a \sim \text{Gamma}(K_j + a^B, \sum_{k=1}^{K_j} \psi_{ajk}/2 + b^B)$, where K_j is the number of wavelet coefficients at resolution level j.

Step 6. For each a, j, update $\pi_{aj}|\{\gamma_{ajk}\}_k \sim \text{Beta}(\sum_k \gamma_{ajk} + a^\pi, K_j - \sum_k \gamma_{ajk} + b^\pi)$.

Repeat Steps 1-6 until reaching a pre-specified maximum number of iterations.

1.2 Some Derivations for the MCMC Algorithm

(1) **The conditional Bayes factor in Step 2.**

From Step 2, it is easy to see that the conditional odds can be written as:

$$\frac{f(\gamma_{ajk} = 1|\mathbf{d}_{jk}^+, \mathbf{B}_{(-a)jk}^*, \mathbf{\Sigma}_{jk}^+(\boldsymbol{\theta}_{jk}), \boldsymbol{\nu}^2)}{f(\gamma_{ajk} = 0|\mathbf{d}_{jk}^+, \mathbf{B}_{(-a)jk}^*, \mathbf{\Sigma}_{jk}^+(\boldsymbol{\theta}_{jk}), \boldsymbol{\nu}^2)} = \frac{f(\mathbf{d}_{jk}^+|\gamma_{ajk} = 1, \mathbf{B}_{(-a)jk}^*, \mathbf{\Sigma}_{jk}^+(\boldsymbol{\theta}_{jk}), \boldsymbol{\nu}^2)}{f(\mathbf{d}_{jk}^+|\gamma_{ajk} = 0, \mathbf{B}_{(-a)jk}^*, \mathbf{\Sigma}_{jk}^+(\boldsymbol{\theta}_{jk}), \boldsymbol{\nu}^2)} \cdot \frac{f(\gamma_{ajk} = 1)}{f(\gamma_{ajk} = 0)}$$

$$= \text{Conditional Bayes Factor} \times \text{Prior Odds},$$

in which, $f(\mathbf{d}_{jk}^+|\gamma_{ajk} = 0, \mathbf{B}_{(-a)jk}^*, \mathbf{\Sigma}_{jk}^+(\boldsymbol{\theta}_{jk}), \boldsymbol{\nu}^2) \propto |\mathbf{\Sigma}_{jk}^+|^{-1/2} \exp\{-\frac{1}{2}(\tilde{\mathbf{d}}_{jk}^+)^T(\mathbf{\Sigma}_{jk}^+)^{-1}\tilde{\mathbf{d}}_{jk}^+\}$, where $\tilde{\mathbf{d}}_{jk}^+ = \mathbf{d}_{jk}^+ - \mathbf{X}_{(-a)jk}^+\mathbf{B}_{(-a)jk}^*$. In the numerator of the conditional Bayes factor,

$$
\begin{aligned}
& f(\mathbf{d}_{jk}^{+}|\gamma_{ajk}=1,\mathbf{B}_{(-a)jk}^{*},\boldsymbol{\Sigma}_{jk}^{+}(\boldsymbol{\theta}_{jk}),\boldsymbol{\nu}^{2}) \\
= & \int f(\mathbf{d}_{jk}^{+}|\mathbf{b}_{jk}^{*},\boldsymbol{\Sigma}_{jk}^{+}(\boldsymbol{\theta}_{jk}),\boldsymbol{\nu}^{2}) f(B_{ajk}^{*}|\gamma_{ajk}=1,\psi_{ajk})dB_{ajk}^{*} \\
\propto & \ |\boldsymbol{\Sigma}_{jk}^{+}|^{-1/2}\psi_{ajk}^{-1/2}\tilde{K}^{-1/2}\exp\left\{-\frac{1}{2}(\tilde{\mathbf{d}}_{jk}^{+})^{T}\left[(\boldsymbol{\Sigma}_{jk}^{+})^{-1}-(\boldsymbol{\Sigma}_{jk}^{+})^{-1}\mathbf{X}_{(a)jk}^{+}(\mathbf{X}_{(a)jk}^{+})^{T}(\boldsymbol{\Sigma}_{jk}^{+})^{-1}/\tilde{K}\right]\tilde{\mathbf{d}}_{jk}^{+}\right\},
\end{aligned}
$$

where $\tilde{K} = (\mathbf{X}_{(a)jk}^{+})^{T}(\boldsymbol{\Sigma}_{jk}^{+})^{-1}\mathbf{X}_{(a)jk}^{+} + \psi_{ajk}^{-1} = V_{ajk}^{-1} + \psi_{ajk}^{-1}$, and $\mathbf{X}_{(a)jk}^{+}$ is the ath column of $\mathbf{X}_{jk}^{+}$. Based on this, the conditional Bayes factor can be further simplified to

$$
(1+\psi_{ajk}/V_{ajk})^{-1/2}\exp\left\{\zeta_{ajk}^{2}(1+V_{ajk}/\psi_{ajk})^{-1}/2\right\}.
$$

(2) **The Inverse-Gaussian distributions for updating scaling parameters in Step 4.**

The standard inverse Gaussian distribution (with mean μ, variance μ^3/s) has density $f(x|\mu,s) = [s/(2\pi x^3)]^{1/2}\exp\{-s(x-\mu)^2/(2\mu^2 x)\}$. Consider the scaling parameter λ_{ijk}:

$$
\begin{aligned}
& f(\lambda_{ijk}|d_{ijk},B_{jk}^{*},U_{jk}^{*},(\nu_{jk}^{E})^{2}) \propto f(d_{ijk}|B_{jk}^{*},U_{jk}^{*},\lambda_{ijk})f(\lambda_{ijk}|(\nu_{jk}^{E})^{2}) \\
\propto & (\lambda_{ijk})^{-1/2}\exp\{-\frac{1}{\lambda_{ijk}}[\frac{1}{2}(d_{ijk}-\mathbf{X}_{i}^{T}\mathbf{B}_{jk}^{*}-\mathbf{Z}_{i}^{T}\mathbf{U}_{jk}^{*})^{2}]-\frac{\lambda_{ijk}}{2/(\nu_{jk}^{E})^{2}}\},
\end{aligned}
$$

where $\mathbf{X}_i^T$ denote the ith row of $\mathbf{X}$, and $\mathbf{Z}_i^T$ denote the ith row of $\mathbf{Z}$. Using transformations from λ_{ijk} to λ_{ijk}^{-1}, we find that $(\lambda_{ijk}^{-1}|d_{ijk},B_{jk}^{*},U_{jk}^{*},(\nu_{jk}^{E})^{2})$ is distributed as Inv-Gauss $(\sqrt{\frac{(\nu_{jk}^{E})^{2}}{(d_{ijk}-X_{i}^{T}B_{jk}^{*}-Z_{i}^{T}U_{jk}^{*})^{2}}},(\nu_{jk}^{E})^{2})$. Similar derivation gives the conditional inverse Gaussian distributions for ϕ_{bjk}^{-1} and ψ_{ajk}^{-1}.

(3) **In Step 4, the updating of ψ_{ajk} conditional on $\gamma_{ajk}=0$ needs extra attention.**

According to our prior assumption,

$$
f(\psi_{ajk}|B_{ajk}^{*},(\nu_{jk}^{B})^{2},\gamma_{ajk}=0) \propto f(B_{ajk}^{*}|\gamma_{ajk}=0,\psi_{ajk})f(\psi_{ajk}|(\nu_{jk}^{B})^{2}) = \delta_{0}(B_{ajk}^{*})\cdot f(\psi_{ajk}|(\nu_{jk}^{B})^{2}).
$$

Therefore we have $f(\psi_{ajk}|B_{ajk}^{*}=0,(\nu_{jk}^{B})^{2},\gamma_{ajk}=0) = f(\psi_{ajk}|(\nu_{jk}^{B})^{2})$, which is $\mathrm{Exp}((\nu_{jk}^{B})^{2}/2)$. Since ψ_{ajk} is a prior parameter for the case $\gamma_{ajk}=1$, the above step is only for the purpose of forming a strict Gibbs sampler. Here we have assumed that the prior for ψ_{ajk} is independent of γ_{ajk}, i.e. $f(\psi_{ajk}|(\nu^B_{jk})^2, \gamma_{ajk} = 1) = f(\psi_{ajk}|(\nu^B_{jk})^2, \gamma_{ajk} = 0) = f(\psi_{ajk}|(\nu^B_{jk})^2)$. Similar issues are discussed by Carlin & Chib (1995); they call $f(\psi_{ajk}|B^*_{ajk}, (\nu^B_{jk})^2, \gamma_{ajk} = 0)$ the "pseudo-prior." Our simulations and real data application show acceptable mixing for $\{\psi_{ajk}\}$ and the related parameters.

2. INITIAL VALUES AND EMPIRICAL BAYES PARAMETERS

2.1 Initial Values for MCMC

In the MCMC algorithm, we compute initial values based Henderson's mixed model equations on pages 275-286 of Searle et al. (1992). In particular, we first obtain maximum likelihood estimates (MLE) for the variance components, fixed effect and random effects in model (3), assuming uncorrelated Gaussian distributions for $\mathbf{u}^*_{jk}$ and $\mathbf{e}^*_{jk}$, i.e., $\text{Var}(\mathbf{u}^*_{jk}) = \sigma^2_{jk}\mathbf{I}_m$, $\text{Var}(\mathbf{e}^*_{jk}) = (\sigma^0_{jk})^2\mathbf{I}_n$, $\text{Cov}(\mathbf{u}^*_{jk}, \mathbf{e}^*_{jk}) = 0$. We then initialize the lasso parameters by matching the mean of the exponential mixing distributions with the estimated variance components $\hat{\sigma}^2_{jk}$, $(\hat{\sigma}^0_{jk})^2$ and the estimated variance of B^*_{ajk}. The initial values for the scaling parameters are then obtained by sampling from the distributions specified in Step 4, conditional on other initial values.

2.2 Empirical Bayes Parameters for Gamma

The values of the Gamma hyper-parameters (a^E, b^E), (a^U, b^U), (a^B, b^B) are determined by letting the mode of the Gamma distributions equal to the averaged initial estimators of $\{(\nu^E_{jk})^2\}$, $\{(\nu^U_{jk})^2\}$, $\{(\nu^B_{jk})^2\}$ respectively while controlling the variance to be large, such as 10^3. The initial estimators of $\{(\nu^E_{jk})^2\}$, $\{(\nu^U_{jk})^2\}$, $\{(\nu^B_{jk})^2\}$ are obtained by match the mean (first moment) of the exponential prior with the MLE estimator of the variance components, as described in 2.1. The Beta hyper-parameters for $\{\pi_{aj}\}$ were chosen in a similar way by matching the mode of Beta to the averaged initial estimates $\{\pi_{aj}\}$ while control the variance to be fairly large. The initial estimates $\{\pi_{aj}\}$ were computed from the conditional odds in Step 2 while plugging in the initial

MLE estimates.

To guarantee the stability of the algorithm, some numerical constraints are added when determining the initial values. Since the ν values control the variance of the DE distributions, which, if too large, will result in too small variance of DE, which is not what we want to see in the prior (and likelihood). Therefore we add extra constraints that all estimated ν initial values are scaled by 10 (i.e. divided by 10) and they are all bounded above by .1. This will guarantee that all initial values of ν are small enough. Accordingly the estimated initial values of (a, b) enables a Gamma prior with mode less than .01, variance around 10^3. Note that for a Gamma(a,b) distribution with the mode exactly at .01 and variance exactly at 10^3, the corresponding (a,b) values will be $a = 1.0003$, $b = 0.0316$, with 95% confidence interval $(0, 94.8]$.

2.3 A Sensitivity Study

To numerically test whether the posterior estimates of our Robust FMM model are sensitive to the pre-specified variance of Gamma(a, b), we re-run our real data analysis at 4 different variances for initial values of (a, b), i.e., $\text{Var}(\nu^2) = (500, 1000, 5000, 10000)$. Figure 1 shows the estimated $B_1(t)$ and its 95% credible intervals at each run. We can not see any significant difference between these posterior estimates. Table 1 show the integrated squared posterior mean of $B_a(t)$,$a = 1, ..., 5$ and $U_b(t)$, $b = 1, \ldots, 5$. No significant difference or pattern is found.

3. FURTHER EXPLANATION OF BAYESIAN FDR

Once we apply the MCMC, we are left with posterior samples of the data-space model parameters that we can use to perform Bayesian inference. One common task is to find regions of $\mathcal{T}$ for which $B_a(t)$ or some function of the $B_a(t)$ is significantly nonzero.

Suppose we model the $\log_2$ scale, and are interested in finding regions of $\mathcal{T}$ for which there is at least a δ-fold difference in the mean intensity. From the MCMC procedure, suppose we have G posterior samples of the corresponding fixed effect function $\mathbf{B}(t) = [B(t_1), B(t_2), \ldots, B(t_T)]$ on the $\log_2$ scale, denoted by $\{\mathbf{B}^{(g)}(t), g = 1, \ldots, G\}$.

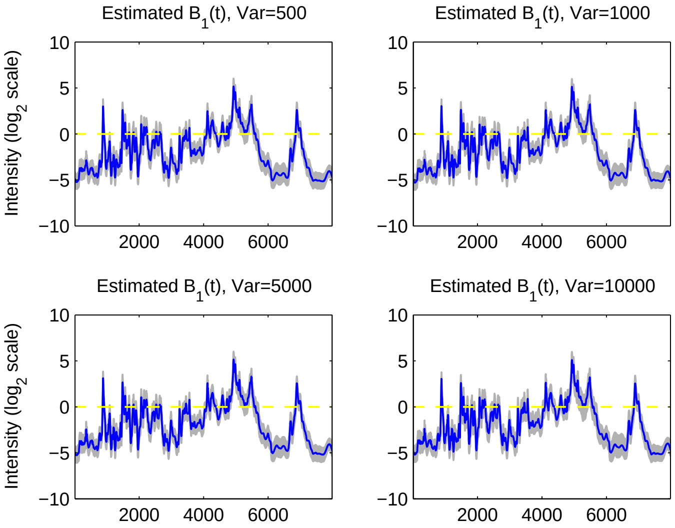

Figure 1: The posterior mean of $B_1(t)$ under different initial setup.

	Variance of Gamma			
	500	1000	5000	10000
$B_1(t)$	266.20	266.35	266.72	266.14
$B_2(t)$	259.44	259.60	260.00	259.46
$B_3(t)$	260.39	260.09	259.97	260.37
$B_4(t)$	267.69	267.72	267.86	267.81
$B_5(t)$	18.98	18.96	18.62	18.57
$U_1(t)$	1.82	1.94	1.88	1.81
$U_2(t)$	1.62	1.51	1.46	1.55
$U_3(t)$	3.10	3.06	2.91	2.81
$U_4(t)$	1.31	1.31	1.44	1.33
$U_5(t)$	2.68	2.89	2.66	2.76

Table 1: Integrated squared posterior mean of $B_a(t)$, $a = 1, ..., 5$ and $U_b(t)$, $b = 1, \ldots, 5$ under difference choices of variance for Gamma(a, b) priors.

From these, we can estimate the point-wise posterior probabilities of at least δ-fold intensity changes at each spectral location as $p(t_l) = \Pr\{|B(t_l)| \geq \log_2(\delta)|\text{Data}\} \approx G^{-1}\sum_g I\{|B^{(g)}(t_l)| \geq \log_2(\delta)\}$ for all $t_l, l = 1, \ldots, T$. Note that $1 - p(t_l)$ can be interpreted as a "local FDR" for detecting a δ-fold expression difference at t_l. Given a desired global FDR bound α $(0 < \alpha < 1)$, we can determine a threshold ϕ_α at which to flag the set of points with $p(t_l) \geq \phi_\alpha$ as differentially expressed.

To obtain ϕ_α, we firstly sort $\{p(t_l), l = 1, \ldots, T\}$ in descending order to obtain $\{p_{(l)}, l = 1, \ldots, T\}$. Then $\phi_\alpha = p_{(s)}$, with $s = \max\{l^* : (l^*)^{-1}\sum_{l=1}^{l^*}\{1 - p_{(l)}\} \leq \alpha\}$. The set of locations $\psi = \{t_l : p(t_l) > \phi_\alpha\}$ is the set of ***discoveries***. The threshold ϕ_α is a cut-point on the posterior probabilities that corresponds to an expected Bayesian FDR of α, in the sense that on average $\geq 100(1-\alpha)\%$ locations of the set ψ should have a true δ-fold difference. That is, if $\mathbb{N}(\psi)$ is the cardinality of the set ψ, defined as $\mathbb{N}(\psi) = \sum_{l=1}^{T} I(t_l \in \psi)$, then $\mathbb{N}(\psi)^{-1}\sum_{t_l \in \psi}\Pr\{|B(t_l)| \leq \log_2(\delta)|\text{Data}\} \leq \alpha$. Morris, et al. (2008) describe the analogous criterion in the continuous space, based on Lebesgue measures.

Based on the set of discoveries ψ, we can further compute the model-based estimates of the FDR, false negative rate (FNR), sensitivity (Sens) and specificity (Spec) for detecting differentially expressed locations. Defining $\psi \cup \psi' = \mathcal{T}$, and $\mathbb{N}(\mathcal{S})$ as the cardinality of set $\mathcal{S}$, the FDR is estimated by $\mathbb{N}(\psi)^{-1}\sum_{t_l \in \psi}\{1 - p(t_l)\}$, the FNR by $\mathbb{N}(\psi')^{-1}\sum_{t_l \in \psi'} p(t_l)$, the Sens by $\{\sum_{l=1}^{T} p(t_l)\}^{-1}\sum_{t_l \in \psi} p(t_l)$, and Spec by $(\sum_{l=1}^{T}\{1 - p(t_l)\})^{-1}\sum_{t_l \in \psi'}\{1 - p(t_l)\}$. We refer to these as "empirical" quantities, since they are not based on a gold standard but are estimated based on the specified model.

We can construct an ROC curve to summarize the overall strength of our results. Instead of specifying α and computing the corresponding cutpoint ϕ_α, we vary the threshold ϕ across the entire range of (0,1), compute Sens and Spec for each, and plot Sens vs. 1-Spec to construct an ROC curve. Again, we refer to this as an *empirical ROC curve*, since it is based on model-based estimates, not a knowledge of the true curve. The area under the empirical ROC curve (AUC) can be computed and can serve as a summary measure of the strength of detected differences for these data. Alternatively, to focus on the most relevant part of the ROC curve with high specificity, we can compute the p-percentile AUC (AUC-p, e.g. for $p = 10$) by finding the area under the portion of the empirical ROC curve with $(1 - \text{Spec}) \leq p\%$ and multiplying this area by $100/p$.

The method described above can be sketched in a diagram in Figure 2, where the solid decreasing line denotes the ordered $p(t_l)$, and on the x-axis, the left side of ϕ_α is the $p(t_l)$ corresponding to the set of discoveries ψ. The areas denoted by A, B, C, D are the estimated proportions for true positives, false positives, false negatives and true negatives, respectively. The threshold ϕ_α is indeed determined by constraining $B/(A + B) \leq \alpha$, and the corresponding FNR is estimated by $C/(C + D)$, Sens by $A/(A + C)$, Spec by $D/(B + D)$.

The Bayesian FDR-based inference described above yields estimates of the statistics FDR, FNR, Sens, Spec, AUC and AUC-p without knowing the true underlying function $B(t)$. We call such estimated statistics the "empirical" quantities. In simulations, where we know the true function $B(t)$, we can compute the true statistics by computing the true false positives, false negatives, true positives and true negatives, of which we also use A, B, C, D to denote the corresponding counts. The counts are determined by fixing $A + B = \mathbb{N}(\psi)$ and $C + D = \mathbb{N}(\psi')$, for the same ψ and ψ' sets obtained when computing empirical statistics. Among the set of ψ and ψ', we can find the number of positions within the curve that truly have δ-fold differences using the true function $B(t)$. The resulting A, B, C, D counts are listed in Table 2. The statistics are computed based on these counts using the same formula as was used for those of Figure 2. We call the statistics computed from Table 2 when true $B(t)$ is known the "realized" quantities.

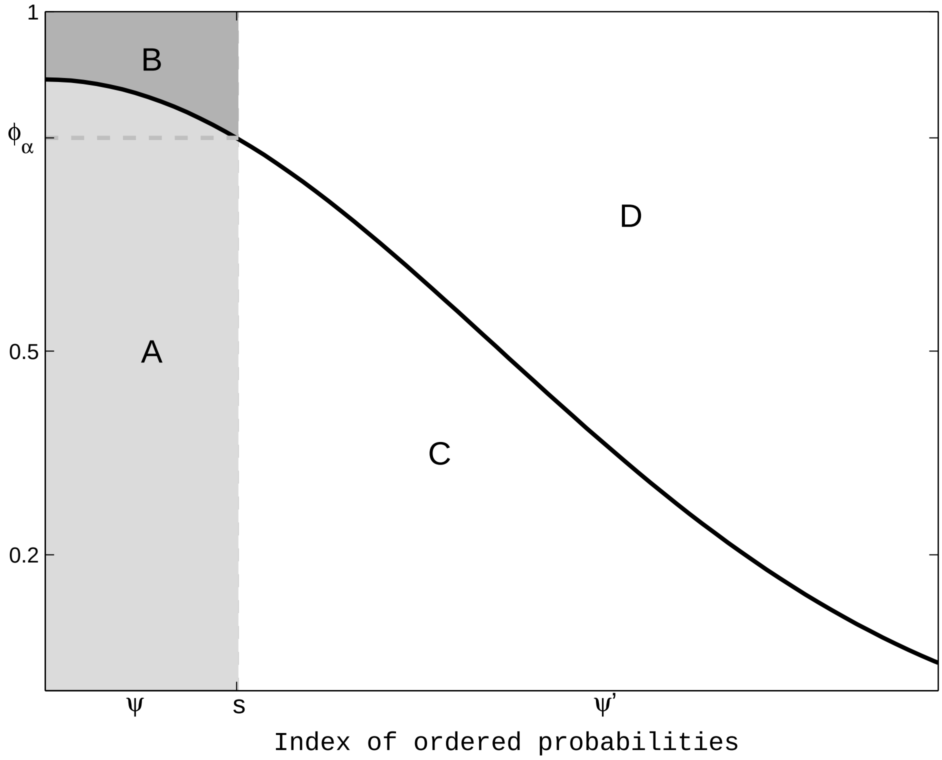

Figure 2: A diagram for Bayesian-FDR based inference.

Table 2: The true FDR inference when true $B(t)$ is known

		True $\lvert B(t_l)\rvert > \delta$		
		Yes	No	Total
$\Pr\{\lvert B(t_l)\rvert > \delta\} > \phi$	Yes	A	B	$\mathbb{N}(\psi)$
	No	C	D	$\mathbb{N}(\psi')$

4. EXTRA SIMULATION RESULTS

4.1 Extra Results for the Main Simulation

In simulation study, we also plotted the posterior mean for one fixed effect function $B_4(t)$ with corresponding 95% point-wise credible interval for both methods and all 5 distributions from one simulation run (see Figure 3). From this plot, we see that as the tails of the random effects and errors get heavier, the R-WFMM provides better estimation and more adaptive regularization than the G-WFMM, in the sense that the G-WFMM retains true spikes better while smoothing out more of the "spurious wiggles". This is most clear in regions with extreme outliers for which the MLE deviates far from the truth. In these regions, the G-WFMM is strongly affected by the outliers, with relatively poor estimation and wide credible intervals, while the R-WFMM does a much better job, with posterior mean estimates close to the truth and relatively small credible intervals.

Using the three summary measures (IMSE, IPVar, ITVar) described in the paper, we computed the ratio of G-WFMM and R-WFMM as measures of relative efficiency. For each measure, we then summarized the mean ratio across all 10 repetitions, and across index a for $B_a(t)$ and across index b for $U_b(t)$, along with the corresponding 90% intervals. Results are presented in Figure 4 with larger numbers indicating greater efficiency for the R-WFMM. Note that for clearer display, the ratios in Figure 4 are plotted after $\log_2$ transforms but are labeled according to the original scale in y-axis.

To evaluate the relative inferential performance, we also computed posterior samples for the organ, cell line, and organ-by-cell line functional effects $C_i(t), i = 1, 2, 3$, defined in Section 4 in the paper, for both the G-WFMM and R-WFMM. We then computed posterior probabilities of 1.5-fold expression changes for all 3 functional effects, and estimated the corresponding thresholds ϕ_{10} to declare significance based on a global FDR of $\alpha = 0.10$, as described in Section 3. Based on these determinations, we computed both the "realized" and "empirical" FDR, FNR, Sens, and Spec, plus the AUC and AUC10 for the realized and empirical ROC curves. The "realized"

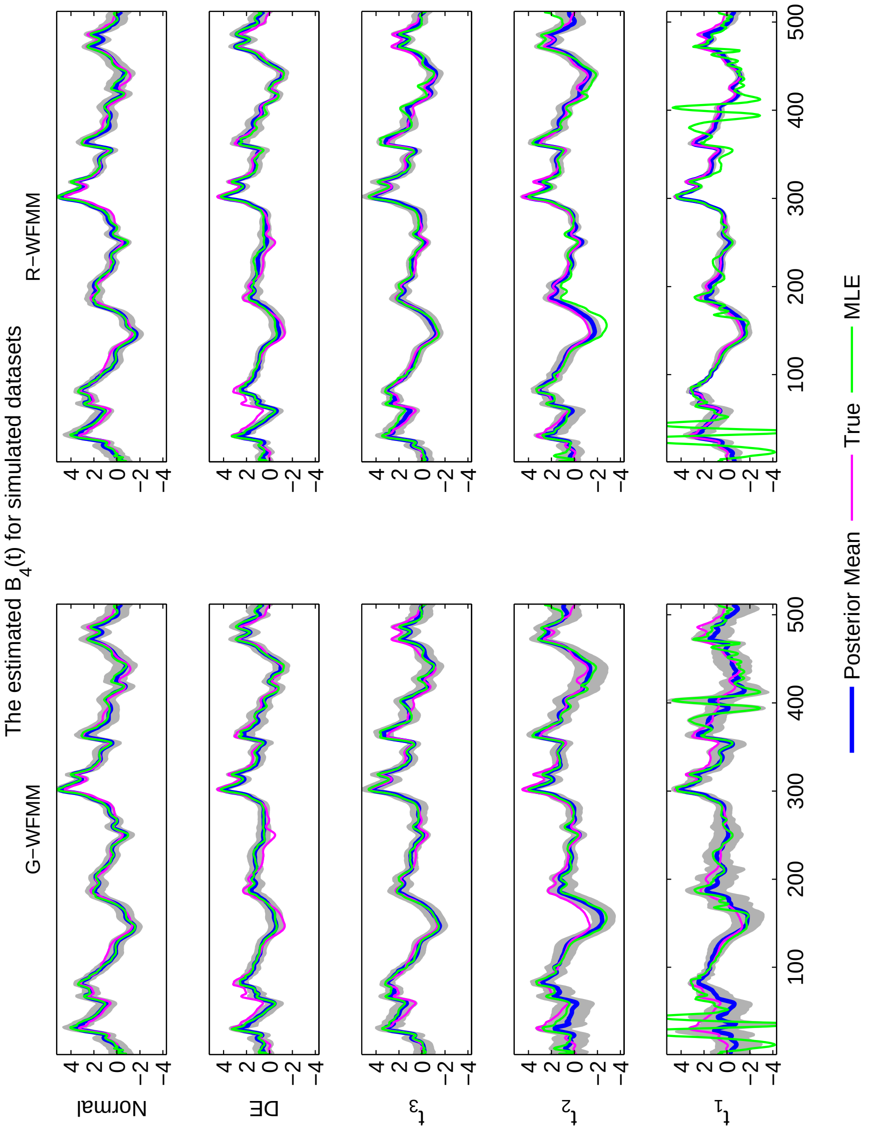

Figure 3: ***Adaptive estimation of*** $\mathrm{B}_4(\mathrm{t})$ ***for G-WFMM and R-WFMM in simulation.*** This plot presents posterior means (blue line) and 95% credible intervals (grey bands) for G-WFMM (left) and R-WFMM (right) for all 5 tail distributional assumptions used in the simulation (rows), along with the true $B_4(t)$ (pink) and naive, unregularized estimates of $B_4(t)$ (green). This plot is for one of the 10 simulations. Similar plots for other parameters for all simulation runs are available as online supplementary material.

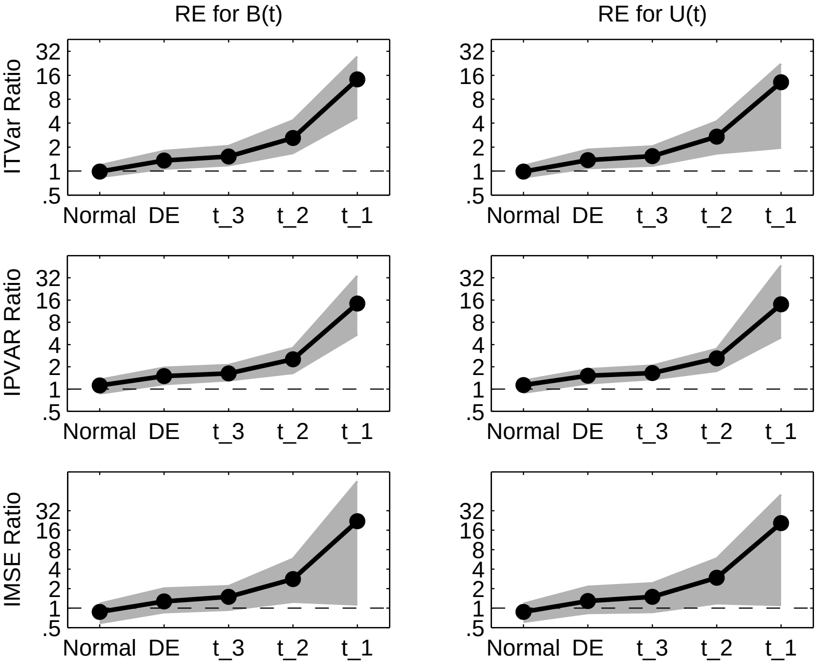

Figure 4: ***Relative efficiency of R-WFMM to G-WFMM for*** $\mathbf{B(t)}$ ***and*** $\mathbf{U(t)}$. The relative efficiencies are obtained by taking the ratio of G-WFMM to R-WFMM for the integrated mean squared error (IMSE), the integrated posterior variance (IPVar), and integrated total variance (ITVar), and summarized by mean, 5th quantile, and 95th quantile of simulation distribution, combining across simulations and fixed effect functions indexed by a. The horizontal axis indicates the distributions used to simulate the random effects and residual errors in the wavelet space.

statistics are computed based on the true $B_a(t)$, whereas the "empirical" quantities are estimated from the model without knowledge of the true $B_a(t)$. Results are given in Table 3.

Using the realized AUC as a summary measure of performance, we see that the R-WFMM considerably outperformed the G-WFMM for all simulation settings with heavier-than-normal tails, with the magnitude of the difference increasing with the heaviness of the tails. This improvement is even more pronounced in the AUC-10, which focuses on the region of the ROC curve with highest specificity, and can also been seen in the individual FDR, FNR, Sens, and Spec statistics. These results were mirrored in the estimated empirical statistics, which did not presume knowledge of the true $B_a(t)$. Note that the G-WFMM yielded slightly higher AUC and AUC-10 than the R-WFMM in the Gaussian simulation. This indicates, as expected, that some inferential price was paid for robust modeling when it was not needed, although the magnitude of this trade-off was not large compared with the improvements seen in setting of heavy-tailed distributions.

Extensive results from the 10 simulation runs are put into 10 folders, named by run1 through run10, available online at (`http://odin.mdacc.tmc.edu/~jmorris/papers.html`). Each folder contains the following files:

- Plots for fixed effects. For example, `B1_SIMU10.pdf` is the 5×2 plots of fixed effect $B_1(t)$ obtained in simulation run 10. The 5×2 plots are like Figure 3 in the main paper, with the columns indicating method (G-WFMM/R-WFMM) and the rows indicating the distribution used for the simulation (Normal,DE,t_3,t_2,t_1).

- Plots for grand mean effects (C_0(t)), organ effects($\mathrm{C}_1(t)$), cell line effects($\mathrm{C}_2(t)$), and Organ-cell-line interaction($\mathrm{C}_3(t)$) effects. For example, `C1_DE_SIMU10.pdf` is the 2×1 plots of organ effects for DE data in simulation run 10.

- Plots for random effects. For example, `U13_t_1_SIMU10.pdf` is the 2×1 plots for random effect $U_{13}(t)$ for t_1 data in simulation run 10.

Table 3: Simulations: Inferential results for G-WFMM and R-WFMM, including sensitivity, specificity, FDR, FNR, and area under the ROC curve (AUC) and partial ROC curve (AUC10), all computed both assuming knowledge of the true quantities $B_a(t)$ (realized) and computed from the model without assuming knowledge of the true quantities (empirical). Summaries combine information across all fixed effect functions $a = 1, \ldots, 5$

		Realized						Empirical					
Tails	Model	AUC	AUC10	Sens	Spec	FDR	FNR	AUC	AUC10	Sens	Spec	FDR	FNR
Normal	G	.818	.346	.288	.964	.251	.214	.816	.377	.245	.984	.097	.321
	R	.783	.301	.347	.939	.325	.204	.837	.420	.325	.978	.097	.300
DE	G	.813	.323	.316	.953	.288	.209	.826	.387	.269	.981	.097	.333
	R	.845	.402	.433	.945	.257	.181	.867	.474	.385	.976	.097	.268
t_3	G	.868	.453	.507	.938	.248	.162	.859	.461	.406	.970	.098	.293
	R	.893	.506	.602	.927	.246	.137	.899	.550	.514	.966	.097	.234
t_2	G	.771	.269	.271	.948	.343	.221	.798	.338	.232	.981	.097	.374
	R	.871	.469	.484	.950	.219	.167	.883	.501	.418	.975	.097	.252
t_1	G	.700	.226	.214	.963	.318	.232	.765	.257	.142	.982	.098	.502
	R	.881	.715	.641	.974	.099	.120	.845	.621	.557	.973	.097	.171

- Sample trace plots of model parameters. For example, `Traceplot2_t_1_SIMU10.pdf` is the second trace plot (5×2) for t_1 data in simulation run 10.

4.2 Details for the Extra Simulation

It is of interest to consider performing simulations with heavy tailed distributions directly in the data domain, such as Cauchy or other non-Gaussian processes which may induce spurious artifacts. This study is aimed to show that with data generated in time domain directly, similar results will be obtained as the result shown in the paper, i.e., we would expect the proposed robust method giving improved estimation than the Gaussian model.

We generate data in time domain as follows: Firstly two time domain "reference" covariances Σ_u, Σ_e are obtained from the reference data. These covariance are of practical forms since they are obtained based on a real dataset. Secondly, the random effects $U_b(t)$ and $E_i(t)$ are generated from multivariate t distributions with ν degree of freedom and scaling covariances Σ_u, Σ_e respectively. The design matrices and fixed effects remain the same as those in the original simulations in the paper. Figure 5 shows the plots of the data generated with $\nu = 1$, and Figure 6 shows the reference covariances and the resulting sample covariances. Note that for multivariate t distribution, denoted as $t(\nu, \mu, \Sigma)$, the mean is μ, the covariance is $\nu/(\nu - 2)\Sigma$. The true covariances do not exist when the degree of freedom $\nu \leq 2$. We applied the G-WFMM method and R-WFMM method to this data. Table 4 shows the resulting IMSE for the fixed and random effects. Note that the IMSE is computed by $\int(\hat{\theta}(t) - \theta_0(t))^2 dt$. From Table 4 we see that using the proposed R-WFMM model, the resulting IMSE for fixed and random effects are significantly smaller than that using the G-WFMM model.

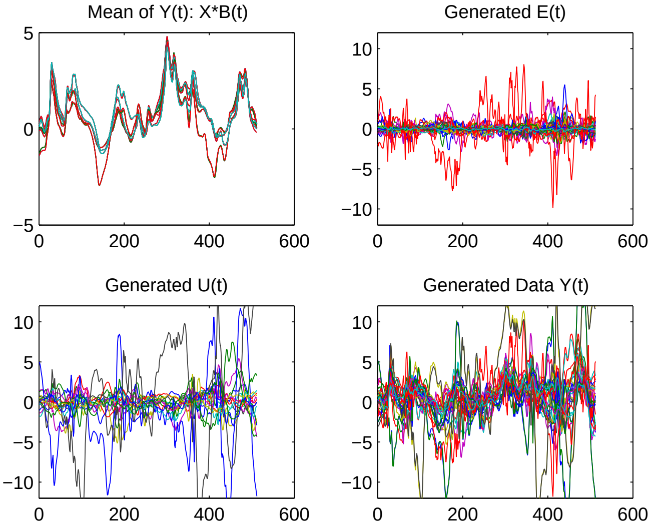

Figure 5: The plot of data generated from multivariate t distribution with degree of freedom 1.

	IMSE								
	$B_1(t)$	$B_2(t)$	$B_3(t)$	$B_4(t)$	$B_5(t)$	$U_1(t)$	$U_2(t)$	$U_3(t)$	$U_4(t)$
G-WFMM	1.21	1.14	0.86	1.02	0.003	1.17	1.28	1.32	1.14
R-WFMM	0.53	0.76	0.47	0.58	0.002	0.53	0.61	0.69	0.49

Table 4: IMSE for the G-WFMM and R-WFMM estimate for $B_a(t)$, $a = 1, ..., 5$ and $U_b(t)$, $b = 1, \ldots, 5$.

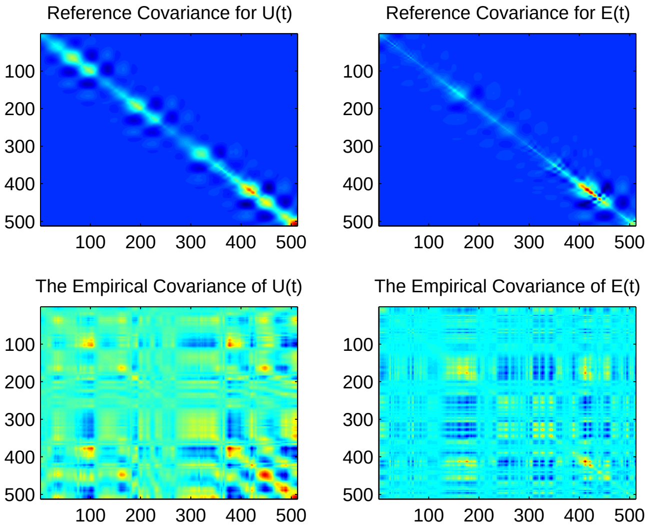

Figure 6: The image plot of reference covariances and sample covariances. The top two panels show the reference covariances for random effect and error that are used to generate data; the bottom two panels show the respective sample covariances estimated from the generated data.

5. EXTRA REAL DATA APPLICATION RESULTS

5.1 More Results Using Bayesian FDR

Here we show more results for real data applications. Figure 7 zooms in to demonstrate one of the regions flagged as significant by the R-WFMM but not the G-WFMM [7450D, 7700D]. The top panel contains the empirical mean spectra and model-based regularized posterior mean estimates for the A375P (solid line) and PC3MM2 (dotted line) cell lines, respectively, the middle panel contains the posterior mean cell line effect functions for the two methods, and the third panel contains the corresponding posterior probability plots (1.5-fold).

In addition to the significant regions for Cell Line effect plotted in Figure 2 in the paper, here we show the significant regions plots for Organ effect in Figure 8, and for Organ-by-Cell Line interaction in Figure 9.

5.2 Results for Outlier Detection

To investigate possible outliers in the data, we computed the statistics $\lambda_{i\cdot\cdot}$ for each individual spectrum, $i = 1, \ldots, 32$, and $\phi_{b\cdot\cdot}$ for each individual rat, $b = 1, \ldots, 16$, and constructed box-plots of these quantities (shown in the top panels of Figure 10). None of the individual curves or rats were flagged as outliers, defined as 1.5×IQR above the median. To check whether the regions of certain curves were outliers, we also computed the functional outlier statistics $\{\lambda_i(t)\}$ and $\{\phi_b(t)\}$ for all spectra and rats, respectively. For each t, we computed the point-wise box-plot statistics, i.e. medians $M_\lambda(t)$ and $M_\phi(t)$ and IQRs given by $IQR_\lambda(t)$ and $IQR_\phi(t)$. We flagged regions of individual spectra with $\lambda_i(t) > M_\lambda(t) + 1.5IQR_\lambda(t)$ and individual rats with $\phi_b(t) > M_\phi(t) + 1.5IQR_\phi(t)$. For illustration, we summarize the results for one spectrum ($i = 21$) with largest $\{\lambda_{i\cdot\cdot}\}$ and the one rat ($b = 4$) with largest $\{\phi_{b\cdot\cdot}\}$, marked by the triangle and square, respectively, in the top panel of Figure 10. The bottom panel contains $M_\lambda(t)$ and $M_\phi(t)$ in black, and the flagged regions of $\lambda_{21}(t)$ and $\phi_4(t)$ in red, and the middle panel plots the corresponding raw spectra (red) along with the others (black). From the left-hand panels, we see that spectrum 21 has unusually high levels of protein expression for some proteins around 4000D, and unusually low levels of expression for several peaks around 5000D and 10000D. From the right-hand panels, we see that rat 4 has unusually low levels of some proteins around 5000D, and unusually high levels for some proteins around 7000D. These outlier functions are useful diagnostics to identify unusual curves or individuals for further investigation.

6. THE GAMMA MIXTURE OF DE PRIOR

Assuming $Y \sim N(0, \lambda)$ and $\lambda \sim \text{Exp}(\nu^2/2)$, we can integrate λ out to get Double Exponential distribution of Y: $Y \sim DE(0, 1/\nu)$. If we further have $\nu^2 \sim \text{Gamma}(a, b)$, we can further integrate ν out to get the density of Y as

$$f(y) = \frac{a}{\sqrt{\pi}} \frac{2^a}{\sqrt{2b}} \Gamma(a + \frac{1}{2}) \exp\left\{\frac{y^2}{8b}\right\} D_{-2a-1}\left(\frac{|y|}{\sqrt{2b}}\right) \tag{1}$$

where $D_\nu(\cdot)$ is the parabolic cylinder function defined as $D_p(z) = e^{-z^2/4}/\Gamma(-p) \int_0^\infty \exp\{-xz - x^2/2\} x^{-p-1} dx$, $p < 0$ (Page 1028, 9.241, 2., Gradshteyn & Ryzhik, 2007). This formula is initially shown by Griffin & Brown (2005) with slightly different notation. They call the distribution associated with (1) the Normal-Exponential-Gamma (NEG) distribution. The NEG distribution is always proper as long as $0 < a, b < \infty$. Here a controls the heaviness of the tails, and b controls the scale.

An equivalent formula of (1) is:

$$f(y) = \frac{b^a}{\Gamma(a)} \int_0^\infty \nu^{2a} \exp\{-\nu|x| - b\nu^2\} d\nu. \tag{2}$$

The above formula is obtained by directly integrating ν out from the hierarchical set: $\text{DE}(0, 1/\nu)$ and $\nu^2 \sim \text{Gamma}(a, b)$.

7. THE EXPONENTIAL-GAMMA MIXING DISTRIBUTION

Upon the request of one referee, we'd like to compare the Exponential-Gamma mixing distribution with the mixing distribution that leads to Cauchy. In the hierarchical setup of scale mixture of normals with Exponential and Gamma priors, we try to collapse the Exponential and Gamma priors first by integrating ν_{jk}^2, and compare that with the mixing distribution that leads to Cauchy. We would like to see if the integrated marginal distribution is of Cauchy type.

Consider mixing distributions with respect to normal kernel. We will evaluate the heaviness of tails using the concept of regular variation. An introduction of regular variation can be found in Andrade & O'Hagan (2006). Assuming $y \sim N(0, \lambda)$, $\lambda \sim \text{Exp}(\nu^2/2)$ and $\nu^2 \sim \text{Gamma}(a, b)$, we can integrate ν^2 out to get the mixing distribution:

$$g(\lambda) = \frac{a}{2b}(1 + \frac{\lambda}{2b})^{-(a+1)}. \tag{3}$$

Since this density function is regularly varying with order $\rho = -(a + 1)$, while on the other hand Inv-Gamma$(1/2, 1/2)$ is regularly varying with order $\rho = -3/2$, the distribution (3) has heavier (right) tail than Inv-Gamma$(1/2, 1/2)$ if $a < 1/2$. Since normal mixture of Inv-Gamma(1/2,1/2) results in Cauchy, we therefore expect that using mixing distribution (3), the resulting distribution will have heavier tails than Cauchy provided $a < 1/2$. In fact, since the resulting distribution has density (1), which, when $\frac{|y|}{\sqrt{2b}}$ is large, can be approximated by

$$f(y) \approx c(\frac{|y|}{\sqrt{2b}})^{-2a-1}. \tag{4}$$

The density in (4) is regularly varying with rate $\rho = -2a - 1$, while Cauchy density is regularly vary with rate $\rho = -2$. Therefore as long as $0 < a < 1/2$, distribution (4) has heavier than Cauchy tails. This is consistent with the result when comparing the tails of mixing distributions.

8. PROOF OF ROBUSTNESS AS ONE CURVE APPROACHES INFINITY

We now provide some preliminary results on the robustness properties of the models. Although it is well known that in simple regression, using heavy tailed distributions will result in robust estimates, to our knowledge, no formal definition has been made for the robustness in functional data regression. We believe that both the definition and theoretical investigation deserve another intensive study. Results shown here are only preliminarily and are built under the particular assumptions made in this paper.

Essentially, we'd like to show that under the hierarchical setup of our model, the posterior estimates are (asymptotically) not influenced by outlying observations. We approach this by showing that as one outlying observation goes to infinity, the posterior distribution approaches to that depending only on the non-outlying data. As the model is constructed through wavelet transform, we just need to show the equivalent properties for wavelet coefficients.

Lemma 1. *Consider a simple model $d_i = \mu + \epsilon_i$, $i = 1, \ldots, n, n+1$. The prior distribution for μ, $\pi(\mu)$ is proper. The likelihood is assumed through the following hierarchical setup: $d_i|\mu \sim N(\mu, \lambda_i)$, $\lambda_i \sim Exp\{\nu^2/2\}$, $\nu^2 \sim Gamma(a, b)$, where $0 < a, b < \infty$ and a, b are known constants. Then the resulting likelihood distribution for this model is is outlier-prone, which means that the posterior distribution*

$$Pr(\mu < c \mid d_1, d_2, \ldots, d_{n+1}) \longrightarrow Pr(\mu < c \mid d_1, d_2, \ldots, d_n), \quad \text{as } |d_{n+1}| \to \infty, \tag{5}$$

for all c and $\{d_1, \ldots, d_n\}$ and for all proper $\pi(\mu)$.

Remark: The terminology *outlier prone* is defined in O'Hagan (1979), which essentially means that a data-generating distribution *"have well-behaved and "thick" tail, so that when the observation becomes large the information it carries is weak."* The counterpart *outlier resistance* means that *"a posterior distribution that necessarily "increases" when an observation increases".* It was shown that normal distribution is in the outlier resistant family, while t-distribution is outlier prone.

Proof: We will follow the results of O'Hagan (1979). Firstly, provided that the likelihood $f(\cdot)$ is bounded, if the outlier proneness holds for some n, it holds for n=1. Since the posterior distribution can be written as:

$$\begin{aligned} & \int_{\mu<c} f(\mu|d_1, \ldots, d_n, d_{n+1}) d\mu \\ \propto & \int_{\mu<c} \prod_{i=1}^{n+1} f(d_i|\mu)\pi(\mu) d\mu \\ \propto & \int_{\mu<c} f(d_{n+1}|\mu) dG(\mu|d_1, \ldots, d_n) \end{aligned}$$

where $f(d_i|\mu)$ is the likelihood and $G(\mu|d_1,\ldots,d_n)$ is the posterior distribution based on the first n observations. If (5) holds for any proper prior $\pi(\mu)$, we can treat the $G(\mu|d_1,\ldots,d_n)$ as a "new" proper prior, corresponding to a single observation d_{n+1}. It thus holds for $n=1$. Similar arguments can be found on Page 362 of O'Hagan(1979). Secondly, under the hierarchical setup, if we integrate out the intermediate parameters λ_i and ν^2, we find that the likelihood (residual) distribution $f(d-\mu)$ takes the following form:

$$f(y)=\frac{a}{\sqrt{\pi}}\frac{2^a}{\sqrt{2b}}\Gamma(a+\frac{1}{2})\exp\left\{\frac{y^2}{8b}\right\}D_{-2a-1}\left(\frac{|y|}{\sqrt{2b}}\right) \tag{6}$$

where $D_\nu(\cdot)$ is the parabolic cylinder function defined as $D_p(z)=e^{-z^2/4}/\Gamma(-p)\int_0^\infty \exp\{-xz-x^2/2\}x^{-p-1}dx$, $p<0$ (Page 1028, 9.241, 2., Gradshteyn & Ryzhik, 2007). This formula is initially shown by Griffin & Brown (2005) with slightly different notation. They call the distribution associated with (6) the Normal-Exponential-Gamma (NEG) distribution. According to O'Hagan (1979), we simply need to show that $f(y)$ is outlier prone of order one. Note that $f(y)$ in (6) is obviously symmetric and bounded. To show outlier proneness for a symmetric density, we only need to verify the following conditions listed in O'Hagan for the *right* outlier prone. The *left* outlier proneness follows by symmetry.

(i) Given $\epsilon>0, h>0$, there exists A such that if $y>A$, then $|f(y')-f(y)|<\epsilon f(y)$ whenever $|y'-y|<h$.

(ii) (a) $f(y)$ is continuous and positive for all $y\in\mathbb{R}$.

(b) There exist a B such that, for all $y\geq B$, (I) f(y) is decreasing in y. (II) $b(y)=d\text{log}f(y)/dy$ exists and is increasing in y.

The part (ii) (a) is obvious from (6). The part (ii) (b) (I) is also obvious (for $B=0$). We simply need to verify (i) and (ii) (b) (II). To verify (i), we use the asymptotic expansions for the parabolic cylinder function: for $|z|>>1$ and $|z|>>p$, we have

$$D_p(z)\sim e^{-\frac{z^2}{4}}z^p\left(1-\frac{p(p-1)}{2z^2}+o(1/z^2)\right)$$

(Page 1029, 9.246, 1. Gradshteyn & Ryzhik, 2007). Using this, we get $f(y) \sim c_1 y^{-2a-1}(1+c_2/y^2+o(1/y^2))$ for constant c_1, c_2. Take y large so that $y > \max(2\sqrt{b}, 2hc_3/\epsilon, h\sqrt{2c_4/\epsilon})$,

$$\begin{aligned}
\left|\frac{f(y')-f(y)}{f(y)}\right| &= \left|\frac{(y'-y)f'(y)+(y'-y)^2 f''(y)/2+...}{f(y)}\right| \\
&= \left|(y'-y)\frac{f'(y)}{f(y)}+\frac{(y'-y)^2}{2}\frac{f''(y)}{f(y)}+...\right| \\
&\approx |(y'-y)(-2a-1)y^{-1}+O((y'-y)^2)O(y^{-2}))| \\
&< c_3 h y^{-1} + c_4 h^2 y^{-2} \\
&< \epsilon,
\end{aligned}$$

where c_3, c_4 are some constants.

To show (ii) (b) (II), assume y is positive and extremely large,

$$\begin{aligned}
b(y) = \frac{d\log f(y)}{dy} = \frac{f'(y)}{f(y)} &\approx \frac{c_1(-2a-1)y^{-2a-2}+c_1c_2(-2a-3)y^{-2a-4}+o(y^{-2a-4})}{c_1 y^{-2a-1}+c_1c_2 y^{-2a-3}+o(y^{-2a-3})} \\
&\approx -(2a+1)y^{-1}
\end{aligned}$$

therefore $b(y)$ is increasing in y. This shows that the distribution in (6) is outlier prone, therefore (5) holds.

Proposition 1. *Consider the model $Y_i(t) = B(t) + E_i(t)$, $i = 1, \dots, n, n+1$. Assuming $Y_i(t), B(t), E_i(t)$ are in $L^2[T]$, $T \subseteq \mathbb{R}$, associated with L^2 norm. The corresponding wavelet domain model can be written as $D_i = B + E_i^*$, with $D_i = \{d_{i,jk}\}_{jk}$, $B = \{b_{jk}\}_{jk}$, $E_i^* = \{\epsilon_{i,jk}\}_{jk}$, where $j = 1, \dots, J$ is the index for scales and $k = 1, \dots, k_j$ is the index for locations. For each (j,k), assuming the following hierarchical setup for the likelihood, independently across j, k:*

$$\begin{aligned}
(d_{i,jk}|b_{jk}) &\sim N(0, \lambda_{i,jk}), \\
\lambda_{i,jk} &\sim Exp(\nu_{jk}^2/2), \\
\nu_{jk}^2 &\sim Gamma(a_{jk}, b_{jk}).
\end{aligned}$$

In addition, assuming priors: $b_{jk} \sim \pi(\theta_{jk})$, where $\pi(\theta_{jk})$ are proper and are independent across j, k. Then given $A \in \mathcal{B}$, where $\mathcal{B}$ is the σ-algebra generated by Borel sets in $L^2[T]$, as $||Y_{n+1}(t)|| \to \infty$, we have either

$$Pr(B(t) \in A|Y_1(t), \dots, Y_n(t), Y_{n+1}(t)) \longrightarrow Pr(B(t) \in A|Y_1(t), \dots, Y_n(t))$$

or

$$Pr\left(B(t) \in A|Y_1(t), \dots, Y_n(t), Y_{n+1}(t)\right) \longrightarrow Pr\left(B(t) \in A|Y_1(t), \dots, Y_n(t), \widetilde{Y}_{n+1}(t)\right).$$

where $\widetilde{Y}_{n+1}(t) \neq Y_{n+1}(t)$ *and* $||\widetilde{Y}_{n+1}(t)|| < \infty$.

Remark: Proposition 1 says that as the norm of the outlying observation approaches infinity, the posterior distribution of the mean function $B(t)$ approaches to a posterior that either depends only on the non-outlying observations, or depends on the non-outlying observations as well as the "partial" outlier, where the "partial" outlier are composed of those finite-valued wavelet components.

Proof: By Parseval's identity, we have $||Y_{n+1}(t)||^2 = \sum_{j,k} d^2_{n+1,jk}$, for which the right hand side is a finite sum when J is finite. Therefore $||Y_{n+1}(t)|| \to \infty$ is equivalent to: (1) all components in $\{d_{n+1,jk}\}_{jk}$ approach infinity, or (2) a subset of the components $\{d_{n+1,jk}\}_{jk}$ approach infinity. On the other hand, because of the map $Y_i(t) \to \{d_{i,jk}\}_{jk}$ is an isometric isomorphism, there exist a sequence of subsets $\{C_{jk}\}_{jk}$, with $C_{jk} \subset \mathbb{R}$ such that $\Pr\left(B(t) \in A|Y_1(t), \dots, Y_n(t), Y_{n+1}(t)\right) = \Pr\left(\{b_{jk} \in C_{jk}\}_{jk}|D_1, \dots, D_n, D_{n+1}\right)$. In addition, since we assume independent likelihood and priors across j, k, we further have

$$\Pr\left(\{b_{jk} \in C_{jk}\}_{jk}|D_1, \dots, D_n, D_{n+1}\right) = \prod_{jk} \Pr\left(b_{jk} \in C_{jk}|d_{1,jk}, \dots, d_{n,jk}, d_{n+1,jk}\right)$$

. Now we discuss the two cases:

- If all components $d_{n+1,jk} \to \infty$, then by Lemma 1., we have

  $$\Pr(b_{jk} \in C_{jk} \mid d_{1,jk}, \dots, d_{n,jk}, d_{n+1,jk}) \longrightarrow \Pr(b_{jk} \in C_{jk} \mid d_{1,jk}, \dots, d_{n,jk}).$$

  for all j, k. Therefore,

  $$\prod_{j,k} \Pr(b_{jk} \in C_{jk} \mid d_{1,jk}, \dots, d_{n,jk}, d_{n+1,jk}) \longrightarrow \prod_{j,k} \Pr(b_{jk} \in C_{jk} \mid d_{1,jk}, \dots, d_{n,jk}).$$

  which implies that $\Pr\left(B(t) \in A|Y_1(t), \dots, Y_n(t), Y_{n+1}(t)\right) \longrightarrow \Pr\left(B(t) \in A|Y_1(t), \dots, Y_n(t)\right)$.

- If a subset of the components $\{d_{n+1,jk}\}_{jk}$ approaches infinity, denote the (j,k) index of this subset by $s = \{(j_p, k_q)\}_q$. Then we have

$$\prod_{(j,k)\in s} \Pr(b_{jk} \in C_{jk} \mid d_{1,jk}, \ldots, d_{n,jk}, d_{n+1,jk}) \longrightarrow \prod_{(j,k)\in s} \Pr(b_{jk} \in C_{jk} \mid d_{1,jk}, \ldots, d_{n,jk}).$$

Therefore,

$$\prod_{j,k} \Pr(b_{jk} \in C_{jk} \mid d_{1,jk}, \ldots, d_{n,jk}, d_{n+1,jk}) \longrightarrow$$
$$\left[\prod_{(j,k)\notin s} \Pr(b_{jk} \in C_{jk} \mid d_{1,jk}, \ldots, d_{n,jk}, d_{n+1,jk})\right] \left[\prod_{(j,k)\in s} \Pr(b_{jk} \in C_{jk} \mid d_{1,jk}, \ldots, d_{n,jk})\right] \quad (7)$$

The right hand side of (7) is equivalent to $\Pr\left(B(t) \in A | Y_1(t), \ldots, Y_n(t), \tilde{Y}_{n+1}(t)\right)$, where $\tilde{Y}_{n+1}(t) = \sum_{(j,k)\notin s} d_{n+1,jk}\phi_{jk}(t)$ and $\{\phi_{jk}(t)\}_{jk}$ represents the wavelet basis. In addition, $||\tilde{Y}_n(t)||^2 = \sum_{(j,k)\notin s} d^2_{n+1,jk} < \infty$.

REFERENCES

[1] Andrade, J. A. A. and O'Hagan, A. (2006). Bayesian robustness modeling using regularly varying distributions. *Bayesian Analysis* **1,** 169-188.

[2] Carlin, B. P. and Chib, S. (1995). Bayesian Model Choice via Markov Chain Monte Carlo methods. *J. R. Statist. Soc. B* **57,** 473-484

[3] Gradshteyn I. S. and Ryzhik I. M. (2007). *Table of Integrals, Series, and Products*, Seventh Edit. Academic Press.

[4] Griffin J. E. and Brown P. J. (2005). Alternative prior distribution for variable selection with very many more variables than observations. CRiSM Working Paper No. 05-10, University of Warwick. Available online at `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.126.3430&rep=rep1&type=pdf`.

[5] Morris, J. S. and Carroll, R. J. (2006). Wavelet-based functional mixed models. *J. R. Statist. Soc. B* **68,** 179-199

[6] O'Hagan A. (1979) On outlier rejection phenomena in Bayes inference. *J. R. Statist. Soc. B*, **41**, 358-367.

[7] Park, T. and Casella, G. (2008). The Bayesian Lasso. J. Am. Statist. Ass. **103,** 681-686

[8] Searle S. R. and Casella, G. and McCulloch C. E. (1992). *Variance Components.*

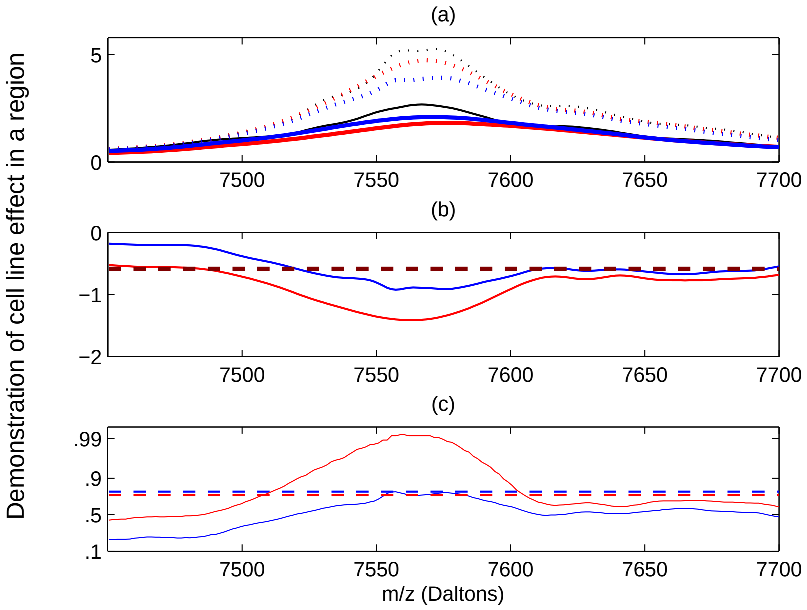

Figure 7: Cell line effect on the region [7450D,7700D]. (a) Empirical mean spectra for each cell line and the corresponding model-based regularized posterior mean functions from each model. Dotted lines: cell line PC3MM2; Solid lines: cell line A375P; Blue color: estimate of G-WFMM; Red color: estimate of R-WFMM; Black color: empirical means. (b) The posterior mean estimates for cell line effect functions. Blue line: G-WFMM; Red line: R-WFMM. (c) Posterior probability discovery plot of 1.5-fold expression differences. Sold lines: the point-wise probabilities; Dashed lines: the threshold obtained using Bayesian FDR based inference, $\alpha = 0.10$. Blue color: G-WFMM; Red color: R-WFMM.

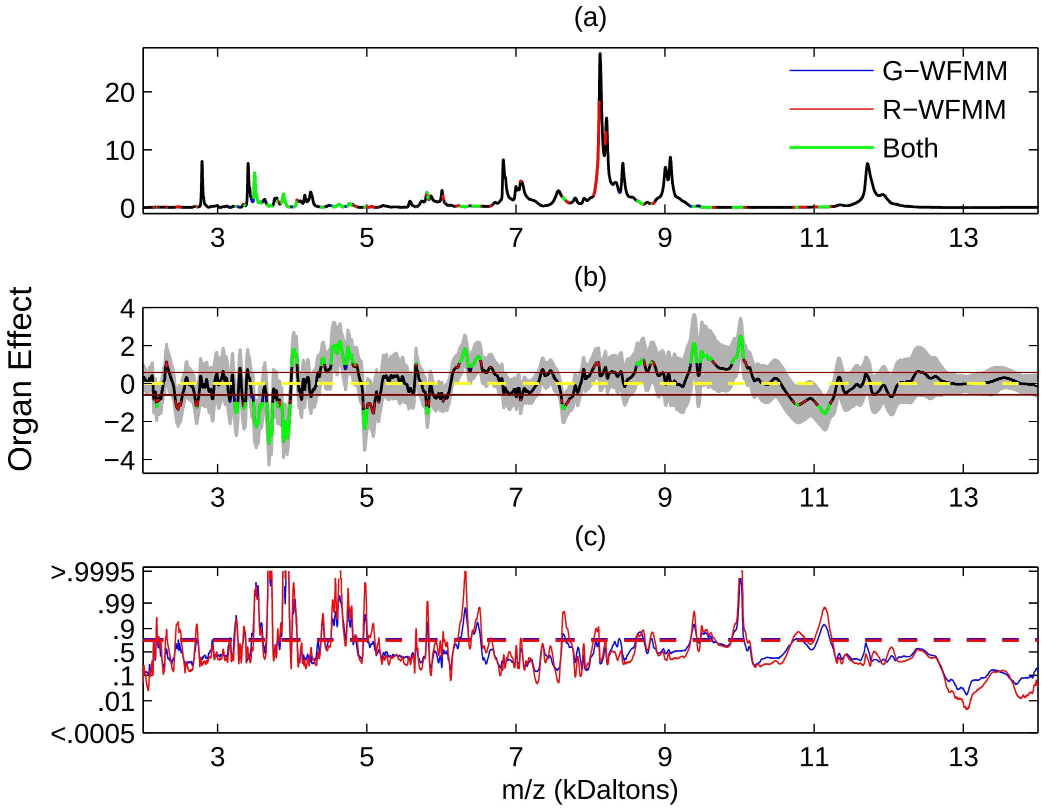

Figure 8: Significant regions of 1.5 fold differences for Organ effect $C_1(t)$ flagged out by G-WFMM and R-WFMM. (a) The regions flagged on the grand mean function $C_0(t)$, plotted in the original scale. (b) The same regions flagged on the organ effect function $C_1(t)$, plotted in log2 scale. In both (a) and (b), Blue color: regions flagged by G-WFMM only; Red color: regions detected by R-WFMM only; Green color: regions detected by both methods; Black color: regions detected by neither methods. (c) The corresponding posterior probability estimates and the thresholds obtained using Bayesian FDR-based inference, with $\alpha = 0.10$. In (c), Blue color represents G-WFMM, red color represents R-WFMM.

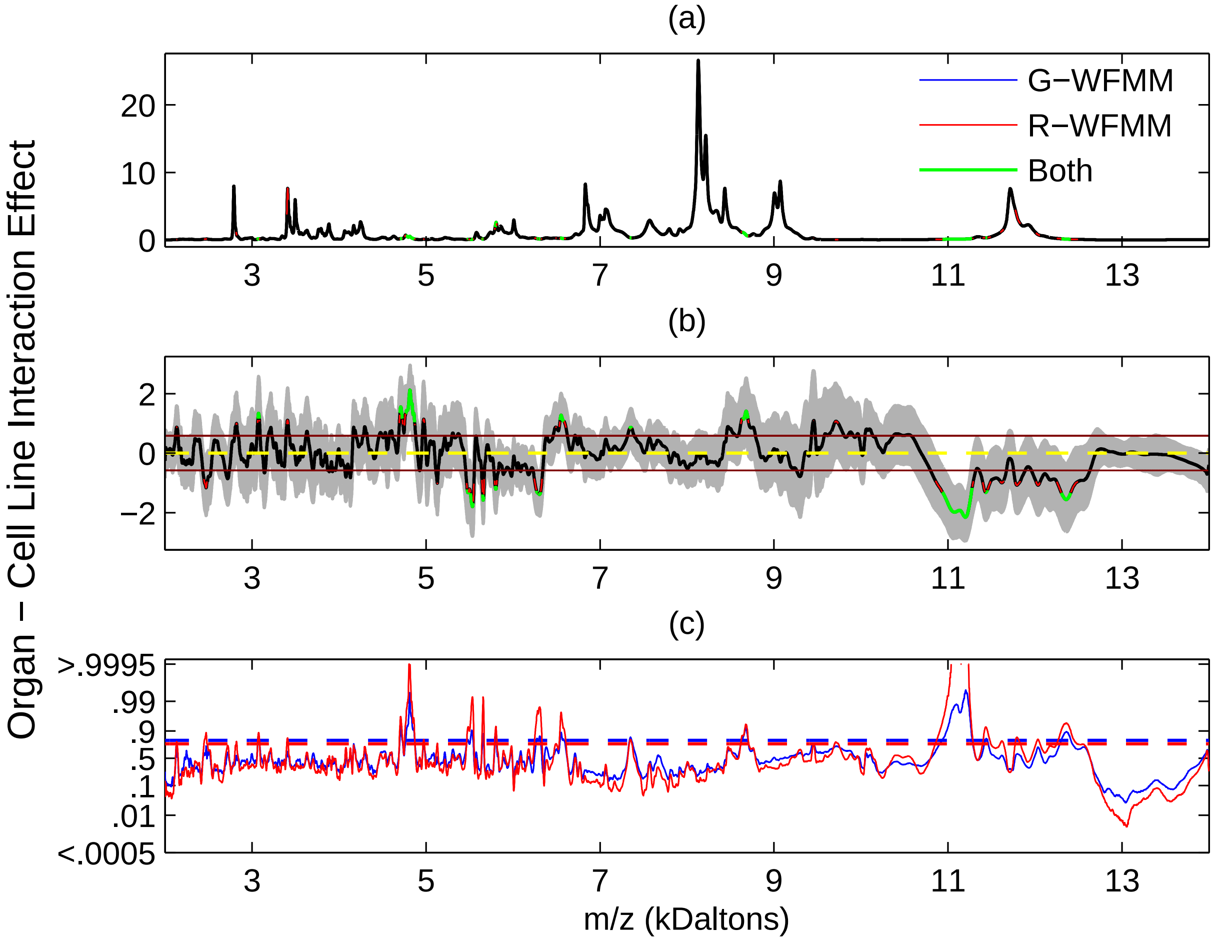

Figure 9: Significant regions of 1.5 fold differences for Organ-by-Cell Line interaction effect $C_3(t)$ flagged out by G-WFMM and R-WFMM. (a) The regions flagged on the grand mean function $C_0(t)$, plotted in the original scale. (b) The same regions flagged on the organ effect function $C_3(t)$, plotted in log2 scale. In both (a) and (b), Blue color: regions flagged by G-WFMM only; Red color: regions detected by R-WFMM only; Green color: regions detected by both methods; Black color: regions detected by neither methods. (c) The corresponding posterior probability estimates and the thresholds obtained using Bayesian FDR-based inference, with $\alpha = 0.10$. In (c), Blue color represents G-WFMM, red color represents R-WFMM.

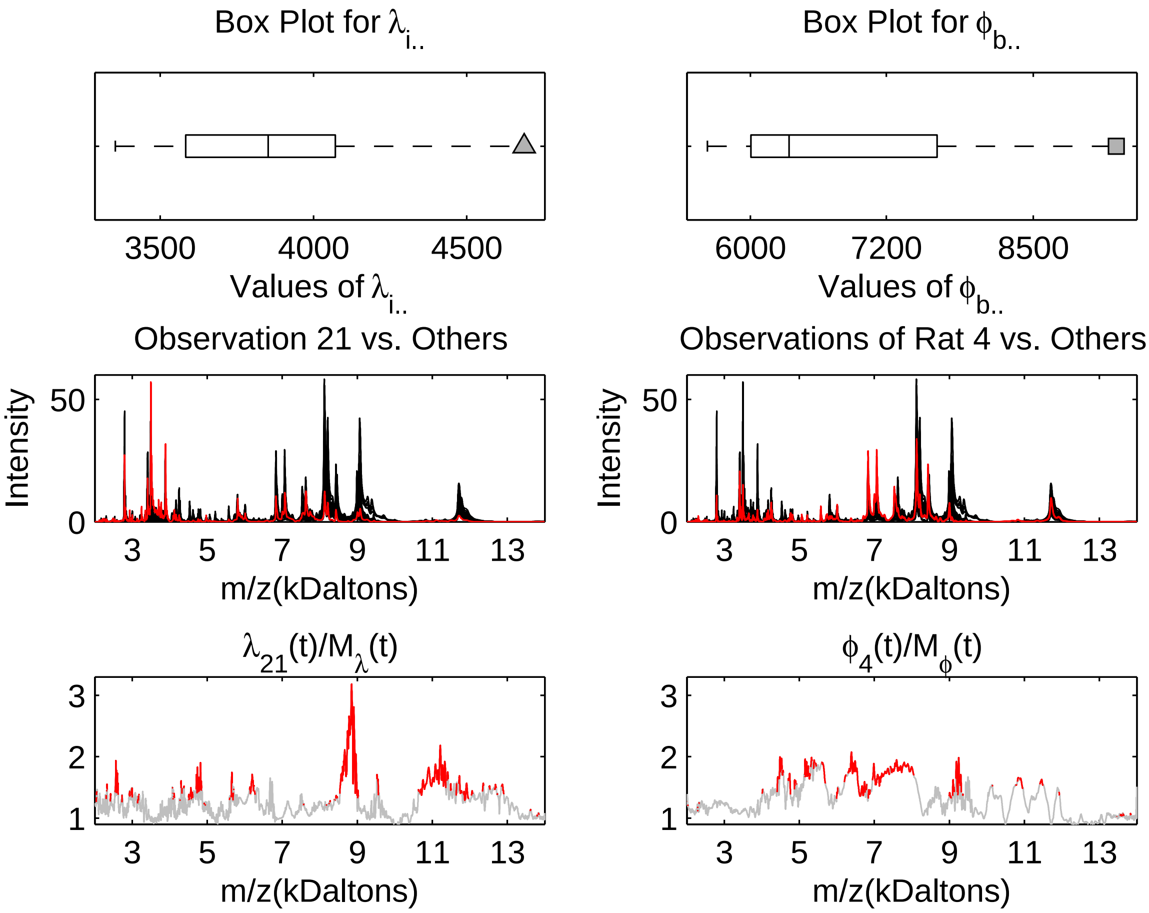

Figure 10: ***Detecting outliers using scaling parameters from R-WFMM.*** Top panels are box plots of scalar outlier scores for individual curves $\lambda_{i..}$ (left) and rats (random effects) $\phi_{b..}$ (right). The middle panels plot all spectra (black), with highlighted spectra in red, which are spectrum 21 (left) and the spectra corresponding to rat 4 (right), the spectrum and rat with highest scalar outlier scores. The bottom panels summarize the point-wise outlier scores. In the bottom left panel, the gray line is the point-wise ratio of $\lambda_{21}(t)$ vs. $M_\lambda(t)$, where the latter is the median of the $\lambda_i(t)$. The red color highlights the portions of spectrum 21 that are detected as outliers. The bottom right is plotted in a similar way for the ratio of $\phi_4(t)$ vs. $M_\phi(t)$, where the latter is the median for the rat effects.