Web-based Supplementary Materials for "Robust Classification of Functional and Quantitative Image Data using Functional Mixed Models"

Hongxiao Zhu

Department of Statistical Science, Duke University, Durham, NC 27708, U.S.A.

Philip J. Brown

School of Mathematics, Statistics and Actuarial Science, University of Kent, U.K.

Jeffrey S. Morris

jefmorris@mdanderson.org

Department of Biostatistics, University of Texas M. D. Anderson Cancer Center,Houston, TX 77230, U.S.A.

Web Appendix A

Details of the MCMC Algorithm for R-WFMM

We perform a Markov Chain Monte Carlo algorithm to draw samples from the posterior of the wavelet-space model parameters of R-WFMM, 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. Note that when training the data, we combine the design matrix V and X together, and both are treated as fixed effects. Therefore in the following algorithm, the first q rows of B^* will be G^*.

Step 0. Initialize $\{\nu^E_{jk}\}, \{\nu^U_{jk}\}, \{\nu^B_{jk}\}$ 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 $\mathbf{\Lambda}^{-1/2}_{jk} = \text{diag}\{\lambda_{ijk}\}^n_{i=1}$, 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 $\boldsymbol{\mu}_{u_{jk}^{*}} = \{(\mathbf{Z}_{jk}^{+})^{T}\mathbf{Z}_{jk}^{+} + \mathbf{\Phi}_{jk}^{-1}\}^{-1}(\mathbf{Z}_{jk}^{+})^{T}(\mathbf{d}_{jk}^{+} - \mathbf{X}_{jk}^{+}\mathbf{B}_{jk}^{*})$ and $\mathbf{V}_{u_{jk}^{*}} = \{(\mathbf{Z}_{jk}^{+})^{T}\mathbf{Z}_{jk}^{+} + \mathbf{\Phi}_{jk}^{-1}\}^{-1}$, with $\mathbf{\Phi}_{jk} = \text{diag}\{\phi_{bjk}\}_{b=1}^{m}$. 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.

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

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^{ψ}_{aj}.

Step 5. Update the lasso parameters $\{\nu^E_{jk}\}$, $\{\nu^U_{jk}\}$, $\{\nu^B_{jk}\}$ from their complete conditional distributions. Their squared values are conjugate gammas, i.e., $(\nu^E_{jk})^2|\{\lambda_{ijk}\}_i \sim \text{Gamma}(n + a^E, \sum_{i=1}^n \lambda_{ijk}/2 + b^E)$, $(\nu^U_{jk})^2|\{\phi_{bjk}\}_b \sim \text{Gamma}(m + a^U, \sum_{b=1}^m \phi_{bjk}/2 + b^U)$, and $(\nu^B_{jk})^2|\{\psi_{ajk}\}_a \sim \text{Gamma}(K_j + a^B, \sum_{k=1}^{K_j} \psi_{ajk}/2 + b^B)$, 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.

Web Appendix B

Proof of Proposition 1

Proof: For the case of $b_1 \neq b_2$, from Corollary 1 of Nadarajah and Kotz (2005), we have for $Z > 0$,

$$\begin{aligned} f(z) = dF(z)/dz &= \left[\mu/2 + \mu^2/(4(\lambda-\mu)) - \mu^2/(4(\lambda+\mu))\right]\exp\{-\mu z\} \\ &+ \left[\lambda\mu/(4(\mu+\lambda)) - \lambda\mu/(4(\lambda-\mu))\right]\exp\{-\lambda z\}. \end{aligned}$$

For $Z < 0$,

$$\begin{aligned} f(z) = dF(z)/dz &= \left[\mu/2 - \mu^2/(4(\lambda+\mu)) - \mu^2/(4(\mu-\lambda))\right]\exp\{\mu z\} \\ &+ \left[\lambda\mu/(4(\mu-\lambda)) + \lambda\mu/(4(\lambda+\mu))\right]\exp\{\lambda z\}. \end{aligned}$$

Now since the density of DE is defined differently with Nadarajah and Kotz (2005), the λ in Corollary 1 of Nadarajah and Kotz (2005) corresponds to $1/b_1$ in our Proposition 1, and μ corresponds to $1/b_2$ in our Proposition 1. Replacing $\lambda = 1/b_1$, $\mu = 1/b_2$, we get for $z > 0$,

$$\begin{aligned} f(z) &= \left[\frac{1}{2b_2} + \frac{1/b_2^2}{4(1/b_1 - 1/b_2)} - \frac{1/b_2^2}{4(1/b_1 + 1/b_2)}\right]\exp\{-z/b_2\} \\ &+ \left[\frac{1/(b_1b_2)}{4(1/b_1 + 1/b_2)} - \frac{1/(b_1b_2)}{4(1/b_1 - 1/b_2)}\right]\exp\{-z/b_1\}. \end{aligned}$$

For $z < 0$,

$$\begin{aligned} f(z) &= \left[\frac{1}{2b_2} - \frac{1/b_2^2}{4(1/b_1 + 1/b_2)} - \frac{1/b_2^2}{4(1/b_2 - 1/b_1)}\right]\exp\{z/b_2\} \\ &+ \left[\frac{1/(b_1b_2)}{4(1/b_2 - 1/b_1)} + \frac{1/(b_1b_2)}{4(1/b_2 + 1/b_1)}\right]\exp\{z/b_1\}. \end{aligned}$$

Simplifying the above formula and combine the cases of $z > 0$ and $z < 0$, we get the result of Proposition 1 for $b_1 \neq b_2$.

In case of $b_1 = b_2 = b$, we see that X_1 and X_2 are i.i.d. DE$(0, b)$. Firstly note that the DE distribution is a special case of Normal Gamma (NG) distribution with the shape parameter equals 1, and the inverse scale parameter $1/(2b^2)$, i.e., a DE$(0, b)$ is a NG$(1, 1/(2b^2))$. An NG distribution

is defined as the marginal distribution of a scale mixture of Gamma, i.e., $\theta \sim \text{NG}(\lambda, 1/(2\gamma^2))$, if $\theta \sim N(0, \Psi)$ and $\Psi \sim \text{Gamma}(\lambda, 1/(2\gamma^2))$. The formula of the general $NG(\lambda, 1/(2\gamma^2))$ density can be found in equation (3) of [2]. The NG distribution has the property that the sum of two i.i.d. $\text{NG}(1, 1/(2b^2))$ is distributed as $\text{NG}(2, 1/(2b^2))$, with the shape parameter doubled. Using the equation (3) of Griffin and Brown (2010) and applying the formula 10.2.17 in page 444 of Abramowitz and Stegun (1972) for the simplified Bessel function, we obtain the density for the case of $b_1 = b_2 = b$ shown in Proposition 1.

Web Appendix C

Classification Performance using WFMM when Ignoring the Random Effects

In the pancreatic cancer application, in order to see what happens if one ignores the random effects (time block effects), we refitted the GWFMM and RWFMM model without using the block information (both in training and prediction steps). The models are fitted under the setup where 90% wavelet compression is used, and for both 4-fold cross validation setups(one for in-block classification and one for out-of-block classification). We denoted the methods of not using random effects as "GWFMM-90, no RE" and "RWFMM-90, no RE" for the Gaussian model and Robust model, respectively. The results are compared with all other FMM classification outputs in Table 1. The corresponding empirical ROC curves are plotted in Figure 3. Both Table 1 and Figure 3 show that for the in-block classification cases, the prediction performance is systematically degraded if ignoring the random effects. For the out-of-block classification case, the performance of GWFMM is similar (with AUC .818 vs. .815) if ignoring the random effects, whearas the RWFMM model performs worse when ignoring the random effects. In summary, we found that at least for this data, if random effects present, ignoring them will most likely lead to worse prediction. The prediction power can be systematically improved if taking into account these effects using the proposed FMM framework.

Web Appendix D

Allowing the Covariance to Vary by Class

In both the G-WFMM and R-WFMM discussed in Sections 3.1 and 3.2, we assumed that the distribution of $\mathbf{U}(t)$ and $\mathbf{E}(t)$ in model (1) were common for all classes. In some settings, one may wish the random effect and/or residual error covariances to vary across class, which would yield more flexible classification rules. This has been previously described for the G-WFMM (Morris and Carroll, 2006) and involves expanding the variance components $\{q^*_{j,k}\}, \{s^*_{j,k}\}$ to $\{q^{*,\,c}_{j,k}\}, \{s^{*,\,c}_{j,k}\}, c = 1, \ldots, q$. For prediction, the posterior predictive probability needs to be adjusted so that the corresponding variance components of $c = j$ are used when the likelihood conditions on class label $c = j$. In the R-WFMM, we allow the population scale parameters $\{\nu^E_{jk}\}$ and $\{\nu^U_{jk}\}$ to be class specific, i.e., $\{\nu^{E,c}_{jk}\}, \{\nu^{U,c}_{jk}\}, c = 1, \ldots, q$. Correspondingly, their Gamma hyper-prior parameters (α^E, β^E) and (α^U, β^U) would also be indexed by c. This involves only minor changes of the previously described MCMC algorithm. Similarly, for prediction, we need to adjust the DE likelihood by plugging in the corresponding population scale parameters when conditioning on a particular $c = j$.

References

[1] Abramowitz, M. and Stegun, I. A. (1972). *Handbook of Mathematical Functions With Formulas, Graphs, and Mathematical Tables* Dover, New York.

[2] Griffin, J. E. and Brown, P. J. (2010). Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis* **5,** 171–188.

[3] Nadarajah, S. and Kotz, S. (2005). On the linear combination of Laplace random variables. *Probability in the Engineering and Informational Sciences* **19,** 463–470.

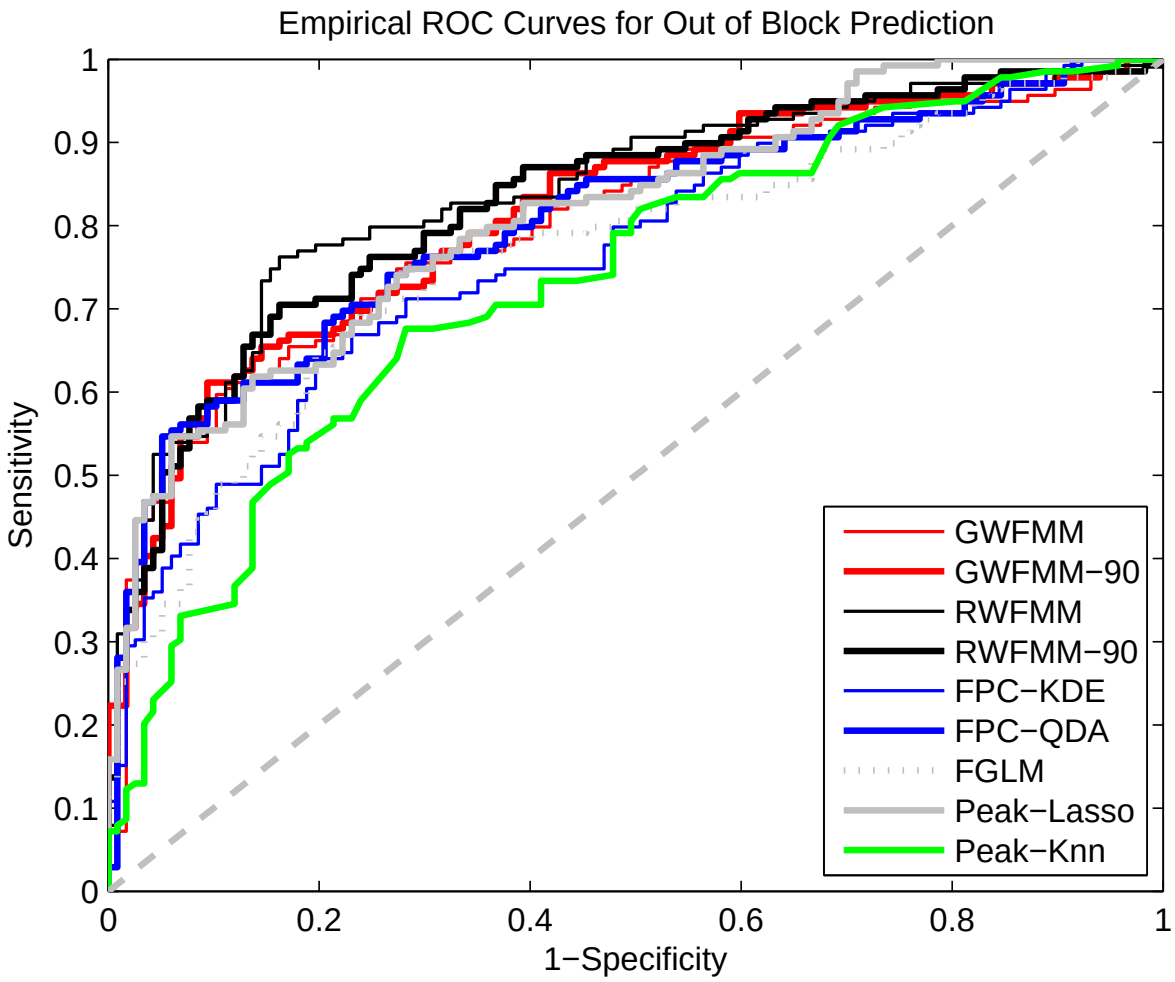

Figure 1: (Web figure) ROC plot for out of block prediction.

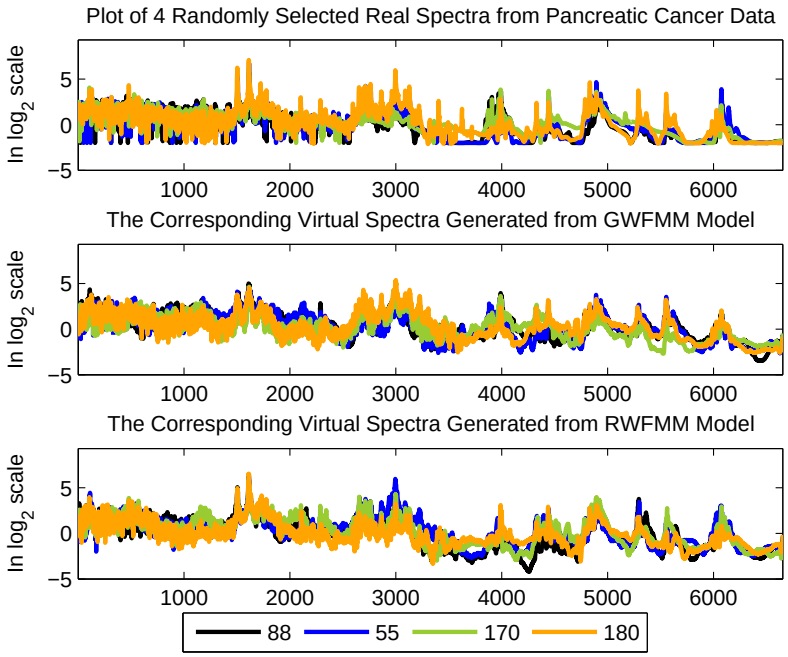

Figure 2: (Web figure) "Virtual Sepctra" plot.

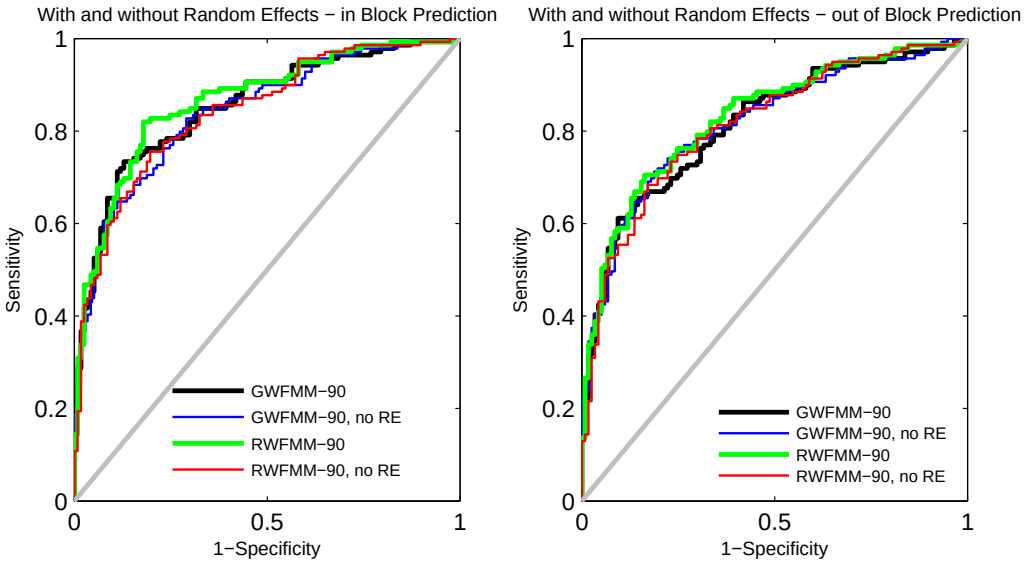

Figure 3: (Web figure) Empirical ROC curves for FMM classification either using random effects or ignoring random effects.

Table 1: (Web Table) Comparison of FMM Classification Results with and without Random Effects

	Methods	Model Name	AUC	MisR	Sens	Spec
In Block	FMM	GWFMM	0.816	0.270	0.669	0.812
		$GWFMM_{90}$	0.854	0.211	0.719	0.880
		$GWFMM_{90}$-NoRE	0.842	0.246	0.691	0.829
		RWFMM	0.850	0.231	0.705	0.846
		$RWFMM_{90}$	0.865	0.215	0.727	0.855
		$RWFMM_{90}$-NoRE	0.843	0.242	0.712	0.821
Out Block	FMM	GWFMM	0.802	0.273	0.612	0.863
		$GWFMM_{90}$	0.815	0.254	0.655	0.855
		$GWFMM_{90}$-NoRE	0.818	0.246	0.698	0.821
		RWFMM	0.838	0.266	0.619	0.872
		$RWFMM_{90}$	0.830	0.242	0.705	0.829
		$RWFMM_{90}$-NoRE	0.814	0.266	0.698	0.786