Authors response to reviewers of the manuscript submitted to *Statistics and its Interface* (SII1407-004)

"Stratified Psychiatry via Convexity-Based Clustering with Applications Towards Moderator Analysis" (Tarpey, Petkova, and Zhu)

We thank the Editor, Associate Editor and the two referees for their careful assessment of our manuscript. We have edited the manuscript based on the reviews of our paper. Below we provide responses to the comments and questions from the reviewers.

Response to Reviewer 1

The reviewer pointed out six typographical errors. Thank you for carefully reading our manuscript. We corrected all typos the reviewer identified including a few other typos that we noticed while revising the paper.

Q1. *In section 3, the authors explain the convexity-based clustering in case of two sub-populations, where the squared difference between two posterior probabilities measures the strength for classifying observations into sub-populations. What quantity measures the strength in case of more than two sub-populations?*

A: This question raises an important point, which is how to generalize the partitioning method in the case of more than two groups (or treatments). The convexity-based clustering can be generalized to more than two groups by re-defining the original quantity λ to be vector valued which follows a general framework discussed in Bock (2003). Specifically, if there are say T groups, we can generalize the original definition of the scalar λ function to

$$\boldsymbol{\lambda}(\boldsymbol{x}) = (\lambda_1, \ldots, \lambda_T)' = (\frac{\pi_1 f_1(\boldsymbol{x})}{f(\boldsymbol{x})}, \ldots, \frac{\pi_T f_T(\boldsymbol{x})}{f(\boldsymbol{x})})'.$$

Similarly, the support points w_j in the convexity-based clustering become vector-valued:

$$\boldsymbol{w}_j = (\frac{\pi_1 P_1(B_j)}{P(B_j)}, \ldots, \frac{\pi_T P_T(B_j)}{P(B_j)})'.$$

The original convex function $\phi(x) = (1-2x)^2$ generalizes to

$$\phi(\boldsymbol{x}) = \phi(x_1, \ldots, x_T) = \sum_{h=1}^{T} (1-2x_h)^2.$$

The minimum support plane partition defined by the tangent function in equation (10) becomes

$$D_j = \{\boldsymbol{\lambda}(\boldsymbol{x}) : t(\boldsymbol{\lambda}(\boldsymbol{x}); \boldsymbol{w}_j) = \max_h t(\boldsymbol{\lambda}(\boldsymbol{x}); \boldsymbol{w}_h)\}.$$

In the original setting for two treatments, the D_j are intervals on the real line. In the more general setting of more than two treatments (or groups), essentially the same algebraic derivation shows that the sets D_j define a Voronoi partition in $\Re^{T-1}$ (the multivariate generalization of intervals in one-dimension). From these relations, following the same derivation, criterion (16) in the paper generalizes to finding a partition that maximizes

$$\mathcal{C} = \sum_{j=1}^{k} \sum_{h=1}^{T} \frac{(2\pi_h P_h(B_j) - P(B_j))^2}{P(B_j)},$$

which is equivalent to the original criterion (up to a multiplicative constant) when $T = 2$.

We have added a paragraph right before Section 3.1 providing a brief explanation of the more general setting of $T > 2$ groups.

Q2. *Will the sample size ratio between the two arms affect clustering or finding moderators?*

A: The quality of estimating the component densities f_h and the clustering (and hence finding treatment effect moderators) is of course enhanced with larger sample. In this regard, the clustering can be affected due to sample sizes. However, the ratio of the sizes of samples from different sub-populations will not affect the clustering if the true prior probabilities $\pi_t, t = 1, \ldots, T$, are known. If these quantities are estimated from the samples, the relative size of the sample will affect the clustering results.

Q3. *In case of drop-out, is it appropriate to apply imputation for estimating densities $f_1(x)$ and $f_2(x)$?*

A: We interpret this question to refer to the specific example presented in Section 5. The clustering in that section is based on the shapes of depression trajectories of subjects in the clinical trial. The subject-specific shapes are represented by their random intercept, slope and concavity obtained from mixed effects models with orthogonal quadratic polynomials. Best linear unbiased predictor (BLUP) of those quantities are obtainable for all patients under the assumption that conditional on the prior observations of the outcome, the missing data are missing at random. If there is a reason to suspect that that condition is violated, one might try imputing the missing outcomes based on more covariates than just the previous observations of the outcome. Producing multiple imputed data sets, one can then obtain the BLUPs of the random intercepts, slopes and concavity and apply the clustering algorithm to each imputed data set. Comparing the multiple clustering results would be a nice way of assessing the effect of the missing data on the clustering inferences.

We have added a brief paragraph to the discussion section to comment on this possibility.

Response to Reviewer 2

Q1. *In the section 5, the reason why the orthogonal-quadratic polynomial was used to fit the data should be discussed. Why not using Fourier analysis to fit function to data?*

A: The reviewer asks an important question, which prompted us to include the following paragraph after the first paragraph of Section 5, to clarify our decision:

"Other basis functions could have been used to fit the longitudinal trajectories (e.g. Fourier basis functions or B-splines). We choose to use orthogonal quadratic polynomials for the following reasons. (i) Most clustering algorithms are based on a minimal Euclidean distance. When the data are curves in function space, the L^2 distance

between curves corresponds to the usual Euclidean distance between regression coefficients when an orthonormal basis is used to fit the curves. Additionally, differences in clustering that occur due to the choice of the basis functions used to represent the curves are minimized when using an orthogonal basis function representation (Tarpey, 2007). (ii) Quadratic functions are easily interpretable and provide a good fit to the data over this relatively short longitudinal evaluation period (6 weeks). In particular, with orthogonal quadratic polynomials, the coefficient of the linear polynomial corresponds to the average quadratic slope of the parabola, which is an overall measure of improvement throughout the trial (Tarpey, 2003; Tarpey *et al.*, 2003). Also, the coefficient of the quadratic polynomial is a simple measure of the trajectory's curvature which has important interpretations in clinical settings, particularly when modeling placebo response."

Q2. *In the section 5, can you share the sensitivity, specificity and accuracy when applying a classification analysis to this data.*

A: This is an interesting question and we appreciate that the reviewer brought it up. The partitioning can be used to address various clinical questions, for which the proposed classification can have different sensitivity and specificity values. Here we illustrate the sensitivity and specificity due to the clustering with one question that has high significance for clinical practice. In particular, if a drug-treated subject responds to treatment, the clinician can ask if the response is due to the specific effects of the drug or due to non-specific placebo effects?

Fluoxetine is considered by many clinicians as a first line medication for treating depression. The typical course of treatment for an acute depressive episode is about 6 to 8 weeks, and patients who respond are urged to continue taking the medication, on average for another 6 to 9 months (Stewart *et al.*, 1998). Even though the side effects of antidepressants like Fluoxetine are mild, many patients prefer not to take

Table 1: Percent of Subjects Classified to Each Cluster

	Fluoxetine, $n = 196$			Placebo, $n = 162$		
Cluster	% Responders	% Non-Responders	% Total	% Responders	% Non-Responders	%Total
1	29	6	35	5	0	5
2	31	14	45	24	4	28
3	4	12	16	10	31	41
4	0	4	4	0	26	26
Overall	64	36	100	39	61	100

the drug, if their symptoms have improved. If a patient is responding to non-specific aspects of the treatment (i.e., a placebo responder (PR)), stopping the medication is not likely to affect his/her depression status. If, however, a patient responds to the specific aspects of the drug (i.e., s/he is specific drug responder (SDR)), stopping the medication too early will likely cause the patient to relapse to depression. Thus, if a clinician knows that the patient has responded to non-specific aspects of the treatment, i.e., the patient is PR, the clinician would be inclined to stop the antidepressant and schedule frequent visits with this patient to closely monitor symptoms increase and possible relapse. On the other hand, if the clinician is quite certain that the patient has had a specific response to the drug, i.e., the patient is SDR, the clinician would try to motivate the patient to continue taking the medication and warn the patient about the dangers of stopping treatment too soon.

Of course, in clinical practice subjects are always prescribed an active drug and if a patient responds, the clinician does not know whether the response is to specific or to non-specific aspects of the treatment. The overall results from the Fluoxetine/placebo clinical trial (Table 1 in the main text, also reproduced here) showed 64% response for drug-treated subjects and among them some would be SDRs and the rest would be PRs. Also there was 39% response among subjects treated with placebo and they all would be PRs. According to the randomized clinical trials convention, the

specific drug effect is 64% − 39% = 25%. Thus, of all drug-treated responders only 25/64 = 39% would be SDRs.

The clinician might try to identify those patients that responded to the specific aspects of the drug (i.e., SDRs) and urge them to stay on the medication, while allowing the remaining responders to discontinue treatment if they want. Without a "crystal ball" that tells the clinician which responders are SDR's and which are PR's, the clinician could make the decisions at random. In this scenario, if the decisions are made at random, one can expect that the clinician would randomly call 39% of the responders SDRs and motivate them to stay on the active medication. For the problem of classifying drug-treated responders as either SDR or PRs, such a procedure to identify SDR would have a sensitivity of 0.39, specificity 0.61 and an overall proportion of correct identification $P_{accurate} = 0.52$, see Table 2 in this response.

Table 2: Among drug-treated responders (64% of all drug-treated subjects), identifying specific drug responders (SDR) without the benefit of the clustering: sensitivity and specificity of the random selection of 39% of drug treated responders. (PR: placebo responders)

Drug Treated Responders Only		Truth (%) SDR	PR	Total
Call (%)	SDR	15	24	39
	PR	24	37	61
	Total	39	61	100
Sensitivity		15/39=.39		
Specificity		37/61=.61		
Accuracy		15+37=52%		

A simple alternative that utilizes the clustering results, is to consider all drug treated responders with trajectories in Cluster 1 as SDR and all other drug treated responders as PRs. Table 3 in this response presents some of the information from Table 1, combining clusters 2, 3 and 4 and conveniently rearranged. Notice that of all drug treated responders 45% are in Cluster 1 and the other 55% are in Clusters 2, 3 or 4, see column 4 of Table 3.

Table 3: Distribution of drug-treated responders (DTR) and placebo-treated responders (PTR) across clusters. their classification as specific drug responders (SDR) or drug-treated placebo responders (PR) overall, in Cluster 1 and in the combined Clusters 2, 3 and 4. $\%DRT = \%SDR + \%PR$.

	% Responders		% SDR of all drug treated subjects	% of DTR	Prevalence of SDR among DTR (%)
	Drug DTR	Pbo PTR			
Cluster 1	29	5	$29 - 5 = 24$	29/64=45	$24/29 = 0.83$
Clusters 2, 3 & 4	35	34	$35 - 34 = 1$	35/64=55	$1/35 = 0.03$
Overall	64	39	$64 - 39 = 25$	64/64=100	$25/64 = 0.39$

The first row of Table 3 corresponds to Cluster 1. Twenty nine percent of the overall 65% drug-treated responders are in Cluster 1. In Cluster 1, there are 5 of the total 39% placebo-treated responders. It stands to reason that of the 29% drug-treated responders with trajectories in Cluster 1, $29 - 5 = 24\%$ are likely to be SDRs, since 5% of the placebo-treated subjects responded and belong to the same cluster, see column 3 of Table 3. This results in a $24/29 = 0.83$ prevalence of SDR among drug-treated responders in Cluster 1, see column 5. Suppose we declare all drug treated responders with symptom curves in Cluster 1 as SDRs. For drug-treated subjects who's symptom curves fall in Cluster 1, the sensitivity of such procedure would be 1, the specificity would be 0, and the accuracy proportion would be $P_{accuracy} = 0.83$.

The second row of Table 3 shows the combined results for Clusters 2, 3 and 4. Thirty five of the total of 64% drug-treated responders, constituting 55% of all DTRs, are in those three clusters. Additionally, of the 39% placebo-treated responders, 34% are in these three clusters. Therefore, we can infer that among the 35% drug treated responders in Clusters 2, 3 and 4, only $(35 - 34)/35 = 1/35 = 3\%$ would be SDRs. Given this small percentage of SDRs in these clusters, we can classify all drug treated responders with symptom curves falling in these three clusters to be (primarily) PRs. For drug-treated responders, whose symptom curves fall in Clusters 2, 3 or 4, this

decision rule will have sensitivity 0, specificity 1, and accuracy proportion $P_{accuracy} = 0.97$. Note that overall then we have $29-5 = 24\%$ SDR in Cluster 1 plus $35-34 = 1\%$ SDR in Clusters 2, 3, and 4 giving a SDR rate of $(24+1)/64$ or 39% (see column 7 of Table 4). Taking into account the prevalence of drug-treated responders in Cluster 1 vs. Clusters 2, 3 and 4 (45% vs. 55%), the procedure for identifying SDR based on the clustering, has a sensitivity 0.96, specificity 0.87 and overall accuracy $P_{accuracy} = 0.91$. Table 4 summarizes this information.

Table 4: Among drug-treated responders (64% of all drug-treated subjects), identifying specific drug responders (SDR) using the clustering: sensitivity and specificity of a procedure that declares responders with symptom curves in Cluster 1 as SDRs and all other responders, as placebo responders (PR).

		Drug Treated Responders								
		Cluster 1 45%			Clusters 2, 3 & 4 55%			All 100%		
		Truth			Truth			Truth		
		SDR	PR	Total	SDR	PR	Total	SDR	PR	Total
Call	SDR	.83	.17	1	0	0	0	.3735	.0765	.45
	PR	0	0	0	.03	.97	1	.0165	.5335	.55
	Total	.83	.17	1	.03	.97	1	.39	.61	1
Sensitivity		$.83/.83 = 1$			$0/.03 = 0$			$.3735/.39 = .9577$		
Specificity		$0/.17 = 0$			$.97/.97 = 1$			$.5335/.61 = .8746$		
Accuracy		.83			.97			$.3735 + .5335 = .9070$		

Although this response provides substantial detail, we feel this explanation regarding sensitivity and specificity is a helpful motivation for the approach in the paper and would make a useful contribution to the paper. Thus, if the Editors and reviewers concur, we would be happy to include this material, for example by adding a subsection 6.1 and then putting the moderator importance plots in subsection 6.2.

Q3 *The completed code, program or R package should be provided as the supplementary so that the reader can apply and repeat similar study.*

A: We have prepared a package in R to perform the convexity-based clustering. We (the authors) will gladly make the code available to anyone that requests it.

References

Bock, H.-H. (2003). Convexity-based clustering criteria: theory, algorithms, and applications in statistics. *Statistical Methods & Applications* **12**:293–317.

Stewart, J. W., Quitkin, F. M., McGrath, P. J., Amsterdam, J., Fava, M., Fawcett, J., Reimherr, F., Rosenbaum, J., Beasley, C., and Roback, P. (1998). Use of pattern analysis to predict differrential relapse of remitted patients with major depression during 1 year of treatment with fluoxetine or placebo. *Archives of General Psychiatry* **55**:334–343.

Tarpey, T. (2003). Estimating the average slope. *Journal of Applied Statistics* **30**:389–395.

Tarpey, T. (2007). Linear transformations and the k-means clustering algorithm: Applications to clustering curves. *The American Statistician* **61**:34–40.

Tarpey, T., Petkova, E., and Ogden, R. T. (2003). Profiling placebo responders by self-consistent partitioning of functional data. *Journal of the American Statistical Association* **98**:850–858.