$\{\mathbf{XA}^{(i)}\}_{i=1}^{B_1}$ as in random-ensemble classification and derive the corresponding column basis $\{\mathbf{Q}^{(i)}\}_{i=1}^{B_1}$. Can we construct an aggregated column basis $\tilde{\mathbf{Q}}$ from $\{\mathbf{Q}^{(i)}\}_{i=1}^{B_1}$ such that $\tilde{\mathbf{Q}}$ will converge to the top k left singular space of $\mathbf{X}$ as B_1 increases to ∞?

Finally, we further stress an important advantage of the ensemble classifier proposed: its full adaptivity to the distributed computing architecture. To implement the random-projection ensemble classification in a distributed computing system, we first let each node computer solve for classification on randomly projected data. Then, according to the estimated risk of the base classifier on each node computer, we can screen out the good projections as described in Section 3 of the paper and construct the final ensemble classifier. Note that the algorithm does not require high communication cost since random projections are small.

Yang Feng (*Columbia University, New York*)

I congratulate Dr Cannings and Professor Samworth on their novel and stimulating contributions to classification using random-projection ensembles (RPEs). It is quite a general framework and we expect to see many follow-up works on the idea combined with some popular classifiers.

Regarding the choice of B_2, the authors did a careful theoretical analysis through assumption 2 and theorem 3. In assumption 2, I wonder whether β should depend on the sample size n or whether the authors believe that there is a universal β for all n. If β in fact turns out to decrease as n increases, we would need to conduct a more delicate analysis regarding the implications on the results of theorem 3 as $n \to \infty$.

Here, I propose a variant of the RPE approach. In this variant, the random projections are not generated independently; instead, the selected B_1 random projections are chosen sequentially and designed to be mutually orthogonal. The intuition is that, by making the random projections mutually orthogonal, the additional contribution of the newly recruited projections could be more significant than those without such constraints. I expect the variant to have a competitive performance when B_1 is small and the problem is high dimensional. A detailed modification is outlined as follows.

First, generate $\mathbf{A}_1$ the same way as the RPE. Now, suppose that we have found the projections $\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_k$, for some k. Then combine the corresponding random projections into the matrix $\mathbf{P}_k = (\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_k)_{p\times(dk)}$. To search for $\mathbf{A}_{k+1}$, first generate B_2 random projections $\{\tilde{\mathbf{A}}_{k+1,b_2}\}_{b_2=1}^{B_2}$ according to the Haar measure on $\mathcal{A}$, and then define $\mathbf{A}_{k+1,b_2} = (I - \mathbf{P}_k(\mathbf{P}_k^{\mathrm{T}}\mathbf{P}_k)^{-1}\mathbf{P}_k^{\mathrm{T}})\tilde{\mathbf{A}}_{k+1,b_2}$ as the orthogonal projection of $\tilde{\mathbf{A}}_{k+1,b_2}$ onto the space $\mathbf{P}_k^{\perp}$, which is the orthogonal complement of $\mathbf{P}_k$. Afterwards, we can follow the same procedure to find the optimal $\mathbf{A}_{k+1}$ by using the new random-projections candidates. At the ensemble step, I propose to use a weighted voting scheme based on the error rate on the test data $\{\mathrm{err}_{b_1}\}_{b_1=1}^{B_1}$ as follows:

$$v_n(x) := \frac{\sum_{b_1=1}^{B_1} w_{b_1} I\{C_n^{\mathbf{A}_{b_1}}(x) = 1\}}{\sum_{b_1=1}^{B_1} w_{b_1}},$$

where $w_{b_1} = \log\{(1 - \mathrm{err}_{b_1})/\mathrm{err}_{b_1}\}$. The final classifier can be created with a data-driven choice of the threshold α by taking into account the weights.

Michael P. B. Gallaugher and Paul D. McNicholas (*McMaster University, Hamilton*)

We congratulate Cannings and Samworth on a very well-written, enjoyable, and interesting contribution. Data collected today are often high dimensional and effective classification techniques for such data are most welcome. In the simulations and the real data analyses, the authors compare the proposed ensemble classifiers with the respective base classifiers as well as 'state of the art' techniques. We note the absence of mixture discriminant analysis, which was introduced in this self-same journal over 20 years ago (Hastie and Tibshirani, 1996) and subsequently studied by others (e.g. Fraley and Raftery (2002)). More general discriminant analysis techniques could also be considered, where a flexible non-Gaussian density is used for each class (see McNicholas (2016), section 9.2, for some discussion). It may also be interesting to consider discriminant analysis using a mixture of factor analysers model (Ghahramani and Hinton, 1997) or an extension thereof (see McNicholas (2016), chapter 3).

For brevity, we consider only mixture discriminant analysis, where the idea is to allow each class to be modelled by using a Gaussian mixture model. For the eye state data set, we take 10 training–test splits with 1000 observations in the training set, similar to the situation in the '$n = 1000$' column of Table 3. Using mixture discriminant analysis via the `mclust` package (Fraley *et al.*, 2017), we obtained an average misclassification rate, for the observations considered unlabelled, of around 0.18; this is a better result than two of the three random-projection classifiers considered. We also note that the mice data set contains