note, the recently proposed compressed regression approach by Guhaniyogi and Dunson (2015) is even closer to the authors' proposal, in the sense that it projects data into an ensemble of directions and uses model averaging to arrive at a final regression model. The focus of Guhaniyogi and Dunson (2015) is on regression itself though, but we also wonder about the authors' view on this. Finally, the practitioner could be left with the question: 'How likely is it for the ensemble classifier to improve over the base classifier on the original feature vectors?'.

Roberto Casarin and Lorenzo Frattarolo (*University Ca' Foscari of Venice*) **and Luca Rossini** (*University Ca' Foscari of Venice and Free University of Bozen-Bolzano*)

Cannings and Samworth are to be congratulated on their excellent research, which has culminated in the development of a characterization of the approximation errors in random-projection methods when applied to classification. We believe that the approach can find many applications in economics such as credit scoring (e.g. Altman (1968)) and can be extended to more general types of classifiers. In this discussion we would like to draw the authors' attention to copula-based discriminant analysis (Han *et al.*, 2013; He *et al.*, 2016).

We consider $X|Y=r$ distributed as a p-dimensional meta-Gaussian distribution and $S|Y=r \sim \mathcal{N}_p(0, \Sigma_r)$, where Σ_r is the linear correlation between variables. Given a $p \times d$ random projection A, $AS|Y=r \sim \mathcal{N}_d(0, \Sigma_r^A)$, where $\Sigma_r^A = A\Sigma_r A^{\mathrm{T}}$. If we assume that the information in the marginals is not relevant for the classification, the Bayes decision boundary depends only on the transformed variables $s_i = \Phi^{-1}\{F(x_i)\}$ with Φ and F the univariate normal and the marginal cumulative distribution functions respectively (Fang *et al.*, 2002), s_i and x_i the ith element of s and x, and the correlation of the two groups

$$\Delta(s; \pi_0, \Sigma_0, \Sigma_1) = \log\left(\frac{\pi_1}{\pi_0}\right) - \frac{1}{2}\log\left\{\frac{\det(\Sigma_1)}{\det(\Sigma_0)}\right\} - \frac{1}{2}s^{\mathrm{T}}(\Sigma_1^{-1} - \Sigma_0^{-1})s. \tag{46}$$

Analogously the classifier in the random-projection ensemble will depend only on the random projection of the transformed variables and their covariances. We use the empirical distribution function to obtain the sample version of the transformed variables $S_i = (S_{1i}, \ldots, S_{pi})$, with

$$S_{ji} = \Phi^{-1}\left(\frac{1}{n+1}\sum_{k=1}^{n} \mathbb{1}_{\{X_{jk} \leqslant X_{ji}\}}\right), \qquad i = 1, \ldots, n, \qquad j = 1, \ldots, p. \tag{47}$$

The estimator of Σ_r^A is obtained by maximizing the pseudolikelihood:

$$\hat{\Sigma}_r^A = \frac{1}{n}\sum_{i=1}^{n} AS_i S_i^{\mathrm{T}} A^{\mathrm{T}} \mathbb{1}_{\{Y_i^A = r\}} \qquad \text{for } r = 0, 1$$

where the asymptotic normality is guaranteed by results in Genest *et al.* (1995) and recently in Segers *et al.* (2014). We propose the following robust quadratic discriminant analysis random-projection ensemble classifier:

$$C_n^{\text{A-RQDA}}(s) := \begin{cases} 1 & \Delta(s; \hat{\pi}_0, \hat{\Sigma}_0^A, \hat{\Sigma}_1^A) \geqslant 0, \\ 0 & \text{otherwise.} \end{cases} \tag{48}$$

We are very pleased to thank the authors for their work.

Emre Demirkaya and Jinchi Lv (*University of Southern California, Los Angeles*)

Dr Cannings and Professor Samworth are to be congratulated for their innovative and valuable contribution to the important problem of high dimensional classification. Dimension reduction plays a key role in high dimensional classification, enabling the enhancement of both statistical efficiency and scalability (Fan and Fan, 2008). Through a simple yet ingenious two-level design of using random projections, Cannings and Samworth achieved these goals by proposing the general framework of random-projection ensemble classification with an elegant theory to deal with high dimensionality and to boost the power of existing classification procedures. The general philosophy of random-projection ensemble learning laid out in the paper can also be applicable to many other statistical learning tasks such as clustering and regression.

Our discussion will focus on the perspective of interaction network learning. Understanding large-scale interaction network structures among features can be of fundamental importance in many scientific studies. The problem of interaction network learning has received growing recent interest (Hall and Xue, 2014; Jiang and Liu, 2014; Fan *et al.*, 2015; Kong *et al.*, 2016). Recently, Fan *et al.* (2015)

Table 12. Percentages of retaining all important interactions in model 1 of Fan *et al.* (2015) by RAPID over various settings based on 100 replications when the threshold is chosen to be $[cn/\log(n)]$ following the suggestion in Fan and Lv (2008) with $n = 100$ the sample size for each class

c	*Results (%) for* $p = 100$		*Results (%) for* $p = 500$	
	$d = 5$	$d = 10$	$d = 5$	$d = 10$
0.5	0.99	0.97	0.93	0.95
1	1	0.97	0.97	0.95

introduced innovated interaction screening for high dimensional non-linear classification which depends on large precision matrix estimation. An interesting question is whether we can avoid estimating large precision matrices (Fan and Lv, 2016). To provide a partial answer to this question, we borrow the idea in the current paper and suggest a possible extension called random-projection interaction delineation (RAPID).

To illustrate the idea of RAPID, we adopt the framework in Fan *et al.* (2015) and consider a two-class Gaussian classification problem with heterogeneous precision matrices. In view of the Bayes rule, important interactions correspond to non-zero entries of precision matrix difference Ω. RAPID starts by randomly projecting p-dimensional feature vectors to low dimensions d and building classifiers with quadratic discriminant analysis following Cannings and Samworth. Each selected random projection returns a $d \times d$ symmetric matrix from the quadratic form, which can be lifted back to the original p dimensions through the given random projection. Each of B_1 such matrices can be used as a proxy for the original Ω. RAPID then evaluates the significance of each entry by using the t-statistics and ranks the interactions by the magnitude of these t-statistics. A simulation study shows that RAPID can enjoy a nice sure screening property (Fan and Lv, 2008) for interaction screening; see Table 12 for details. It would be interesting to investigate the theoretical properties of this and further extensions.

Josh Derenski, Yingying Fan and Gareth M. James (*University of Southern California, Los Angeles*)

Cannings and Samworth propose a method of classification involving many random projections of the data onto a lower dimensional space and then utilize a base classifier on the projected data to build an ensemble classification rule. They develop theoretical results involving arbitrary base classifiers and highlight the results when applied to particular base classifiers. In addition, they demonstrate the method's strong prediction accuracy with examples involving artificially generated data, and others involving real data.

The random-projection ensemble classifier may also be useful in determining the relative importance of the covariates. The authors suggest that the projections provide weights that can be used as a metric for determining the relative importance of variables. In a similar spirit, using sparse random projections may also assist in determining variable importance. Indeed, after the matrices have been generated and those that yield the smallest test error have been chosen, a variable is selected if the corresponding entries in the selected projection matrices are non-zero. The importance of a variable can be measured by, say, the frequency of the variable being selected.

The authors' proposed method has the flavour of a bagging algorithm, where the data are randomly sampled, a classifier is applied to each new data set and the results are averaged at the end. Hence, it is possible that prediction accuracy could be improved by applying a boosting-type approach. For example, rather than applying the same classifier to each random permutation, one could reweight the observations at each stage, placing higher weight on observations that were misclassified at the previous iteration. This would be somewhat analogous to standard boosting and would potentially provide a similar level of improvement in classification accuracy to that which boosting often has over bagging. Taking this theme one step further, one could choose the random projection conditionally on the performance of the classification method on the previous projection of the data, and then aggregate the results as in boosting.

The extensions suggested above also enable studying the random-projection ensemble classifier under different methodologies for choosing the projection matrices. The authors suggest the possibility of