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

choosing these matrices under different regimes, and there might not be a universally optimal way for selecting these matrices. For example, one sampling scheme might perform better when the goal is inference and another when the goal is prediction. Both these extensions enable the study of this possibility.

Robert J. Durrant (*University of Waikato, Hamilton*)

I thank Cannings and Samworth for an interesting paper, which I am sure will be of interest not only to statisticians but also to researchers in communities such as machine learning.

This paper is initially motivated by the Johnson–Lindenstrauss lemma (JLL), which gives high probability guarantees for the approximate preservation of Euclidean geometry of randomly projected data in $\mathbb{R}^d$ compared with the original data in the embedding space $\mathbb{R}^p$, $d \ll p$. Here I shall discuss some apparent implications of the JLL on the rejection sampling scheme for projection matrices described in this paper. In particular, it is my experience with random projection that for (linear) classification centring and normalizing a set of observations is usually a sensible preprocessing step to apply before random projection, and it appears that may be worthwhile here also. Below follows some informal argument supporting this view.

First note that projection using a sub-Gaussian random-projection matrix implies not only an ϵ–2δ guarantee on norm preservation, but also an ϵ–2δ guarantee on dot product preservation, i.e., under the same conditions as the JLL, for any $\epsilon, \delta \in (0, 1]$ with probability at least $1 - 2\delta$ over the random draws of $A \in \mathbb{R}^{d \times p}$ where the A_{ij} are independently and identically distributed sub-Gaussian with mean 0 and variance σ_A^2 it holds that

$$d\sigma_A^2 \cdot (v^{\mathrm{T}} w - \epsilon \|v\| \|w\|) \leqslant v^{\mathrm{T}} A^{\mathrm{T}} A w \leqslant d\sigma_A^2 \cdot (v^{\mathrm{T}} w + \epsilon \|v\| \|w\|).$$

For any fixed $v, w \in \mathbb{R}^p$ and random A this confidence interval depends on the Euclidean norms $\|v\|$ and $\|w\|$ independently of the angle between these vectors. Thus the JLL implies that, *even for two pairs of vectors with the same angle*, absent normalization, some dot products will be preserved better than others.

In particular, instantiating w as an observation and v as any unit norm classifier learned in $\mathbb{R}^p$, we see that observations with large norms are more likely to be classified differently with respect to v following projection to a fixed dimension $d < p$—i.e. by a sign change in the dot product—than those with small norms. Assuming v was reasonably accurate in the first place this means they will largely be *mis*classified for many instances of projection matrix A, and the corresponding projection matrix instances may risk being rejected—not necessarily because they fail to capture meaningful structure in the data (for the classification task)—instead because of systematic issues introduced by our choice of data representation. Thus it seems that it could be a reasonable step to add data normalization before projection to the authors' algorithm as described here.

Jianqing Fan and Ziwei Zhu (*Princeton University*)

We congratulate Dr Cannings and Professor Samworth for such a brilliant and thought-provoking paper. We believe that it will stimulate extensive research on statistical inference based on randomly projected data.

The authors aim to handle the curse of high dimensionality in classification problems through voting among multiple classifiers based on random data sketches. One of the most attractive aspects of their theories is that the excessive risk of the proposed ensemble classifier depends only on the dimension of the projected data d rather than the dimension of the original data p. To achieve this, the theory requires sufficient dimension reduction conditions. This exact low dimensional structure assumption can be sometimes stringent and some relaxations of the condition are welcome.

Besides overcoming the curse of dimensions, we emphasize that random projection is an accurate and efficient way of dimension reduction when data have (approximately) low dimensional structure. For example, consider the rank k approximation of $\mathbf{X} \in \mathbb{R}^{n \times p}$. Let $\mathbf{A} \in \mathbb{R}^{p \times (k+s)}$ be a random matrix with independently and identically distributed standard Gaussian entries and $\mathbf{Q} \in \mathbb{R}^{n \times (k+s)}$ be the orthonormal column basis of $\mathbf{XA}$. As shown in theorem 1 of Halko *et al.* (2011), for any $s > 1$,

$$E\|\mathbf{X} - \mathbf{Q}\mathbf{Q}^{\mathrm{T}}\mathbf{X}\|_{\mathrm{op}} \leqslant \left\{1 + \frac{4\sqrt{(k+s)}}{s-1}\sqrt{\min(n, p)}\right\}\sigma_{k+1},$$

where σ_{k+1} is the $(k+1)$th singular value of $\mathbf{X}$. Since σ_{k+1} is the theoretical minimum of rank k approximation error, this result implies that the column space of the random sketch $\mathbf{XA}$ can capture the top k left singular space of $\mathbf{X}$. It will thus be interesting to investigate the $\sin(\Theta)$ distance between the column space of $\mathbf{Q}$ and the top k left singular space of $\mathbf{X}$. Furthermore, suppose that we create B_1 independent sketches