Rebecca Nugent, Alessandro Rinaldo, Aarti Singh and Larry Wasserman (*Carnegie Mellon University, Pittsburgh*)

Meinshausen and Bühlmann argue for using stability-based methods. We suspect that the methods that are introduced in the current paper will generate much interest.

Stability methods have gained popularity lately. See Lange *et al.* (2004) and Ben-Hur *et al.* (2002) for example. There are cases where stability can lead to poor answers (Ben-David *et al.*, 2006). Some caution is needed.

General view of stability

Let $\{\hat{\theta}_h : h \in H\}$ be some class of procedures indexed by a tuning parameter h. We think of larger h as corresponding to larger bias. Our view of the stability approach is to use the least biased procedure subject to having an acceptable variability. This has a Neyman–Pearson flavour to it since we optimize what we cannot control subject to bounds on what we can control. The advantage is that variance is estimable whereas bias, generally, is not. There is no notion of approximating the 'truth' so it is not required that the model be correct. In contrast, Meinshausen and Bühlmann seem to be more focused on finding the 'true structure'.

Rinaldo and Wasserman (2010) applied this idea to finding stable density clusters as follows. Randomly split the data into three groups $X = (X_1, \ldots, X_n)$, $Y = (Y_1, \ldots, Y_n)$ and $Z = (Z_1, \ldots, Z_n)$. Construct a kernel density estimator $\hat{p}_h$ from X (with bandwidth h) and construct a kernel density estimator $\hat{q}_h$ from Y. Define the instability by

$$\Xi(h) = \hat{P}_Z(\{\hat{p}_h > \lambda\} \Delta \{\hat{q}_h > \lambda\})$$

where $\hat{P}_z$ is the empirical distribution based on Z. Under certain conditions, Rinaldo and Wasserman (2010) showed the following theorem.

Theorem 3. Let h_* be the diameter of $\{p > \lambda\}$ and let d be the dimension of the support of X_i. Then:

(a) $\Xi(0) = 0$ and $\Xi(h) = 0$, for all $h \geqslant h_*$;
(b) $\sup_{0<h<h_*}[\mathbb{E}\{\Xi(h)\}] < \frac{1}{2}$;
(c) As $h \to 0$, $\mathbb{E}\{\Xi(h)\} \asymp h^d$;
(d) for each $h \in (0, h_*)$,

$$c_1^n (h_* - h)^{d(n+1)} \leqslant \mathbb{E}\{\Xi(h)\} \leqslant 2c_2^n (h_* - h)^{n+1}$$

for constants c_1 and c_2.

We suggest using

$$\hat{h} = \inf[h : \sup_{t>h} \{\Xi(t)\} \leqslant \alpha], \tag{36}$$

where $\Xi(h)$ measures the variability and α is a user-defined acceptable amount of variability. Currently, we are generalizing the results to hold under weaker conditions and to hold uniformly over cluster trees rather than a single level set. The same ideas can be applied to graphs.

True structure?

The authors spend time discussing the search for true structure. In general, we feel that there is too much emphasis on finding true structure. Consider the linear model. It is a virtual certainty that the model is wrong. Nevertheless, we all use the linear model because it often leads to good predictions. The search for good predictors is much different from the search for true structure. The latter is not even well defined when the model is wrong, which it always is.

Adam J. Rothman, Elizaveta Levina and Ji Zhu (*University of Michigan, Ann Arbor*)

We congratulate the authors on developing a clever and practical method for improving high dimensional variable selection, and establishing an impressive array of theoretical performance guarantees. We are particularly interested in stability selection in graphical models, which is illustrated with one brief example in the paper. To investigate the performance of stability selection combined with the graphical lasso a little further, we performed the following simple simulation. The data are generated from the $N_p(0, \Omega^{-1})$ distribution, where $\Omega_{ii} = 1$, $\Omega_{i,i-1} = \Omega_{i-1,i} = 0.3$ and the rest are 0. We selected $p = 30$ and $n = 100$, and performed 50 replications. Stability selection with pointwise control was implemented with bootstrap samples of size $n/2$ drawn 100 times.

We selected four different values of the tuning parameter λ for the graphical lasso, which correspond

to the marked points along the receiver operating characteristic (ROC) curves for the graphical lasso in Fig. 18. The ROC curve showing false positive and true positive rates of detecting 0s in Ω for the graphical lasso was obtained by varying the tuning parameter λ and averaging over replications. For each fixed λ, we applied stability selection varying π_{thr} within the recommended range of 0.6–0.9, which resulted in an ROC curve for stability selection. The ROC curves show that stability selection reduces the false positive rate, as it should, and shifts the graphical lasso result down along the ROC curve; essentially, it is equivalent to the graphical lasso with a larger λ. Figs 18(a) and 18(b) have λs which are too small, and stability selection mostly improves on the graphical lasso result, but it does appear somewhat sensitive to the exact value of λ: if λ is very small (Fig. 18(a)), stability selection only improves on the graphical lasso for large values of π_{thr}. In Figs 18(c) and 18(d), λ is just right or too large, and then applying stability selection makes the overall result worse. This example confirms that stability selection is a useful computational tool to improve on the false positive rate of the graphical lasso when tuning over the

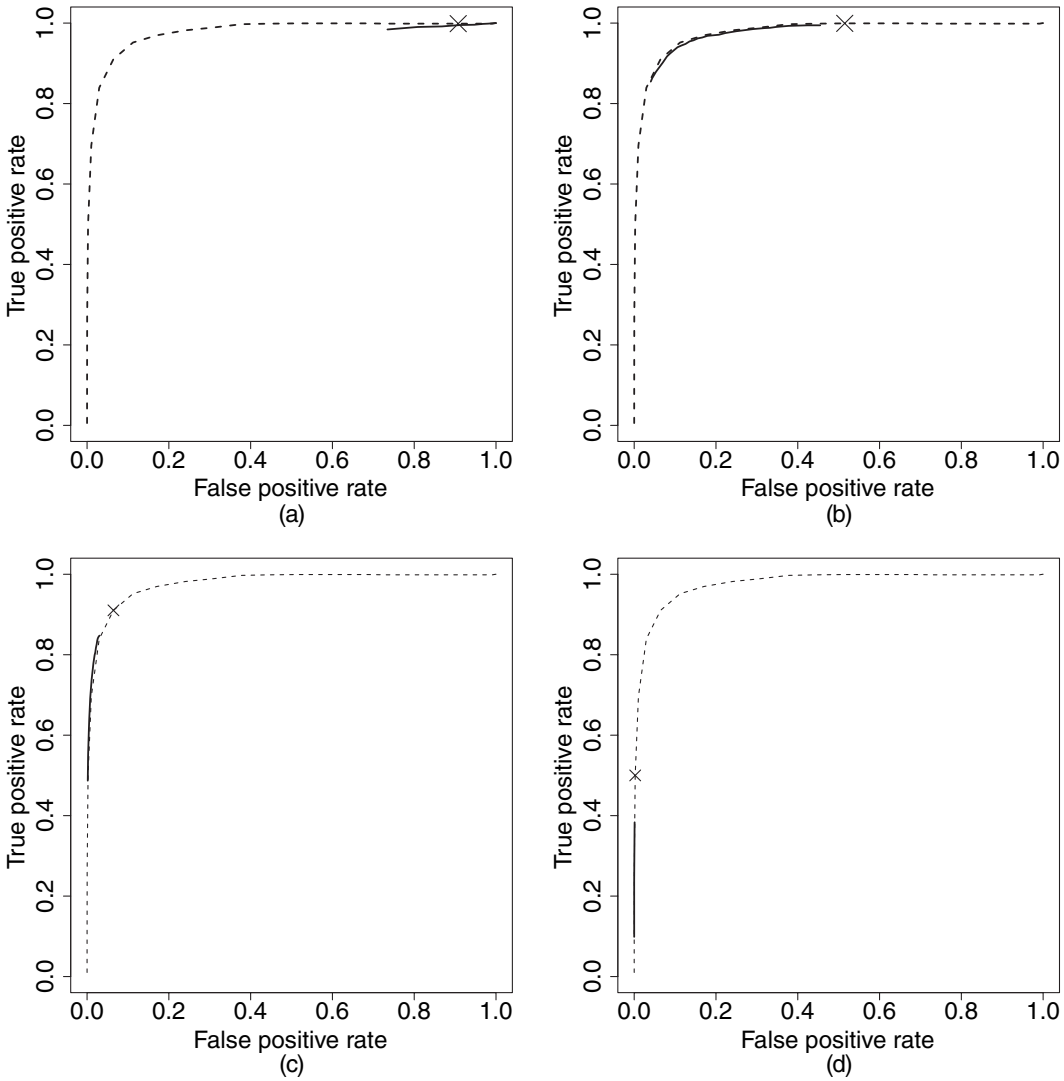

Fig. 18. Graphical lasso ROC curve (-------) and four different stability selection ROC curves (———) obtained by varying π_{thr} from 0.6 to 0.9 for fixed values of λ of (a) 0.01, (b) 0.06, (c) 0.23 and (d) 0.40: × marks the point on the graphical lasso ROC curve corresponding to the fixed λ

full range of λ is more expensive than doing bootstrap replications. However, since it does seem somewhat sensitive to the choice of a suitable small λ, it seems that combining it with some kind of initial crude cross-validation could result in even better performance. It would be interesting to consider whether there are particular types of the inverse covariance matrix that benefit from stability selection more than others, and whether any theoretical results can be obtained specifically for such structures; in particular, it would be interesting to know whether stability selection can perform better than the graphical lasso with oracle λ.

A. B. Tsybakov (*Centre de Recherche en Economie et Statistique, Université Paris 6 and Ecole Polytechnique, Paris*)

I congratulate the authors on a thought-provoking paper, which pioneers many interesting ideas. My question is about the comparison with other selection methods, such as the adaptive lasso or thresholded lasso (TL). In the theory these methods have better selection properties than those stated in theorem 2. For example, consider the TL $\tilde{\beta}_k = \hat{\beta}_k\, I\{|\hat{\beta}_k| > c\tau\sqrt{\|\hat{\beta}\|_0}\}$ where $\hat{\beta}$ is the lasso estimator with λ as in Bickel *et al.* (2009), $\tau = \sqrt{\{\log(p)/n\}}$ and $c > 0$ is such that $\|\hat{\beta} - \beta\|_2 \leqslant cs^{1/2}\tau$ with high probability under the restricted eigenvalue condition of Bickel *et al.* (2009). Then a two-line proof using expression (7.9) in Bickel *et al.* (2009) shows that, with the same probability, under the RE condition $\tilde{\beta}$ selects S correctly whenever $\min_{k \in S} |\beta_k| > Cs^{1/2}\tau$ for some $C > 0$ depending only on σ^2 and the eigenvalues of $X'X/n$. Since also c depends only on X and σ^2 (see Bickel *et al.* (2009)), c can be evaluated from the data. The restricted eigenvalue condition is substantially weaker than assumption 1 of theorem 2 and $\min_{k \in S} |\beta_k|$ need not be as large as greater than $C's^{3/2}\tau$, as required in theorem 2. We may interpret it as the fact that stability selection is successful if the relevant β_k are very large and the Gram matrix is very nice, whereas for smaller β_k and less diagonal Gram matrices it is safer to use the TL. Of course, here we compare only the 'upper bound', but it is not clear why stability selection does not achieve at least similar behaviour to that of the TL. Is it only technical or is there an intrinsic reason?

Cun-Hui Zhang (*Rutgers University, Piscataway*)

I congratulate the authors for their correct call for attention to the utility of randomized variable selection and great effort in studying its effectiveness.

In variable selection, a false variable may have a significant observed association with the response variable by representing a part of the realized noise through luck or by correlating with the true variables. A fundamental challenge in such structure estimation problems with high dimensional data is to deal with the competition of many such false variables for the attention of a statistical learning algorithm.

The solution proposed here is to simulate the selection probabilities of each variable with a randomized learning algorithm and to estimate the structure by choosing the variables with high simulated selection probabilities. The success of the proposed method in the numerical experiments is very impressive, especially in some cases at a level of difficulty that has rarely been touched on earlier. I applaud the authors for raising the bar for future numerical experiments in the field.

On the theoretical side, the paper considers two assumptions to guarantee the success of the method proposed:

(a) many false variables compete among themselves at random so each false variable has only a small chance of catching the attention of the randomized learning algorithm;
(b) the original randomized learning algorithm is not worse than random guessing.

The first assumption controls false discoveries whereas the second ensures a certain statistical power of detecting the true structure. Under these two assumptions, theorem 1 asserts in a broad context the validity of an upper bound for the total number of false discoveries. This result has the potential for an enormous influence, especially in biology, text mining and other areas that are overwhelmed with poorly understood large data.

Because of the potential for great influence of such a mathematical inequality in the practice of statistics, possibly by many non-statisticians, we must proceed with equally great caution. In this spirit, I comment on the two assumptions as follows.

Assumption (a) is the exchangeability condition in theorem 1. As mentioned in the paper, it is a consequence of the exchangeability of X_N given X_S in linear regression. The stronger condition implies a correlation structure for the design as

$$\begin{pmatrix} \Sigma_S & \mathrm{diag}(\rho_S)\mathbf{1} \\ \mathbf{1}\,\mathrm{diag}(\rho_S) & (1-\rho_N)I_N + \rho_N\mathbf{1} \end{pmatrix},$$