Journal of the Royal Statistical Society
SERIES B Statistical Methodology

J. R. Statist. Soc. B (2013)
75, *Part* 1, *pp.* 55–80

Variable selection with error control: another look at stability selection

Rajen D. Shah and Richard J. Samworth

University of Cambridge, UK

[Received May 2011. Revised January 2012]

Summary. Stability selection was recently introduced by Meinshausen and Bühlmann as a very general technique designed to improve the performance of a variable selection algorithm. It is based on aggregating the results of applying a selection procedure to subsamples of the data. We introduce a variant, called complementary pairs stability selection, and derive bounds both on the expected number of variables included by complementary pairs stability selection that have low selection probability under the original procedure, and on the expected number of high selection probability variables that are excluded. These results require no (e.g. exchangeability) assumptions on the underlying model or on the quality of the original selection procedure. Under reasonable shape restrictions, the bounds can be further tightened, yielding improved error control, and therefore increasing the applicability of the methodology.

Keywords: Complementary pairs stability selection; r-concavity; Subagging; Subsampling; Variable selection

1. Introduction

The problem of variable selection has received a huge amount of attention over the last 15 years, motivated by the desire to understand structure in massive data sets that are now routinely encountered across many scientific disciplines. It is now very common, e.g. in biological applications, image analysis and portfolio allocation problems as well as many others, for the number of variables (or predictors) p that are measured to exceed the number of observations n. In such circumstances, variable selection is essential for model interpretation.

In a notable recent contribution to the now vast literature on this topic, Meinshausen and Bühlmann (2010) proposed stability selection as a very general technique designed to improve the performance of a variable selection algorithm. The basic idea is that, instead of applying one's favourite algorithm to the whole data set to determine the selected set of variables, one instead applies it several times to random subsamples of the data of size $\lfloor n/2 \rfloor$ and chooses those variables that are selected most frequently on the subsamples. Stability selection is therefore intimately connected with bagging (Breiman, 1996, 1999) and subagging (Bühlmann and Yu, 2002).

A particularly attractive feature of stability selection is the error control that is provided by an upper bound on the expected number of falsely selected variables (Meinshausen and Bühlmann (2010), theorem 1). Such control is typically unavailable when applying the original selection procedure to the whole data set and allows the practitioner to select the threshold τ for the proportion of subsamples for which a variable must be selected for it to be declared significant.

Address for correspondence: Richard Samworth, Statistical Laboratory, Centre for Mathematical Sciences, University of Cambridge, Wilberforce Road, Cambridge, CB3 0WB, UK.
E-mail: r.j.samworth@statslab.cam.ac.uk

 1369–7412/13/75055

However, the bound does have a couple of drawbacks. Firstly, it applies to the 'population version' of the subsampling process, i.e. to the version of the procedure that aggregates results over the non-random choice of all $\binom{n}{\lfloor n/2 \rfloor}$ subsamples. Even for n as small as 15, it is unrealistic to expect this version to be used in practice, and in fact choosing around 100 random subsamples is probably typical. More seriously, the bound is derived under a very strong exchangeability assumption on the selection of noise variables (as well as a weak assumption on the quality of the original selection procedure, namely that it is not worse than random guessing).

In this paper, we develop the methodology and conceptual understanding of stability selection in several respects. We introduce a variant of stability selection, where the subsamples are drawn as complementary pairs from $\{1, \ldots, n\}$. Thus the subsampling procedure outputs index sets $\{(A_{2j-1}, A_{2j}) : j = 1, \ldots, B\}$, where each A_j is a subset of $\{1, \ldots, n\}$ of size $\lfloor n/2 \rfloor$, and $A_{2j-1} \cap A_{2j} = \emptyset$. We call this variant complementary pairs stability selection (CPSS).

At first glance it would seem that CPSS would be expected to yield very similar results to the original version of stability selection. However, we show that CPSS in fact has the following properties.

(a) The Meinshausen–Bühlmann bound holds for CPSS regardless of the number of complementary pairs B chosen—even with $B = 1$.
(b) There is a corresponding bound for the number of important variables excluded by CPSS.
(c) Our results have no conditions on the original selection procedure and in particular do not require the strong exchangeability assumption on the selection of noise variables. Indeed, we argue that even a precise definition of 'signal' and 'noise' variables is not helpful in trying to understand the properties of CPSS, and we instead state the bounds in terms of the expected number of variables chosen by CPSS that have low selection probability under the base selection procedure, and the expected number of high selection probability variables that are excluded by CPSS. See Section 2 for further discussion.
(d) The bound on the number of low selection probability variables chosen by CPSS can be significantly sharpened under mild shape restrictions (e.g. unimodality or r-concavity) on the distribution of the proportion of times that a variable is selected in both A_{2j-1} and A_{2j}. We discuss these conditions in detail in Sections 3.2 and 3.3 respectively and compare both the original and the new bounds to demonstrate the marked improvement.

Our improved bounds are based on new versions of Markov's inequality that hold for random variables whose distributions are unimodal or r-concave. However, it is important to note at this point that the results are not just a theoretical contribution; they allow the practitioner to reduce τ (and therefore to select more variables) for the same control of the number of low selection probability variables chosen by CPSS. In Section 3.4, we give recommendations on how a practitioner can make use of the bounds in applying CPSS.

In Section 4.1, we present the results of an extensive simulation study that was designed to illustrate the appropriateness of our shape restrictions, and to compare stability selection and CPSS with their base selection procedures. Section 4.2 gives an application of the methodology to a colon cancer data set.

A review of some of the extensive literature on variable selection can be found in Fan and Lv (2010). Work that is related more specifically to stability selection includes Bach (2008), who studied the 'bolasso' (short for bootstrapped enhanced lasso). This involves applying the lasso to bootstrap (with replacement) samples from the original data, rather than subsampling without replacement. A final estimate is obtained by applying the lasso to the intersection of the set of variables selected across the bootstrap samples. Various researchers, particularly in the machine learning literature, have considered the *stability* of a feature selection algorithm, i.e. the

insensitivity of the output of the algorithm to variations in the training set; such studies include Lange *et al.* (2003), Kalousis *et al.* (2007), Kuncheva (2007), Loscalzo *et al.* (2009) and Han and Yu (2010). Saeys *et al.* (2008) considered obtaining a final feature ranking by aggregating the rankings across bootstrap samples.

2. Complementary pairs stability selection

To keep our discussion rather general, we assume only that we have vector-valued data $z_1, \ldots, z_n$ which we take to be a realization of independent and identically distributed random elements $Z_1, \ldots, Z_n$. Informally, we think of some of the components of Z_i as being 'signal variables', and others as being 'noise variables', though for our purposes it is not necessary to define these notions precisely. Formally, we let $S \subseteq \{1, \ldots, p\}$ and $N := \{1, \ldots, p\} \backslash S$, thought of as the index sets of the signal and noise variables respectively. A *variable selection procedure* is a statistic $\hat{S}_n := \hat{S}_n(Z_1, \ldots, Z_n)$ taking values in the set of all subsets of $\{1, \ldots, p\}$, and we think of $\hat{S}_n$ as an estimator of S. As a typical example, we may often write $Z_i = (X_i, Y_i)$ with the covariate $X_i \in \mathbb{R}^p$ and the response $Y_i \in \mathbb{R}$, and our (pseudo-) log-likelihood might be of the form

$$\sum_{i=1}^{n} L(Y_i, X_i^{\mathrm{T}} \beta), \tag{1}$$

for some $\beta \in \mathbb{R}^p$. In this context, we regard $S := \{k : \beta_k \neq 0\}$ as the signal indices, and $N = \{k : \beta_k = 0\}$ as noise indices. Examples from graphical modelling can also be cast within our framework. Note, however, that we do not require a (pseudo-) log-likelihood of the form (1).

We define the selection probability of a variable index $k \in \{1, \ldots, p\}$ under $\hat{S}_n$ as

$$p_{k,n} = \mathbb{P}(k \in \hat{S}_n) = \mathbb{E}(\mathbb{1}_{\{k \in \hat{S}_n\}}). \tag{2}$$

We take the view that, for understanding the properties of stability selection, the selection probabilities $p_{k,n}$ are the fundamental quantities of interest. Since an application of stability selection is contingent on a choice of base selection procedure $\hat{S}_n$, all we can hope is that it selects variables having high selection probability under the base procedure and avoids selecting those variables with low selection probability. Indeed this turns out to be so; see theorem 1 below.

Of course, $\mathbb{1}_{\{k \in \hat{S}_n\}}$ has a Bernoulli distribution with parameter $p_{k,n}$, so we may view $\mathbb{1}_{\{k \in \hat{S}_n\}}$ as an unbiased estimator of $p_{k,n}$ (though $p_{k,n}$ is not a model parameter in the conventional sense). The key idea of stability selection is to improve on this simple estimator of $p_{k,n}$ through subsampling.

For a subset $A = \{i_1, \ldots, i_{|A|}\} \subset \{1, \ldots, n\}$ with $i_i < \ldots < i_{|A|}$, we shall write

$$\hat{S}(A) := \hat{S}_{|A|}(Z_{i_1}, \ldots, Z_{i_{|A|}}).$$

Definition 1 (CPSS). Let $\{(A_{2j-1}, A_{2j}) : j = 1, \ldots, B\}$ be randomly chosen independent pairs of subsets of $\{1, \ldots, n\}$ of size $\lfloor n/2 \rfloor$ such that $A_{2j-1} \cap A_{2j} = \emptyset$. For $\tau \in [0, 1]$, the CPSS version of a variable selection procedure $\hat{S}_n$ is $\hat{S}_{n,\tau}^{\mathrm{CPSS}} = \{k : \hat{\Pi}_B(k) \geqslant \tau\}$, where the function $\hat{\Pi}_B : \{1, \ldots, p\} \to \{0, 1/(2B), 1/B, \ldots, 1\}$ is given by

$$\hat{\Pi}_B(k) := \frac{1}{2B} \sum_{j=1}^{2B} \mathbb{1}_{\{k \in \hat{S}(A_j)\}}. \tag{3}$$

Note that $\hat{\Pi}_B(k)$ is an unbiased estimator of $p_{k,\lfloor n/2 \rfloor}$, but, in general, a biased estimator of $p_{k,n}$. However, by means of the averaging that is involved in expression (3), we hope that $\hat{\Pi}_B(k)$ will have reduced variance compared with $\mathbb{1}_{\{k \in \hat{S}_n\}}$, and that this increased stability will more than compen-

sate for the bias incurred. Indeed, this is so in other situations where bagging and subagging have been successfully applied, such as classification trees (Breiman, 1996) or nearest neighbour classifiers (Hall and Samworth, 2005; Biau *et al.*, 2010; Samworth, 2011).

An alternative to subsampling complementary pairs would be to use bootstrap sampling. We have found that this gives very similar estimates of $p_{k,n}$, though most of our theoretical arguments do not apply when the bootstrap is used (the approach in Section 3.3.1 is an exception in this regard). In fact, taking subsamples of size $\lfloor n/2 \rfloor$ can be thought of as the subsampling scheme that most closely mimics the bootstrap (e.g. Dümbgen *et al.* (2012)).

It is convenient at this stage to define another related selection procedure based on sample splitting.

Definition 2 (simultaneous selection). Let $\{(A_{2j-1}, A_{2j}) : j = 1, \ldots, B\}$ be randomly chosen independent pairs of subsets of $\{1, \ldots, n\}$ of size $\lfloor n/2 \rfloor$ such that $A_{2j-1} \cap A_{2j} = \emptyset$. For $\tau \in [0, 1]$, the simultaneous selection version of $\hat{S}_n$ is $\hat{S}^{\mathrm{SIM}}_{n,\tau} = \{k : \tilde{\Pi}_B(k) \geqslant \tau\}$, where

$$\tilde{\Pi}_B(k) := \frac{1}{B} \sum_{j=1}^{B} \mathbb{1}_{\{k \in \hat{S}(A_{2j-1})\}} \mathbb{1}_{\{k \in \hat{S}(A_{2j})\}}. \tag{4}$$

For our purposes, simultaneous selection is a tool for understanding the properties of CPSS. However, the special case of $B = 1$ of simultaneous selection was studied by Fan *et al.* (2009), and a variant involving all possible disjoint pairs of subsets was considered in Meinshausen and Bühlmann (2010).

3. Theoretical properties

3.1. Worst-case bounds

In theorem 1 below, we show that the expected number of low selection probability variables chosen by CPSS is controlled in terms of the expected number chosen by the original selection procedure, with a corresponding result for the expected number of high selection probability variables not chosen by CPSS. The appealing feature of these results is their generality: they require no assumptions on the underlying model or on the quality of the original selection procedure, and they apply regardless of the number B of complementary pairs of subsets chosen.

For $\theta \in [0, 1]$, let $L_\theta = \{k : p_{k,\lfloor n/2 \rfloor} \leqslant \theta\}$ denote the set of variable indices that have low selection probability under $\hat{S}_{\lfloor n/2 \rfloor}$, and let $H_\theta = \{k : p_{k,\lfloor n/2 \rfloor} > \theta\}$ denote the set of those that have high selection probability.

Theorem 1.

(a) If $\tau \in (\frac{1}{2}, 1]$, then

$$\mathbb{E}|\hat{S}^{\mathrm{CPSS}}_{n,\tau} \cap L_\theta| \leqslant \frac{\theta}{2\tau - 1} \mathbb{E}|\hat{S}_{\lfloor n/2 \rfloor} \cap L_\theta|.$$

(b) Let $\hat{N}^{\mathrm{CPSS}}_{n,\tau} = \{1, \ldots, p\} \backslash \hat{S}^{\mathrm{CPSS}}_{n,\tau}$ and $\hat{N}_n = \{1, \ldots, p\} \backslash \hat{S}_n$. If $\tau \in [0, \frac{1}{2})$, then

$$\mathbb{E}|\hat{N}^{\mathrm{CPSS}}_{n,\tau} \cap H_\theta| \leqslant \frac{1 - \theta}{1 - 2\tau} \mathbb{E}|\hat{N}_{\lfloor n/2 \rfloor} \cap H_\theta|.$$

In many applications, and for a good base selection procedure, we imagine that the set of selection probabilities $\{p_{k,\lfloor n/2 \rfloor} : k = 1, \ldots, p\}$ is positively skewed in $[0, 1]$, with many selection probabilities being very low (predominantly noise variables), and with just a few being large

(including at least some of the signal variables). To illustrate theorem 1, part (a), consider a situation with $p = 1000$ variables and where the base selection procedure chooses 50 of them. Then theorem 1, part (a), shows that on average CPSS with $\tau = 0.6$ selects no more than a quarter of the below-average selection probability variables chosen by $\hat{S}_{\lfloor n/2 \rfloor}$.

Our theorem 1, part (a), is analogous to theorem 1 of Meinshausen and Bühlmann (2010). The differences are that we do not require the condition that $\{\mathbb{1}_{\{k \in \hat{S}_{\lfloor n/2 \rfloor}\}} : k \in N\}$ is exchangeable, nor that the original procedure is no worse than random guessing, and our result holds for all B. The price that we pay is that the bound is stated in terms of the expected number of low selection probability variables chosen by CPSS, rather than the expected number of noise variables, which we do for the reasons that were described in Section 2. If the exchangeability and random guessing conditions mentioned above do hold, then, writing $q := \mathbb{E}|\hat{S}_{\lfloor n/2 \rfloor}|$, we recover

$$\mathbb{E}|\hat{S}^{\mathrm{CPSS}}_{n,\tau} \cap N| \leqslant \frac{1}{2\tau - 1} \frac{q}{p} \mathbb{E}|\hat{S}_{\lfloor n/2 \rfloor} \cap L_{q/p}| \leqslant \frac{1}{2\tau - 1} \frac{q^2}{p}.$$

The final bound here was obtained in theorem 1 of Meinshausen and Bühlmann (2010) for the population version of stability selection.

3.2. *Improved bounds under unimodality*

Despite the attractions of theorem 1, the following observations suggest that there may be scope for improvement. Firstly, we expect we should be able to obtain tighter bounds as B increases. Secondly, and more importantly, examination of the proof of theorem 1, part (a), shows that our bound relies on first noting that

$$1 + \tilde{\Pi}_B(k) \geqslant 2\hat{\Pi}_B(k), \tag{5}$$

and then applying Markov's inequality to $\tilde{\Pi}_B(k)$. For equality in Markov's inequality, $\tilde{\Pi}_B(k)$ must be a mixture of point masses at 0 and $2\tau - 1$, but Fig. 1 suggests that the distribution of $\tilde{\Pi}_B(k)$, which is supported on $\{0, 1/B, 2/B, \ldots, 1\}$, can be very different from this. Indeed, our experience, based on extensive simulation studies, is that when θ is close to q/p (which is where the bound in theorem 1, part (a), is probably of most interest), the distribution of $\tilde{\Pi}_B(k)$ over $k \in L_\theta$ is remarkably consistent over different data-generating processes, and Fig. 1 is typical. It is therefore natural to consider placing shape restrictions on the distribution of $\tilde{\Pi}_B(k)$ which encompass what we see in practice, and which yield stronger versions of Markov's inequality. As a first step in this direction, we consider the assumption of unimodality.

Theorem 2. Suppose that the distribution of $\tilde{\Pi}_B(k)$ is unimodal for each $k \in L_\theta$. If $\tau \in \{\frac{1}{2} + 1/B, \frac{1}{2} + 3/(2B), \frac{1}{2} + 2/B, \ldots, 1\}$, then

$$\mathbb{E}|\hat{S}^{\mathrm{CPSS}}_{n,\tau} \cap L_\theta| \leqslant C(\tau, B)\theta\, \mathbb{E}|\hat{S}_{\lfloor n/2 \rfloor} \cap L_\theta|,$$

where, when $\theta \leqslant 1/\sqrt{3}$,

$$C(\tau, B) = \begin{cases} \dfrac{1}{2\{2\tau - 1 - 1/(2B)\}} & \text{if } \tau \in (\min\{\frac{1}{2} + \theta^2, \frac{1}{2} + 1/(2B) + \frac{3}{4}\theta^2\}, \frac{3}{4}] \\ \dfrac{4\{1 - \tau + 1/(2B)\}}{1 + 1/B} & \text{if } \tau \in (\frac{3}{4}, 1]. \end{cases}.$$

The proof of theorem 2 is based on a new version of Markov's inequality (theorem 3 in Appendix A) for random variables with unimodal distributions supported on a finite lattice. There is also an explicit expression for $C(\tau, B)$ when $\theta > 1/\sqrt{3}$, which follows from theorem 3 in the same way,

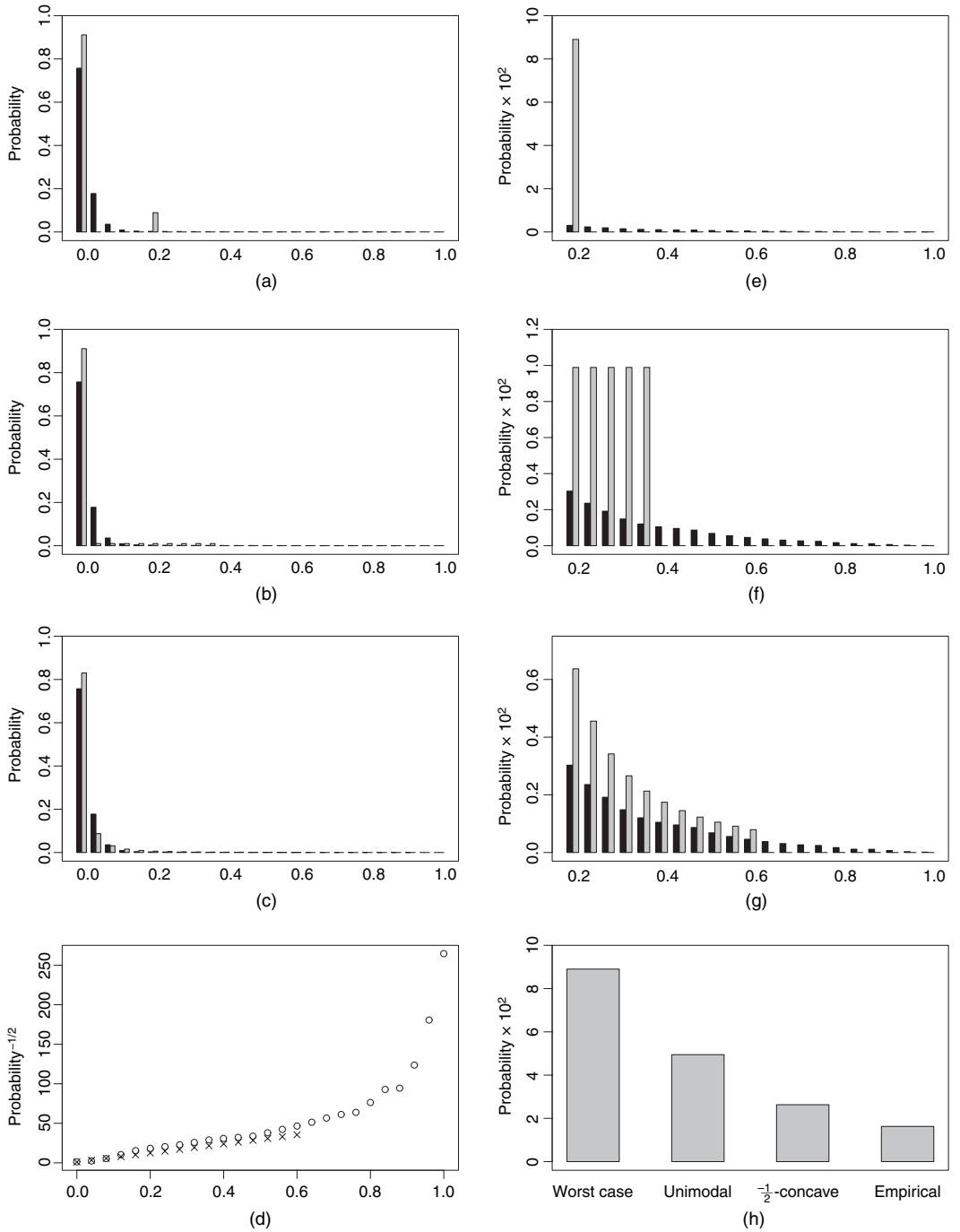

Fig. 1. Typical example of (a)–(c) the full probability mass function and (e)–(g) zoomed in from 0.2 onwards of $\tilde{\Pi}_{25}(k)$ for $k \in L_{g/p}$ (▮), alongside (a), (e) the unrestricted, (b), (f) unimodal and (c), (g) $-\frac{1}{2}$-concave distributions (▮), which have maximum tail probability beyond 0.2 (this situation corresponds to selecting $\tau = 0.6$), and (d) the observed mass function (○) and the extremal $-\frac{1}{2}$-concave mass function (×) on the $x^{-1/2}$-scale and (h) tail probabilities from 0.2 onwards for each of the distributions

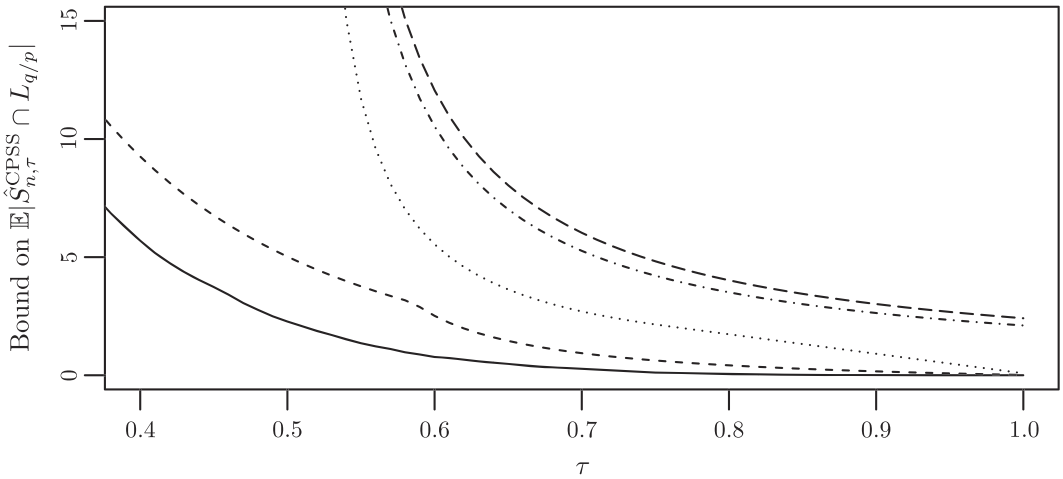

Fig. 2. Comparison of the bounds on $\mathbb{E}|\hat{S}^{\mathrm{CPSS}}_{n,\tau} \cap L_{q/p}|$ for various values of the threshold τ: – – –, original bound from theorem 1 of Meinshausen and Bühlmann (2010); - - - -, our worst-case bound; · · · · · · ·, unimodal bound; -------, *r*-concave bound (8); ———, true value of $\mathbb{E}|\hat{S}^{\mathrm{CPSS}}_{n,\tau} \cap L_{q/p}|$ for a simulated example (in this case $p = 1000$, $q = 50$ and the number of signal variables was 8)

but we do not present it here because it is a little more complicated, and because we anticipate the bound when θ is (much) smaller than $1/\sqrt{3}$ being of most use in practice. See Section 3.4 for further discussion.

Fig. 2 compares the bounds that are provided by theorem 1 and theorem 2 as a function of τ, for the illustration discussed after the statement of theorem 1.

3.3. Further improvements under r-concavity

The unimodal assumption allows for a significant improvement in the bounds that are attainable from a naive application of Markov's inequality. However, Fig. 1 suggests that further gains may be realized by placing tighter constraints on the family of distributions for $\tilde{\Pi}_B(k)$ that we consider, to match better the empirical distributions that we see in practice.

A very natural constraint to impose on the distribution of $\tilde{\Pi}_B(k)$ is log-concavity. By this, we mean that, if f denotes the probability mass function of $\tilde{\Pi}_B(k)$, then the linear interpolant to $\{(i, f(i/B)) : i = 0, 1, \ldots, B\}$ is a log-concave function on $[0, 1]$. Log-concavity is a shape constraint that has received a large amount of attention recently (e.g. Walther (2002), Dümbgen and Rufibach (2009) and Cule *et al.* (2010)), and at first sight it seems reasonable in our context, because, if the summands in expression (4) were independent, then we would have $\tilde{\Pi}_B(k) \sim (1/B)\,\mathrm{Bin}(B, p^2_{k,\lfloor n/2 \rfloor})$, which is log-concave.

It is indeed possible to obtain a version of Markov's inequality under log-concavity that leads to another improvement in the bound on $\mathbb{E}|\hat{S}^{\mathrm{CPSS}}_{n,\tau} \cap L_\theta|$. However, we found that, in practice, the dependence structure of the summands in expression (4) meant that the log-concavity constraint was a little too strong. We therefore consider instead the class of *r-concave* distributions, which we claim defines a continuum of constraints that interpolate between log-concavity and unimodality (see propositions 1 and 2 below). This constraint has also been studied recently in the context of density estimation by Seregin and Wellner (2010) and Koenker and Mizera (2010); see also Dharmadhikari and Joag-Dev (1988).

To define the class, we recall that the rth generalized mean $M_r(a, b; \lambda)$ of $a, b \geqslant 0$ is given by

$$M_r(a, b; \lambda) = \{(1 - \lambda)a^r + \lambda b^r\}^{1/r}$$

for $r > 0$. This is also well defined for $r < 0$ if we take $M_r(a, b; \lambda) = 0$ when $ab = 0$, and define $0^r = \infty$. In addition, we may define

$$M_0(a, b; \lambda) := \lim_{r \to 0} M_r(a, b; \lambda) = a^{1-\lambda} b^{\lambda},$$

$$M_{-\infty}(a, b; \lambda) := \lim_{r \to -\infty} M_r(a, b; \lambda) = \min(a, b).$$

We can now define r-concavity.

Definition 3. A non-negative function f on an interval $I \subset \mathbb{R}$ is r concave if, for every $x, y \in I$ and $\lambda \in (0, 1)$, we have

$$f\{(1-\lambda)x + \lambda y\} \geqslant M_r\{f(x), f(y); \lambda\}.$$

Definition 4. A probability mass function f supported on $\{0, 1/B, 2/B, \ldots, 1\}$ is r concave if the linear interpolant to $\{(i, f(i/B)) : i = 0, 1, \ldots, B\}$ is r concave.

When $r < 0$, it is easy to see that f is r concave if and only if f^r is convex. Let $\mathcal{F}_r$ denote the class of r-concave probability mass functions on $\{0, 1/B, 2/B, \ldots, 1\}$. Then each $f \in \mathcal{F}_r$ is unimodal and, as $M_r(a, b; \lambda)$ is non-decreasing in r for fixed a and b, we have $\mathcal{F}_r \supset \mathcal{F}_{r'}$ for $r < r'$. Furthermore, f is unimodal if it is $-\infty$ concave, and f is log-concave if it is 0 concave. The following two results further support the interpretation of r-concavity for $r \in [-\infty, 0]$ as an interpolation between log-concavity and unimodality.

Proposition 1. A function f is log-concave if and only if it is r concave for every $r < 0$.

Proposition 2. Let f be a unimodal probability mass function supported on $\{0, 1/B, 2/B, \ldots, 1\}$ and suppose both that $f(0) < \ldots < f(l/B) = f\{(l+1)/B\} = \ldots = f(u/B)$ and that $f(u/B) > f\{(u+1)/B\} > \ldots > f(1)$, for some $l \leqslant u$. Then f is r concave for some $r < 0$.

In proposition 5 in Appendix A, we present a result that characterizes those r-concave distributions that attain equality in a version of Markov's inequality for random variables with r-concave distributions on $\{0, 1/B, 2/B, \ldots, 1\}$. If we assume that $\tilde{\Pi}_B(k)$ is r concave for all $k \in L_\theta$, using inequality (5), for these variables we can obtain a bound of the form

$$\mathbb{P}\{\hat{\Pi}_B(k) \geqslant \tau\} \leqslant D(p^2_{k,\lfloor n/2 \rfloor}, 2\tau - 1, B, r) \leqslant D(\theta^2, 2\tau - 1, B, r) \tag{6}$$

where $D(\eta, t, B, r)$ denotes the maximum of $\mathbb{P}(X \geqslant t)$ over all r-concave random variables supported on $\{0, 1/B, 2/B, \ldots, 1\}$ with $\mathbb{E}(X) \leqslant \eta$. Although D does not appear to have a closed form, it is straightforward to compute numerically, as we describe in Appendix A.4. The lack of a simple form means that a direct analogue of theorem 2 is not available. We can nevertheless obtain the following bound on the expected number of low selection probability variables chosen by CPSS:

$$\mathbb{E}|\hat{S}^{\mathrm{CPSS}}_{n,\tau} \cap L_\theta| = \sum_{k \in L_\theta} \mathbb{P}\{\hat{\Pi}_B(k) \geqslant \tau\} \leqslant D(\theta^2, 2\tau - 1, B, r)|L_\theta|. \tag{7}$$

Our simulation studies suggest that $r = -\frac{1}{2}$ is a sensible choice to use for the bound. In other words, if f denotes the probability mass function of $\tilde{\Pi}_B(k)$, then the linear interpolant to $\{(i, f(i/B)^{-1/2}) : i = 0, 1, \ldots, B\}$ is typically well approximated by a convex function. This is illustrated in Fig. 1(d) (note that the right-hand tail in this plot corresponds to tiny probabilities).

3.3.1. *Lowering the threshold τ*

The bounds obtained thus far have used the relationship (5) to convert a Markov bound for $\tilde{\Pi}_B(k)$ into a corresponding bound for the statistic of interest, $\hat{\Pi}_B(k)$. The advantage of this

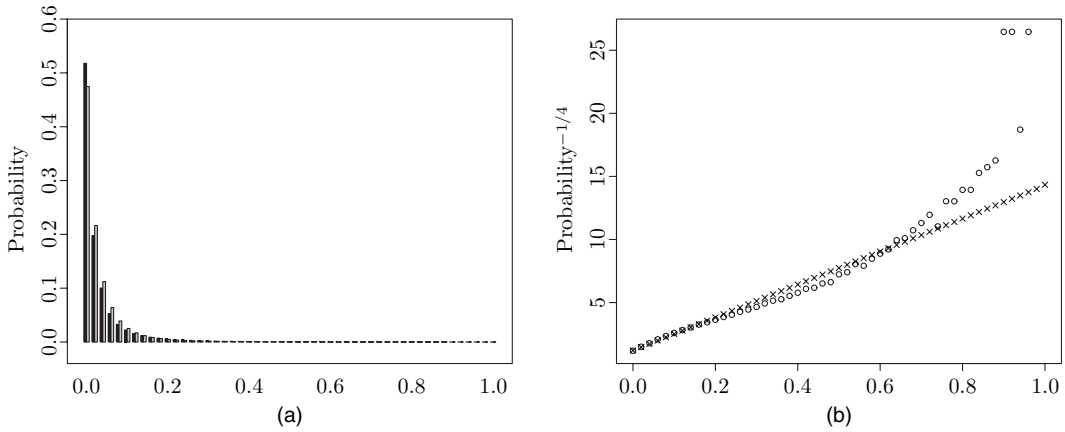

Fig. 3. Typical example of the probability mass function of $\hat{\Pi}_{25}(k)$ for $k \in L_{q/p}$ (▮, ○), alongside the $-\frac{1}{4}$-concave distribution (▮, ×), which has maximum tail probability beyond 0.4

approach is that $\mathbb{E}\{\tilde{\Pi}_B(k)\} = p^2_{k,\lfloor n/2 \rfloor}$ is much smaller than $\mathbb{E}\{\hat{\Pi}_B(k)\} = p_{k,\lfloor n/2 \rfloor}$ for variables with low selection probability, so the Markov bound is quite tight. However, for τ close to $\frac{1}{2}$, inequality (5) starts to become weak, and bounds can only be obtained for $\tau > \frac{1}{2}$ in any case.

To solve this problem, we can apply our versions of Markov's inequality directly to $\hat{\Pi}_B(k)$. We have found, through our simulations, that, for variables with low selection probability, the distribution of $\hat{\Pi}_B(k)$ can be modelled very well as a $-\frac{1}{4}$-concave distribution (Fig. 3). That the distribution of $\hat{\Pi}_B(k)$ is closer to log-concavity than that of $\tilde{\Pi}_B(k)$ is intuitive because, although the summands in expression (3) are not independent, terms involving subsamples which have little overlap will be close to independent. If we assume that $\tilde{\Pi}_B(k)$ is $-\frac{1}{2}$ concave and that $\hat{\Pi}_B(k)$ is $-\frac{1}{4}$ concave for all $k \in L_\theta$, we can obtain our best bound

$$\mathbb{E}|\hat{S}^{\mathrm{CPSS}}_{n,\tau} \cap L_\theta| \leqslant \min\{D(\theta^2, 2\tau - 1, B, -\tfrac{1}{2}), D(\theta, \tau, 2B, -\tfrac{1}{4})\}|L_\theta|, \tag{8}$$

which is valid for all $\tau \in (\theta, 1]$, provided that we adopt the convention that $D(\cdot, t, \cdot, \cdot) = 1$ for $t \leqslant 0$. The resulting improvements in the bounds can been seen in Fig. 2. Note the kink in Fig. 2 for the r-concave bound (8) just before $\tau = 0.6$. This corresponds to the transition from where $D(\theta, \tau, 2B, -\frac{1}{4})$ is smaller to where $D(\theta^2, 2\tau - 1, B, -\frac{1}{2})$ is smaller.

We applied the algorithm that is described in Appendix A.4 to produce tables of values of

$$\min\{D(\theta^2, 2\tau - 1, 50, -\tfrac{1}{2}), D(\theta, \tau, 100, -\tfrac{1}{4})\}$$

over a grid of θ- and τ-values; see Table 1 and Table 2.

3.4. How to use these bounds in practice

The quantities $|L_\theta|$ and $\mathbb{E}|\hat{S}_{\lfloor n/2 \rfloor} \cap L_\theta|$, which appear on the right-hand sides of the bounds, will in general be unknown to the statistician. Thus, when using the bounds, they will typically need to be replaced by p and q respectively. In addition, several parameters must be selected, and in this section we go through each of these in turn and give guidance on how to choose them.

3.4.1. Choice of B

We recommend $B = 50$ as a default value. Choosing B larger than this increases the computational burden, and may lead to the r-concavity assumptions being violated.

Table 1. Values of $\min\{D(\theta^2, 2\tau - 1, 50, -\frac{1}{2}), D(\theta, \tau, 100, -\frac{1}{4})\}$ for $\theta \in \{0.01, 0.02, 0.03, 0.04, 0.05\}$

τ	*Results for the following values of* θ*:*				
	0.01	*0.02*	*0.03*	*0.04*	*0.05*
0.30	6.11×10^{-4}	2.70×10^{-3}	6.51×10^{-3}	1.21×10^{-2}	1.93×10^{-2}
0.31	5.57×10^{-4}	2.47×10^{-3}	5.99×10^{-3}	1.12×10^{-2}	1.79×10^{-2}
0.32	5.08×10^{-4}	2.26×10^{-3}	5.52×10^{-3}	1.03×10^{-2}	1.66×10^{-2}
0.33	4.65×10^{-4}	2.08×10^{-3}	5.10×10^{-3}	9.57×10^{-3}	1.55×10^{-2}
0.34	4.27×10^{-4}	1.92×10^{-3}	4.71×10^{-3}	8.88×10^{-3}	1.44×10^{-2}
0.35	3.92×10^{-4}	1.77×10^{-3}	4.36×10^{-3}	8.25×10^{-3}	1.34×10^{-2}
0.36	3.61×10^{-4}	1.64×10^{-3}	4.05×10^{-3}	7.68×10^{-3}	1.25×10^{-2}
0.37	3.33×10^{-4}	1.51×10^{-3}	3.76×10^{-3}	7.15×10^{-3}	1.17×10^{-2}
0.38	3.08×10^{-4}	1.40×10^{-3}	3.50×10^{-3}	6.67×10^{-3}	1.09×10^{-2}
0.39	2.85×10^{-4}	1.30×10^{-3}	3.26×10^{-3}	6.23×10^{-3}	1.02×10^{-2}
0.40	2.64×10^{-4}	1.21×10^{-3}	3.04×10^{-3}	5.82×10^{-3}	9.59×10^{-3}
0.41	2.45×10^{-4}	1.13×10^{-3}	2.83×10^{-3}	5.45×10^{-3}	9.00×10^{-3}
0.42	2.27×10^{-4}	1.05×10^{-3}	2.65×10^{-3}	5.10×10^{-3}	8.44×10^{-3}
0.43	2.12×10^{-4}	9.81×10^{-4}	2.48×10^{-3}	4.78×10^{-3}	7.93×10^{-3}
0.44	1.97×10^{-4}	9.16×10^{-4}	2.32×10^{-3}	4.48×10^{-3}	7.45×10^{-3}
0.45	1.84×10^{-4}	8.56×10^{-4}	2.17×10^{-3}	4.21×10^{-3}	7.01×10^{-3}
0.46	1.71×10^{-4}	8.01×10^{-4}	2.03×10^{-3}	3.95×10^{-3}	6.60×10^{-3}
0.47	1.60×10^{-4}	7.50×10^{-4}	1.91×10^{-3}	3.72×10^{-3}	6.21×10^{-3}
0.48	1.50×10^{-4}	7.02×10^{-4}	1.79×10^{-3}	3.50×10^{-3}	5.85×10^{-3}
0.49	1.40×10^{-4}	6.58×10^{-4}	1.68×10^{-3}	3.29×10^{-3}	5.52×10^{-3}
0.50	1.31×10^{-4}	6.18×10^{-4}	1.58×10^{-3}	3.10×10^{-3}	5.20×10^{-3}
0.51	1.23×10^{-4}	5.80×10^{-4}	1.49×10^{-3}	2.92×10^{-3}	4.91×10^{-3}
0.52	1.15×10^{-4}	5.45×10^{-4}	1.40×10^{-3}	2.75×10^{-3}	4.63×10^{-3}
0.53	1.08×10^{-4}	5.12×10^{-4}	1.32×10^{-3}	2.59×10^{-3}	4.37×10^{-3}
0.54	1.01×10^{-4}	4.81×10^{-4}	1.24×10^{-3}	2.44×10^{-3}	4.13×10^{-3}
0.55	9.51×10^{-5}	4.52×10^{-4}	1.17×10^{-3}	2.30×10^{-3}	3.90×10^{-3}
0.56	8.93×10^{-5}	4.26×10^{-4}	1.10×10^{-3}	2.17×10^{-3}	3.68×10^{-3}
0.57	8.39×10^{-5}	4.01×10^{-4}	1.04×10^{-3}	2.05×10^{-3}	3.48×10^{-3}
0.58	7.89×10^{-5}	3.77×10^{-4}	9.78×10^{-4}	1.94×10^{-3}	3.29×10^{-3}
0.59	7.41×10^{-5}	3.55×10^{-4}	9.22×10^{-4}	1.83×10^{-3}	2.99×10^{-3}
0.60	6.97×10^{-5}	3.34×10^{-4}	8.69×10^{-4}	1.64×10^{-3}	2.61×10^{-3}
0.61	6.56×10^{-5}	3.15×10^{-4}	7.99×10^{-4}	1.45×10^{-3}	2.30×10^{-3}
0.62	6.16×10^{-5}	2.96×10^{-4}	7.12×10^{-4}	1.29×10^{-3}	2.05×10^{-3}
0.63	5.80×10^{-5}	2.78×10^{-4}	6.38×10^{-4}	1.16×10^{-3}	1.84×10^{-3}
0.64	5.45×10^{-5}	2.51×10^{-4}	5.76×10^{-4}	1.04×10^{-3}	1.66×10^{-3}
0.65	5.13×10^{-5}	2.27×10^{-4}	5.22×10^{-4}	9.46×10^{-4}	1.51×10^{-3}
0.66	4.82×10^{-5}	2.07×10^{-4}	4.75×10^{-4}	8.61×10^{-4}	1.37×10^{-3}
0.67	4.53×10^{-5}	1.89×10^{-4}	4.33×10^{-4}	7.86×10^{-4}	1.25×10^{-3}
0.68	4.23×10^{-5}	1.73×10^{-4}	3.97×10^{-4}	7.20×10^{-4}	1.15×10^{-3}
0.69	3.88×10^{-5}	1.58×10^{-4}	3.64×10^{-4}	6.60×10^{-4}	1.05×10^{-3}
0.70	3.56×10^{-5}	1.45×10^{-4}	3.35×10^{-4}	6.07×10^{-4}	9.68×10^{-4}
0.71	3.28×10^{-5}	1.34×10^{-4}	3.08×10^{-4}	5.59×10^{-4}	8.91×10^{-4}
0.72	3.02×10^{-5}	1.23×10^{-4}	2.84×10^{-4}	5.15×10^{-4}	8.21×10^{-4}
0.73	2.79×10^{-5}	1.14×10^{-4}	2.62×10^{-4}	4.76×10^{-4}	7.58×10^{-4}

(*continued*)

Table 1 (*continued*)

τ	*Results for the following values of* θ*:*				
	0.01	*0.02*	*0.03*	*0.04*	*0.05*
0.74	2.57×10^{-5}	1.05×10^{-4}	2.42×10^{-4}	4.39×10^{-4}	7.00×10^{-4}
0.75	2.37×10^{-5}	9.70×10^{-5}	2.23×10^{-4}	4.06×10^{-4}	6.47×10^{-4}
0.76	2.19×10^{-5}	8.95×10^{-5}	2.06×10^{-4}	3.75×10^{-4}	5.97×10^{-4}
0.77	2.02×10^{-5}	8.27×10^{-5}	1.90×10^{-4}	3.46×10^{-4}	5.52×10^{-4}
0.78	1.87×10^{-5}	7.63×10^{-5}	1.76×10^{-4}	3.20×10^{-4}	5.10×10^{-4}
0.79	1.72×10^{-5}	7.04×10^{-5}	1.62×10^{-4}	2.95×10^{-4}	4.70×10^{-4}
0.80	1.59×10^{-5}	6.48×10^{-5}	1.50×10^{-4}	2.72×10^{-4}	4.34×10^{-4}
0.81	1.46×10^{-5}	5.97×10^{-5}	1.38×10^{-4}	2.51×10^{-4}	3.99×10^{-4}
0.82	1.34×10^{-5}	5.48×10^{-5}	1.27×10^{-4}	2.30×10^{-4}	3.67×10^{-4}
0.83	1.23×10^{-5}	5.03×10^{-5}	1.16×10^{-4}	2.12×10^{-4}	3.37×10^{-4}
0.84	1.13×10^{-5}	4.60×10^{-5}	1.06×10^{-4}	1.94×10^{-4}	3.09×10^{-4}
0.85	1.03×10^{-5}	4.20×10^{-5}	9.71×10^{-5}	1.77×10^{-4}	2.82×10^{-4}
0.86	9.35×10^{-6}	3.82×10^{-5}	8.84×10^{-5}	1.61×10^{-4}	2.57×10^{-4}
0.87	8.47×10^{-6}	3.46×10^{-5}	8.02×10^{-5}	1.46×10^{-4}	2.33×10^{-4}
0.88	7.64×10^{-6}	3.12×10^{-5}	7.24×10^{-5}	1.32×10^{-4}	2.11×10^{-4}
0.89	6.85×10^{-6}	2.80×10^{-5}	6.50×10^{-5}	1.19×10^{-4}	1.89×10^{-4}
0.90	6.10×10^{-6}	2.49×10^{-5}	5.80×10^{-5}	1.06×10^{-4}	1.69×10^{-4}

3.4.2. Choice of θ

As mentioned at the beginning of Section 3.2, $\theta = q/p$ is a natural choice. In other words, we regard the below-average selection probability variables as the irrelevant variables. Other choices of θ are possible, but the use of expressions (6) and (7) to construct the bound suggests that the inequality will be tightest when most of the variables have a selection probability that is close to θ.

3.4.3. Choice of q and threshold τ

One can regard the choice of $q = \mathbb{E}(|\hat{S}_{\lfloor n/2 \rfloor}|)$ (which is usually fixed through a tuning parameter λ) as part of the choice of the base selection procedure. One option is to fix q by varying λ at each evaluation of the selection procedure until it selects q variables. However, if the number of variables selected at each iteration is unknown in advance (e.g. if λ is fixed, or if cross-validation is used to choose λ at each iteration), then q can be estimated by $\Sigma_{k=1}^{p} \hat{\Pi}_B(k)$.

An important point to note is that, although choosing λ or q is usually crucial when carrying out variable selection, this is not so when using CPSS. Our experience is that the performance of CPSS is surprisingly insensitive to the choice of q (see also Meinshausen and Bühlmann (2010)). That is to say, $L_{q/p}$ does not vary much as q varies, and also the final selected sets for different values of q tend to be similar (where different thresholds are chosen to control the selection of variables in $L_{q/p}$ at a prespecified level). Thus, when using CPSS, it is the threshold τ that plays a role similar to that of a tuning parameter for the base procedure. The great advantage of CPSS is that our bounds allow us to choose τ to control the expected number of low selection probability variables selected.

To summarize: we recommend as a sensible default CPSS procedure taking $B = 50$ and $\theta = q/p$. We then choose τ by using the bound (8) with $|L_\theta|$ replaced by p to control the expected number of low selection probability variables chosen.

Table 2. Values of $\min\{D(\theta^2, 2\tau - 1, 50, -\frac{1}{2}), D(\theta, \tau, 100, -\frac{1}{4})\}$ for $\theta \in \{0.06, 0.07, 0.08, 0.09, 0.1\}$

τ	*Results for the following values of* θ*:*				
	0.06	*0.07*	*0.08*	*0.09*	*0.10*
0.30	2.81×10^{-2}	3.82×10^{-2}	4.97×10^{-2}	6.24×10^{-2}	7.63×10^{-2}
0.31	2.61×10^{-2}	3.57×10^{-2}	4.64×10^{-2}	5.84×10^{-2}	7.14×10^{-2}
0.32	2.43×10^{-2}	3.33×10^{-2}	4.35×10^{-2}	5.47×10^{-2}	6.70×10^{-2}
0.33	2.27×10^{-2}	3.12×10^{-2}	4.08×10^{-2}	5.14×10^{-2}	6.30×10^{-2}
0.34	2.12×10^{-2}	2.92×10^{-2}	3.83×10^{-2}	4.83×10^{-2}	5.93×10^{-2}
0.35	1.98×10^{-2}	2.73×10^{-2}	3.59×10^{-2}	4.55×10^{-2}	5.59×10^{-2}
0.36	1.85×10^{-2}	2.57×10^{-2}	3.38×10^{-2}	4.29×10^{-2}	5.28×10^{-2}
0.37	1.74×10^{-2}	2.41×10^{-2}	3.18×10^{-2}	4.04×10^{-2}	4.99×10^{-2}
0.38	1.63×10^{-2}	2.26×10^{-2}	2.99×10^{-2}	3.81×10^{-2}	4.72×10^{-2}
0.39	1.53×10^{-2}	2.13×10^{-2}	2.82×10^{-2}	3.60×10^{-2}	4.46×10^{-2}
0.40	1.43×10^{-2}	2.00×10^{-2}	2.66×10^{-2}	3.40×10^{-2}	4.22×10^{-2}
0.41	1.35×10^{-2}	1.89×10^{-2}	2.51×10^{-2}	3.22×10^{-2}	4.00×10^{-2}
0.42	1.27×10^{-2}	1.78×10^{-2}	2.37×10^{-2}	3.04×10^{-2}	3.79×10^{-2}
0.43	1.19×10^{-2}	1.68×10^{-2}	2.24×10^{-2}	2.88×10^{-2}	3.59×10^{-2}
0.44	1.12×10^{-2}	1.58×10^{-2}	2.11×10^{-2}	2.72×10^{-2}	3.40×10^{-2}
0.45	1.06×10^{-2}	1.49×10^{-2}	2.00×10^{-2}	2.58×10^{-2}	3.23×10^{-2}
0.46	9.98×10^{-3}	1.41×10^{-2}	1.89×10^{-2}	2.44×10^{-2}	3.06×10^{-2}
0.47	9.41×10^{-3}	1.33×10^{-2}	1.79×10^{-2}	2.31×10^{-2}	2.90×10^{-2}
0.48	8.88×10^{-3}	1.26×10^{-2}	1.69×10^{-2}	2.19×10^{-2}	2.76×10^{-2}
0.49	8.38×10^{-3}	1.19×10^{-2}	1.60×10^{-2}	2.08×10^{-2}	2.62×10^{-2}
0.50	7.92×10^{-3}	1.12×10^{-2}	1.52×10^{-2}	1.97×10^{-2}	2.48×10^{-2}
0.51	7.48×10^{-3}	1.06×10^{-2}	1.44×10^{-2}	1.87×10^{-2}	2.36×10^{-2}
0.52	7.07×10^{-3}	1.01×10^{-2}	1.36×10^{-2}	1.77×10^{-2}	2.24×10^{-2}
0.53	6.68×10^{-3}	9.53×10^{-3}	1.29×10^{-2}	1.68×10^{-2}	2.13×10^{-2}
0.54	6.32×10^{-3}	9.02×10^{-3}	1.22×10^{-2}	1.60×10^{-2}	2.02×10^{-2}
0.55	5.98×10^{-3}	8.54×10^{-3}	1.16×10^{-2}	1.52×10^{-2}	1.92×10^{-2}
0.56	5.65×10^{-3}	8.09×10^{-3}	1.10×10^{-2}	1.44×10^{-2}	1.83×10^{-2}
0.57	5.35×10^{-3}	7.66×10^{-3}	1.04×10^{-2}	1.37×10^{-2}	1.73×10^{-2}
0.58	5.06×10^{-3}	7.13×10^{-3}	9.49×10^{-3}	1.22×10^{-2}	1.54×10^{-2}
0.59	4.39×10^{-3}	6.09×10^{-3}	8.10×10^{-3}	1.04×10^{-2}	1.31×10^{-2}
0.60	3.82×10^{-3}	5.30×10^{-3}	7.04×10^{-3}	9.08×10^{-3}	1.14×10^{-2}
0.61	3.37×10^{-3}	4.67×10^{-3}	6.21×10^{-3}	8.00×10^{-3}	1.01×10^{-2}
0.62	3.01×10^{-3}	4.17×10^{-3}	5.54×10^{-3}	7.14×10^{-3}	8.97×10^{-3}
0.63	2.70×10^{-3}	3.74×10^{-3}	4.98×10^{-3}	6.42×10^{-3}	8.06×10^{-3}
0.64	2.44×10^{-3}	3.38×10^{-3}	4.50×10^{-3}	5.80×10^{-3}	7.29×10^{-3}
0.65	2.21×10^{-3}	3.07×10^{-3}	4.08×10^{-3}	5.26×10^{-3}	6.62×10^{-3}
0.66	2.01×10^{-3}	2.79×10^{-3}	3.72×10^{-3}	4.79×10^{-3}	6.03×10^{-3}
0.67	1.84×10^{-3}	2.55×10^{-3}	3.40×10^{-3}	4.38×10^{-3}	5.51×10^{-3}
0.68	1.68×10^{-3}	2.34×10^{-3}	3.11×10^{-3}	4.01×10^{-3}	5.05×10^{-3}
0.69	1.55×10^{-3}	2.14×10^{-3}	2.86×10^{-3}	3.68×10^{-3}	4.64×10^{-3}
0.70	1.42×10^{-3}	1.97×10^{-3}	2.63×10^{-3}	3.39×10^{-3}	4.27×10^{-3}
0.71	1.31×10^{-3}	1.82×10^{-3}	2.42×10^{-3}	3.12×10^{-3}	3.93×10^{-3}
0.72	1.21×10^{-3}	1.68×10^{-3}	2.23×10^{-3}	2.88×10^{-3}	3.63×10^{-3}
0.73	1.11×10^{-3}	1.55×10^{-3}	2.06×10^{-3}	2.66×10^{-3}	3.35×10^{-3}

(*continued*)

Table 2 (*continued*)

τ	*Results for the following values of* θ*:*				
	0.06	*0.07*	*0.08*	*0.09*	*0.10*
0.74	1.03×10^{-3}	1.43×10^{-3}	1.90×10^{-3}	2.46×10^{-3}	3.09×10^{-3}
0.75	9.51×10^{-4}	1.32×10^{-3}	1.76×10^{-3}	2.27×10^{-3}	2.86×10^{-3}
0.76	8.78×10^{-4}	1.22×10^{-3}	1.63×10^{-3}	2.10×10^{-3}	2.64×10^{-3}
0.77	8.12×10^{-4}	1.13×10^{-3}	1.50×10^{-3}	1.94×10^{-3}	2.44×10^{-3}
0.78	7.50×10^{-4}	1.04×10^{-3}	1.39×10^{-3}	1.79×10^{-3}	2.26×10^{-3}
0.79	6.92×10^{-4}	9.61×10^{-4}	1.28×10^{-3}	1.65×10^{-3}	2.08×10^{-3}
0.80	6.38×10^{-4}	8.86×10^{-4}	1.18×10^{-3}	1.53×10^{-3}	1.92×10^{-3}
0.81	5.88×10^{-4}	8.16×10^{-4}	1.09×10^{-3}	1.41×10^{-3}	1.77×10^{-3}
0.82	5.41×10^{-4}	7.51×10^{-4}	1.00×10^{-3}	1.29×10^{-3}	1.63×10^{-3}
0.83	4.97×10^{-4}	6.89×10^{-4}	9.20×10^{-4}	1.19×10^{-3}	1.50×10^{-3}
0.84	4.55×10^{-4}	6.32×10^{-4}	8.43×10^{-4}	1.09×10^{-3}	1.37×10^{-3}
0.85	4.16×10^{-4}	5.77×10^{-4}	7.71×10^{-4}	9.95×10^{-4}	1.25×10^{-3}
0.86	3.79×10^{-4}	5.26×10^{-4}	7.02×10^{-4}	9.07×10^{-4}	1.14×10^{-3}
0.87	3.44×10^{-4}	4.77×10^{-4}	6.37×10^{-4}	8.23×10^{-4}	1.04×10^{-3}
0.88	3.11×10^{-4}	4.31×10^{-4}	5.76×10^{-4}	7.44×10^{-4}	9.37×10^{-4}
0.89	2.79×10^{-4}	3.88×10^{-4}	5.18×10^{-4}	6.69×10^{-4}	8.42×10^{-4}
0.90	2.49×10^{-4}	3.46×10^{-4}	4.63×10^{-4}	5.97×10^{-4}	7.53×10^{-4}

4. Numerical properties

4.1. Simulation study

In this section we investigate the performance and validity of the bounds that were derived in the previous section by applying CPSS to simulated data. We consider both linear and logistic regression and different values of p and n. In each of these settings, we first generate independent explanatory vectors $X_1, \ldots, X_n$ with each $X_i \sim N_p(0, \Sigma)$. We use a Toeplitz covariance matrix Σ with entries

$$\Sigma_{ij} = \rho^{\|i-j|-p/2|-p/2},$$

and we look at various values of ρ in $[0, 1)$. So the correlation between the components decays exponentially with the distance between them in $\mathbb{Z}_p$.

For linear regression, we generate a vector of errors $\varepsilon \sim N_n(0, \sigma^2 I)$ and set

$$Y = X\beta + \varepsilon,$$

where the design matrix X has ith row X_i^{T}. The error variance σ^2 is chosen to achieve different values of the signal-to-noise ratio SNR, which we define here by

$$\mathrm{SNR}^2 = \frac{\mathbb{E}\|X\beta\|^2}{\mathbb{E}\|\varepsilon\|^2}.$$

For logistic regression, we generate independent responses

$$Y_i \sim \mathrm{Bin}(1, p_i), \qquad i = 1, \ldots, n,$$

where

$$\log\left(\frac{p_i}{1-p_i}\right) = \gamma X_i^{\mathrm{T}}\beta.$$

Here γ is a scaling factor which is chosen to achieve a particular Bayes error rate.

In both cases, we fix the p-dimensional vector of coefficients β to have $s \ll p$ non-zero components, $s/2$ of which we choose as equally spaced points within $[-1, -0.5]$ with the remaining $s/2$ equally spaced in $[0.5, 1]$. The indices of the non-zero components, S, are chosen to follow a geometric progression up to rounding, with first term 1 and $(s+1)$th term $p+1$. The values are then randomly assigned to each index in S, but this choice is then fixed for each particular simulation setting.

With $\rho > 0$, this set-up will have several signal variables correlated among themselves, and also some signal correlated with noise. In this way, the framework above includes a very wide variety of different data-generating processes on which we can test the theory of the previous section.

By varying the base selection procedure, its tuning parameters, the values of ρ, n, p, s and also SNR and Bayes error rates, we have applied CPSS in several hundred different simulation settings. For brevity, we present only a subset of these numerical experiments below, but the results from those omitted are not qualitatively different.

In the graphs which follow, we look at CPSS applied to the lasso (Tibshirani, 1996), which we implemented by using the package `glmnet` (Friedman *et al.*, 2010) in R (R Development Core Team, 2010). We follow the original stability selection procedure that was put forward in Meinshausen and Bühlmann (2010) and compare this with the method suggested by our r-concave bound (8). Thus we first choose the level l at which we wish to control the expected number of low selection probability variables (so we aim to have $\mathbb{E}|\hat{S}^{\mathrm{CPSS}}_{n,\tau} \cap L_{q/p}| \leqslant l$). Then we fix $q = \sqrt{(0.8lp)}$ and set the threshold τ at 0.9. This ensures that, according to the original worst-case bound, we control the expected number of low selection probability variables selected at the required level. In the r-concave case, we take our threshold as

$$\tilde{\tau} = \min\{\tau \in \{0, 1/(2B), \ldots, 1\} : \min\{D(q^2/p^2, 2\tau - 1, B, -\tfrac{1}{2}), D(q/p, \tau, 2B, -\tfrac{1}{4})\} \leqslant l/p\}.$$

We also give the results that we would obtain by using the lasso alone, but with the benefit of an oracle which knows the optimal value of the tuning parameter λ, i.e. we take $\hat{S}^{\lambda^*}_n$ as our selected set, where

$$\lambda^* = \inf\{\lambda : \mathbb{E}|\hat{S}^{\lambda}_n \cap L_{q/p}| \leqslant l\},$$

and $\hat{S}^{\lambda}_n$ is the selected set when using the lasso with tuning parameter λ applied to the whole data set.

We present all of our results relative to the performance of CPSS using an oracle-driven threshold τ^*, where τ^* is defined by

$$\tau^* = \min\{\tau \in \{0, 1/(2B), \ldots, 1\} : \mathbb{E}|\hat{S}^{\mathrm{CPSS}}_{n,\tau} \cap L_{q/p}| \leqslant l\}.$$

Referring to Figs 4–7, the heights of the black bars, grey bars and crosses are given by

$$\frac{\mathbb{E}|\hat{S}^{\mathrm{CPSS}}_{n,0.9} \cap S|}{\mathbb{E}|\hat{S}^{\mathrm{CPSS}}_{n,\tau^*} \cap S|}, \qquad \frac{\mathbb{E}|\hat{S}^{\mathrm{CPSS}}_{n,\tilde{\tau}} \cap S|}{\mathbb{E}|\hat{S}^{\mathrm{CPSS}}_{n,\tau^*} \cap S|}, \qquad \frac{\mathbb{E}|\hat{S}^{\lambda^*}_{n} \cap S|}{\mathbb{E}|\hat{S}^{\mathrm{CPSS}}_{n,\tau^*} \cap S|}$$

respectively. Thus the heights of the black and grey bars relate to the loss of power in using the threshold suggested by the corresponding bounds. In all our simulations, we used $B = 50$. Each scenario was run 500 times and, to determine the set $L_{q/p}$, in each scenario, we applied the particular selection procedure $\hat{S}_{\lfloor n/2 \rfloor}$ to 50 000 independent data sets.

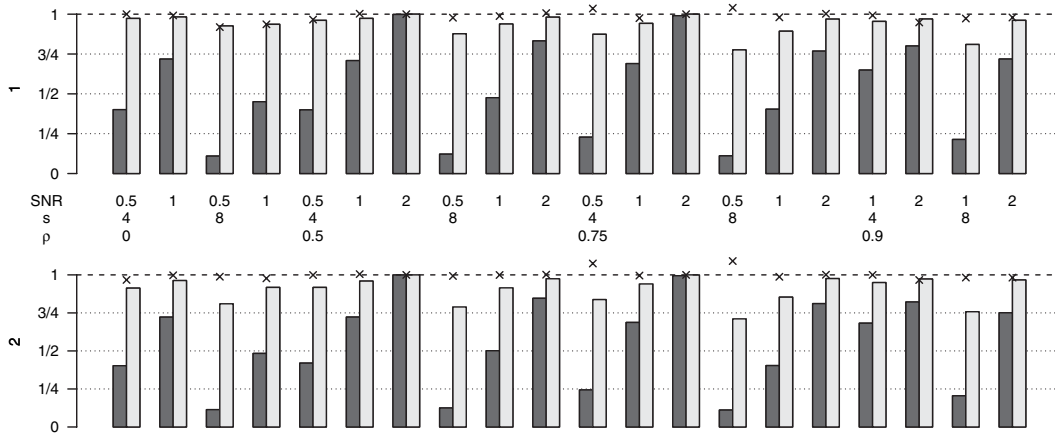

Fig. 4. Linear regression with $n = 200$ and $p = 1000$ (the y-axis label gives the error control level l; higher bars are preferred): ▮, worst-case procedure; ▮, r-concave procedure; ×, theoretical oracle-driven lasso procedure

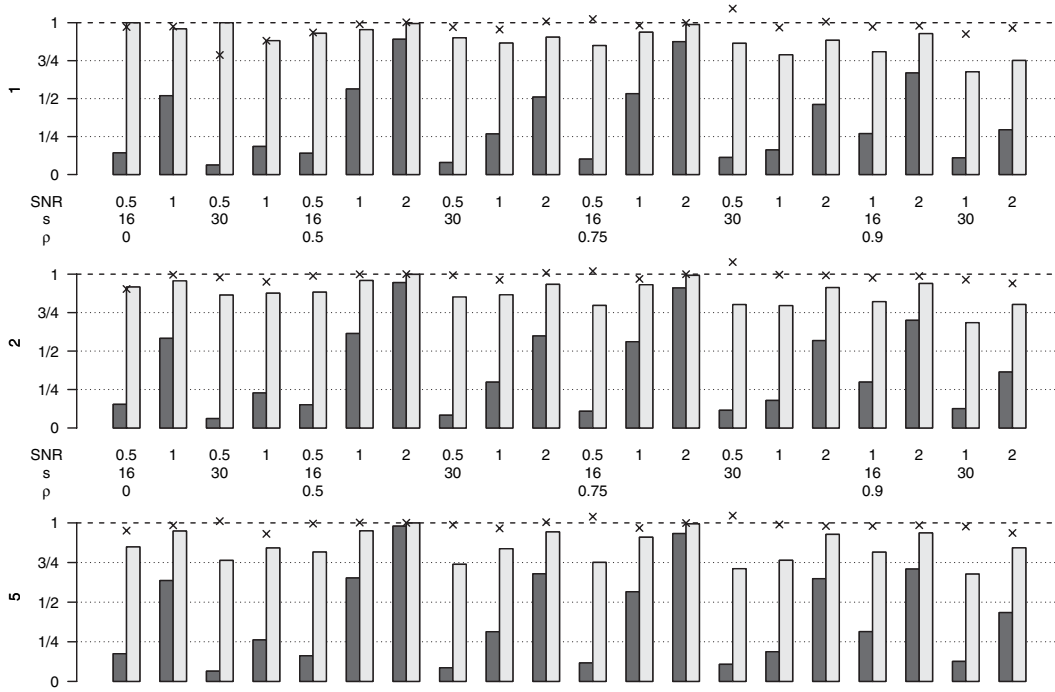

Fig. 5. As in Fig. 4 but with $n = 500$ and $p = 2000$

It is immediately obvious from the results that, using the r-concave bound, we can recover significantly more variables in S than when using the the worst-case bound. Furthermore, though it is not shown in the graphs explicitly, we also achieve the required level of error control in all except one case (where the r-concavity assumption fails). In fact the one particular example is hardly exceptional in that we have $\mathbb{E}|\hat{S}^{\mathrm{CPSS}}_{n,\tilde{\tau}} \cap L_{q/p}| = 1.034 > 1 = l$. Thus, in close accordance with our theory, there are no significant violations of the r-concave bound.

We also see that the loss in power due to using $\tilde{\tau}$ rather than τ^* is very low. In almost all of the scenarios, we can select more than 75% of the signal that we could select with the benefit of

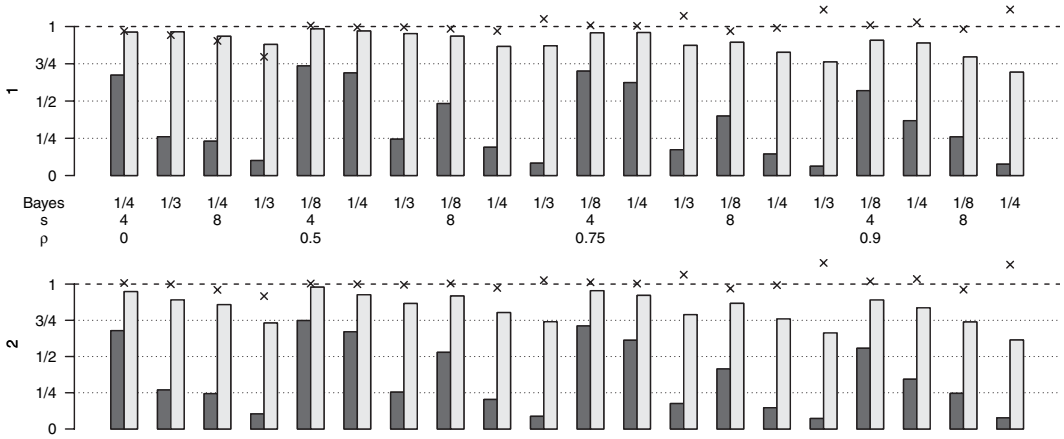

Fig. 6. As in Fig. 4 but with logistic regression ($n = 200$, $p = 1000$)

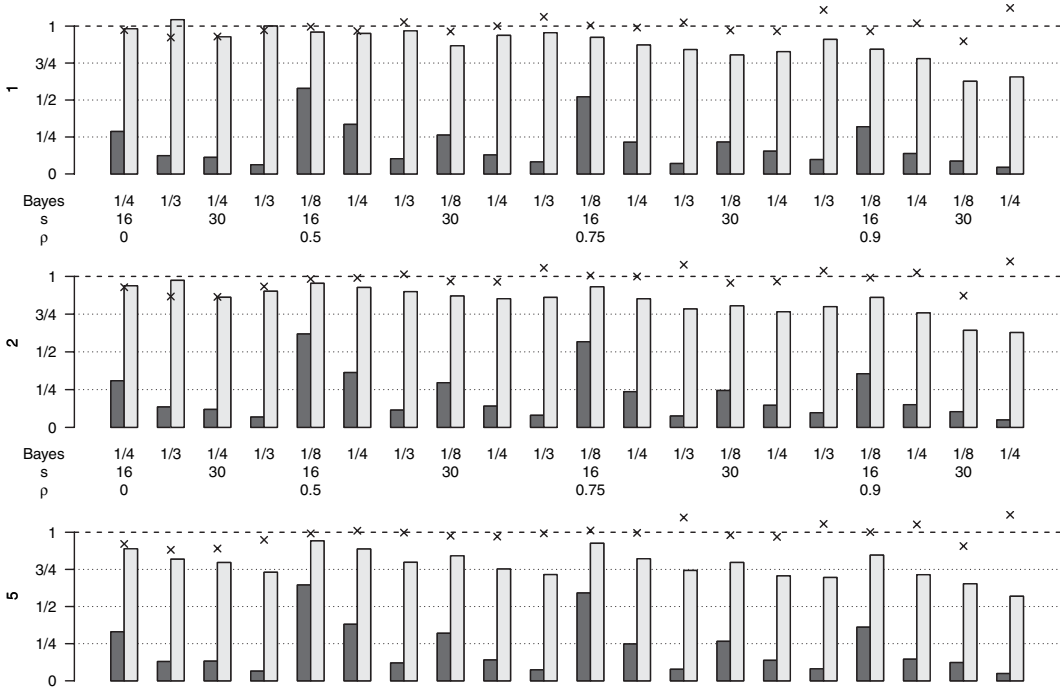

Fig. 7. As in Fig. 4 but with logistic regression ($n = 500$, $p = 2000$)

an oracle, and usually much more than this. It is interesting that the performances of the oracle CPSS and oracle lasso procedures are fairly similar. The key advantage of CPSS is that it allows for error control whereas there is in general no way of determining (or even approximating) the optimal λ^* that achieves the required error control. In fact, the performance of CPSS with our bound is only slightly worse than that of the oracle lasso procedure, and in a few cases, particularly when ρ is small, it is even slightly better. In the cases where $\rho \geqslant 0.75$, we see that CPSS is not quite as powerful. This is because having such large correlations between variables causes $\{p_{k,\lfloor n/2 \rfloor} : k = 1, \ldots, p\}$ to be relatively spread out in $[0, 1]$. As explained in Section 3.4,

we expect our bound to weaken in this situation. However, even when the correlation is as high as 0.9, we recover a sizable proportion of the signal that we would select if we had used the optimal τ^*.

4.2. Real data example

Here we illustrate our CPSS methodology on the widely studied colon data set of Alon *et al.* (1999), which is freely available from `http://microarray.princeton.edu/oncology/affydata/index.html`. The data consist of 2000 gene expression levels from 40 colon tumour samples and 22 normal colon tissue samples, measured by using Affymetrix oligonucleotide arrays. Our goal is to identify a small subset of genes which we are confident are linked with the development of colon cancer. Such a task is important for improving scientific understanding of the disease and for selecting genes as potential drug targets.

The data were first preprocessed by averaging over the expression levels for repeated genes (which had been tiled more than once on each array), log-transforming each gene expression level, standardizing each row to have mean 0 and unit variance, and finally removing the columns corresponding to control genes, so that $p = 1908$ genes remained. The transformation and standardization are very common preprocessing steps to reduce skewness in the data and help to eliminate the effects of systematic variations between different microarrays (see for example Amaratunga and Cabrera (2004) and Dudoit *et al.* (2002)).

We applied CPSS with l_1- (lasso) penalized logistic regression as the base procedure, with $B = 50$, and choosing τ by using both the r-concave bound of Section 3.4 and the original bound of Meinshausen and Bühlmann (2010). We estimated the expected classification error in the two cases by averaging over 128 repetitions of stratified random subsampling validation, taking eight cancerous and four normal observations in each test set. Thus, when applying CPSS, we had $n = 40 + 22 - 12 = 50$. We looked at $q = 8, 10, 12$, and set τ to control $\mathbb{E}|\hat{S}^{\mathrm{CPSS}}_{n,\tau} \cap L_{q/p}| \leqslant l$ with $l = 0.1$ and $l = 0.5$.

Rather than subsampling completely at random when using CPSS, we also stratified these subsamples to include the same proportion of cancerous to normal samples as in the training data that are supplied to the procedure. Without this step, some of the subsamples may not include any samples from one of the classes, and applying $\hat{S}_{\lfloor n/2 \rfloor}$ to such a subsample would give misleading results. Using stratified random subsampling is still compatible with our theory, provided that $\mathbb{E}|\hat{S}^{\mathrm{CPSS}}_{n,\tau} \cap L_\theta|$ is interpreted as an expectation over random data which contain the

Table 3. Improvement in classification error over the naive classifier which always determines the data to be from a cancerous tissue†

q	*Improvement (%) for the worst-case procedure*		*Improvement (%) for the r-concave procedure*	
	$l = 0.1$	$l = 0.5$	$l = 0.1$	$l = 0.5$
8	4.9 (0.5)	11.6 (1.1)	16 (2.3)	17.5 (5.1)
10	0.9 (0.1)	10.6 (0.9)	14.7 (1.6)	15.8 (4.4)
12	0.0 (0.0)	9.4 (0.8)	12.8 (1.1)	15.8 (4.1)

†Thus the classification errors are $33\frac{1}{3}\%$ minus these quantities. We also give the average number of variables selected in parentheses.

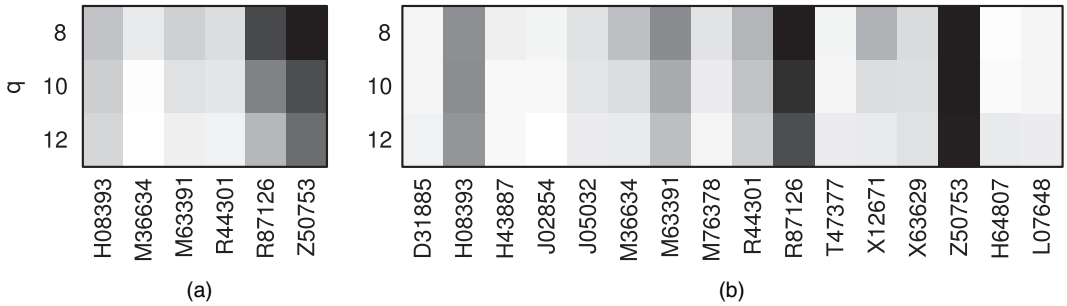

Fig. 8. For (a) $l = 0.1$ and (b) $l = 0.5$, we have plotted the proportion of times a gene was selected by our r-concave CPSS procedure for all genes which were selected at least 5% of the time among the 128 repetitions: ■, gene selected in every repetition: □, gene never selected (thus dark vertical lines indicate that the choice of q has little effect on the end result of CPSS)

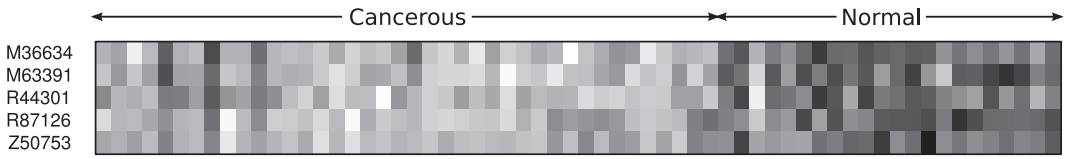

Fig. 9. Heat map of the normalized, centred, log-intensity values of the genes selected when we use the r-concave bound to choose τ such that we control $\mathbb{E}|\hat{S}^{\mathrm{CPSS}}_{n,\tau} \cap L_{q/p}| \leqslant 0.5$

same class proportions as observed in the original data. In general, this approach of stratified random subsampling is useful when the response is categorical.

The results in Table 3 show that, as expected, the new error bounds allow us to select more variables than the conservative bounds of Meinshausen and Bühlmann (2010) for the same level of error control and, as a consequence, the expected prediction error is reduced. Fig. 8 demonstrates the robustness of the selected set to the different values of q. Finally, we also applied CPSS on the entire data set with $q = 8$ and $B = 50$ and using the r-concave bound of Section 3.4 to choose τ to control $\mathbb{E}|\hat{S}^{\mathrm{CPSS}}_{n,\tau} \cap L_{q/p}| \leqslant 0.5$ (Fig. 9). We see that, with just five genes out of 1908, we manage to separate the two classes quite well.

Acknowledgements

The second author gratefully acknowledges the support of a Leverhulme Research Fellowship, as well as Engineering and Physical Sciences Research Council Early Career Fellowship EP/J017213/1.

Appendix A

A.1. Proof of theorem 1

The proof of theorem 1 requires the following lemma.

Lemma 1.

(a) If $\tau \in (\frac{1}{2}, 1]$, then

$$\mathbb{P}(k \in \hat{S}^{\mathrm{CPSS}}_{n,\tau}) \leqslant \frac{1}{2\tau - 1} p^2_{k,\lfloor n/2 \rfloor}.$$

(b) If $\tau \in [0, \frac{1}{2})$, then

$$\mathbb{P}(k \notin \hat{S}^{\mathrm{CPSS}}_{n,\tau}) \leqslant \frac{1}{1-2\tau}(1 - p_{k,\lfloor n/2 \rfloor})^2.$$

Proof.

(a) Let $\mathcal{A} = \{(A_{2j-1}, A_{2j}) : j = 1, \ldots, B\}$ be randomly chosen independent pairs of subsets of $\{1, \ldots, n\}$ of size $\lfloor n/2 \rfloor$ such that $A_{2j-1} \cap A_{2j} = \emptyset$. Then

$$0 \leqslant \frac{1}{B}\sum_{j=1}^{B}(1 - \mathbb{1}_{\{k \in \hat{S}(A_{2j-1})\}})(1 - \mathbb{1}_{\{k \in \hat{S}(A_{2j})\}}) = 1 - 2\hat{\Pi}_B(k) + \tilde{\Pi}_B(k). \tag{9}$$

Now $\mathbb{E}\{\tilde{\Pi}_B(k)\} = \mathbb{E}[\mathbb{E}\{\tilde{\Pi}_B(k)|\mathcal{A}\}] = p^2_{k,\lfloor n/2 \rfloor}$ because $\hat{S}(A_{2j-1})$ and $\hat{S}(A_{2j})$ are independent conditional on $\mathcal{A}$. It follows using expression (9) that

$$\mathbb{P}(k \in \hat{S}^{\mathrm{CPSS}}_{n,\tau}) = \mathbb{P}\{\hat{\Pi}_B(k) \geqslant \tau\} \leqslant \mathbb{P}[\tfrac{1}{2}\{1 + \tilde{\Pi}_B(k)\} \geqslant \tau] = \mathbb{P}\{\tilde{\Pi}_B(k) \geqslant 2\tau - 1\} \leqslant \frac{1}{2\tau - 1} p^2_{k,\lfloor n/2 \rfloor}, \tag{10}$$

where we have used Markov's inequality in the final step.

(b) Define $\hat{\Pi}_B^{\hat{N}_n}$ and $\tilde{\Pi}_B^{\hat{N}_n}$ by replacing $\hat{S}_n$ with $\hat{N}_n := \{1, \ldots, p\} \setminus \hat{S}_n$ in the definitions of $\hat{\Pi}_B$ and $\tilde{\Pi}_B$ respectively. Then, using the bound corresponding to expression (9) and Markov's inequality again,

$$\mathbb{P}(k \notin \hat{S}^{\mathrm{CPSS}}_{n,\tau}) = \mathbb{P}\{\hat{\Pi}_B(k) < \tau\} = \mathbb{P}\{\hat{\Pi}_B^{\hat{N}_n}(k) > 1 - \tau\} \leqslant \mathbb{P}\{\tilde{\Pi}_B^{\hat{N}_n}(k) > 1 - 2\tau\} \leqslant \frac{1}{1-2\tau}(1 - p_{k,\lfloor n/2 \rfloor})^2.$$

We now prove theorem 1 as follows.

(a) Note that

$$\mathbb{E}|\hat{S}_{\lfloor n/2 \rfloor} \cap L_\theta| = \mathbb{E}\Big(\sum_{k=1}^{p} \mathbb{1}_{\{k \in \hat{S}_{\lfloor n/2 \rfloor}\}} \mathbb{1}_{\{p_{k,\lfloor n/2 \rfloor} \leqslant \theta\}}\Big) = \sum_{k=1}^{p} p_{k,\lfloor n/2 \rfloor} \mathbb{1}_{\{p_{k,\lfloor n/2 \rfloor} \leqslant \theta\}}.$$

By lemma 1, it follows that

$$\begin{aligned}\mathbb{E}|\hat{S}^{\mathrm{CPSS}}_{n,\tau} \cap L_\theta| &= \mathbb{E}\Big(\sum_{k=1}^{p} \mathbb{1}_{\{k \in \hat{S}^{\mathrm{CPSS}}_{n,\tau}\}} \mathbb{1}_{\{p_{k,\lfloor n/2 \rfloor} \leqslant \theta\}}\Big) = \sum_{k=1}^{p} \mathbb{P}(k \in \hat{S}^{\mathrm{CPSS}}_{n,\tau}) \mathbb{1}_{\{p_{k,\lfloor n/2 \rfloor} \leqslant \theta\}} \\ &\leqslant \frac{1}{2\tau - 1}\sum_{k=1}^{p} p^2_{k,\lfloor n/2 \rfloor} \mathbb{1}_{\{p_{k,\lfloor n/2 \rfloor} \leqslant \theta\}} \leqslant \frac{\theta}{2\tau - 1}\mathbb{E}|\hat{S}_{\lfloor n/2 \rfloor} \cap L_\theta|.\end{aligned}$$

(b) The proof of part (b) is very similar to that of part (a) and so is omitted.

A.2. Proof of theorem 2

The proof of theorem 2 requires several preliminary results, and we use the following notation. Let G denote the finite lattice $\{0, 1/B, 2/B, \ldots, 1\} = (1/B)\mathbb{Z} \cap [0, 1]$. If f is a probability mass function on G, we write f_i for $f(i/B)$, thereby associating f with $(f_0, f_1, \ldots, f_B) \in \mathbb{R}^{B+1}$.

For $t \in G$, we denote the probability that a random variable distributed according to f takes values that are greater than or equal to t by $\mathcal{T}_t(f) := \Sigma_{i \geqslant Bt} f_i$. We also write $\mathcal{E}(f) := \Sigma_{i=1}^{B}(i/B) f_i$ for the expectation of this random variable and $\mathrm{supp}(f) := \{i/B \in G : f_i > 0\}$ for the support of f.

Let $\mathcal{U}$ be the set of all unimodal probability mass functions f on G, and let $\mathcal{U}_\eta = \{f \in U : \mathcal{E}(f) \leqslant \eta\}$. We consider the problem of maximizing $\mathcal{T}_t$ over $f \in \mathcal{U}_\eta$. Since the cases $\eta = 0$ and $t \leqslant \eta$ are trivial, there is no loss of generality in assuming throughout that $0 < \eta < t$ and $t \in G$, so in particular $t \geqslant 1/B$.

Lemma 2. There is a maximizer of $\mathcal{T}_t$ in $\mathcal{U}_\eta$.

Proof. Since $\mathcal{T}_t : \mathbb{R}^{B+1} \to \mathbb{R}$ is linear and therefore continuous, it suffices to show that $\mathcal{U}_\eta \subset \mathbb{R}^{B+1}$ is closed and bounded. Now $\mathcal{U}_\eta$ is bounded as $\mathcal{U}_\eta \subset [0, 1]^{B+1}$. Moreover, the hyperplane $H = \{(x_0, \ldots, x_B) : x_0 + x_1 + \ldots + x_B = 1\}$ is closed. Also, $\mathcal{E}$ is a continuous function on $\mathbb{R}^{B+1}$, so $\mathcal{E}^{-1}([0, \eta])$ is closed. Now let $O = \{f \in \mathbb{R}^{B+1} : f \text{ is not unimodal}\}$. If $f \in O$ then there must exist $i_1 < i_2 < i_3$ such that $f_{i_2} < \min\{f_{i_1}, f_{i_3}\}$. Clearly this inequality must hold for all g in a sufficiently small open ball about f, so O is open. We see that

$$\mathcal{U}_\eta = H \cap \mathcal{E}^{-1}([0,\eta]) \cap O^c.$$

Thus $\mathcal{U}_\eta$ is an intersection of closed sets and hence is closed.

We shall make frequent use of the following simple proposition in subsequent proofs.

Proposition 3. Suppose that $(x_1,\ldots,x_n) \in \mathbb{R}^n$ and $(y_1,\ldots,y_n) \in \mathbb{R}^n$ satisfy

$$\sum_{i=1}^n x_i = \sum_{i=1}^n y_i,$$

and that there is some $i^* \in \{1,\ldots,n\}$ with $x_i \geqslant y_i$ for all $i \leqslant i^*$ and $x_i \leqslant y_i$ for all $i > i^*$. Then

$$\sum_{i=1}^n i x_i \leqslant \sum_{i=1}^n i y_i,$$

with equality if and only if $x_i = y_i$ for $i = 1,\ldots,n$.

Proof. We have

$$\sum_{i \leqslant i^*} i(x_i - y_i) \leqslant i^* \sum_{i \leqslant i^*} (x_i - y_i) = i^* \sum_{i > i^*} (y_i - x_i) \leqslant \sum_{i > i^*} i(y_i - x_i).$$

The following result characterizes the extremal elements of $\mathcal{U}_\eta$ in the sense of maximizing the tail probability $\mathcal{T}_t$. In particular, it shows that such extremal elements can take only one of two simple forms.

Proposition 4. Any maximizer $f^* \in \mathcal{U}_\eta$ of $\mathcal{T}_t$ satisfies

(a) $\mathcal{E}(f^*) = \eta$,
(b) writing i_M for $B \max\{\mathrm{supp}(f^*)\}$, we have either
  (i) $f_0^* > f_1^* = f_2^* = \ldots = f_{i_{M-1}}^* \geqslant f_{i_M}^*$ or
  (ii) $i_M = t$ and $f_0^* = f_1^* = \ldots = f_{i_{M-1}}^* \leqslant f_{i_M}^*$.

Proof.

(a) Suppose that $f^* \in \mathcal{U}_\eta$ maximizes $\mathcal{T}_t$, but that $\mathcal{E}(f^*) < \eta$. Define $i_m := \min\{\mathrm{supp}(f^*)\}$. As $\eta < \tau$, we must have $i_m < Bt$. Define g by

$$g_i = \begin{cases} 0 & \text{if } i < i_m, \\ f_i^* - \varepsilon_1 & \text{if } i = i_m, \\ f_i^* + \varepsilon_2 & \text{if } i > i_m \end{cases}$$

where $\varepsilon_1, \varepsilon_2 > 0$ are chosen such that $\Sigma_{i=0}^B g_i = 1$, but are sufficiently small that $\mathcal{E}(g) \leqslant \eta$. Then $g \in \mathcal{U}_\eta$ but $\mathcal{T}_t(g) > \mathcal{T}_t(f^*)$, which is a contradiction.

(b) Suppose first that there is a mode of f^* which is at least t. Let $g \in \mathcal{U}_\eta$ be such that $g_i = f_i^*$ for $i \geqslant Bt$ and

$$g_i = \frac{1}{Bt} \sum_{l=0}^{Bt-1} f_l^*$$

for $i < Bt$. As $f_0^* \leqslant f_1^* \leqslant \ldots \leqslant f_{Bt}^*$, we can apply proposition 3 to see that

$$\mathcal{E}(g) \leqslant \mathcal{E}(f^*). \tag{11}$$

But $\mathcal{T}_t(g) = \mathcal{T}_t(f^*)$, so by optimality of f^* we must have equality in expression (11). Thus proposition 3 gives us that $f^* = g$.

Next, define $h \in \mathcal{U}_\eta$ by $h_i = f_i^*$ for $i < Bt$, $h_{Bt} = \mathcal{T}_t(f^*)$, and $h_i = 0$ for $i > Bt$. Then $\mathcal{T}_t(h) = \mathcal{T}_t(f^*)$. Again proposition 3 and the optimality of f^* give that $f^* = h$. Thus f^* satisfies property (b)(ii) of the theorem.

Now suppose that there is no mode of f^* which is at least t, so $f_{Bt}^* \geqslant f_{Bt+1}^* \geqslant \ldots \geqslant f_B^*$. Let $g \in \mathcal{U}_\eta$ satisfy $g_i = f_i^*$ for $i \geqslant Bt$ and $g_1 = \ldots = g_{Bt}$. We must have $g_0 > g_1$; otherwise f^* would have a mode at t. As $\mathcal{T}_t(g) = \mathcal{T}_t(f^*)$, optimality of f^* and proposition 3 imply $f^* = g$.

Finally, let $h \in \mathcal{U}_\eta$ satisfy $h_i = f_i^*$ for $i \leqslant Bt$ and $h_{Bt} = h_{Bt+1} = \ldots = h_{k-1} \geqslant h_k$, where k and h_k are chosen such that $\Sigma_{i=0}^B h_i = 1$. As before, proposition 3 allows us to deduce that $f^* = h$. Thus f^* satisfies property (b)(i) of the theorem.

We can now state Markov's inequality for random variables with unimodal distributions on G, which may be of some independent interest.

Theorem 3 (Markov's inequality under unimodality). Let X be a random variable with a unimodal distribution on $G = \{0, 1/B, 2/B, \ldots, 1\}$, and let $t \in G$. If $\eta := \mathbb{E}(X) \leqslant \frac{1}{3}$, then

$$\mathbb{P}(X \geqslant t) \leqslant \begin{cases} \dfrac{2\eta - t + 1/B}{t + 1/B} & \text{if } t \in (\eta,\ \min\{\frac{3}{2}\eta + 1/(2B), 2\eta\}], \\ \dfrac{\eta}{2t - 1/B} & \text{if } t \in (\min\{\frac{3}{2}\eta + 1/(2B), 2\eta\},\ \frac{1}{2}], \\ \dfrac{2\eta(1 - t + 1/B)}{1 + 1/B} & \text{if } t \in (\frac{1}{2},\ 1]. \end{cases}$$

Let d be defined by

$$d := d(\eta, B) = -2(\eta - \tfrac{1}{2})(6\eta + 1) + \frac{2 - 4\eta}{B} + \frac{(4\eta - 1)^2}{B^2}.$$

If $\eta > \frac{1}{3}$ and $d > 0$, then

$$\mathbb{P}(X \geqslant t) \leqslant \begin{cases} \dfrac{2\eta - t + 1/B}{t + 1/B} & \text{if } t \in (\eta,\ \frac{1}{2} + (1/4\eta)(1 + 1/B - d^{1/2})], \\ \dfrac{2\eta(1 - t + 1/B)}{1 + 1/B} & \text{if } t \in (\frac{1}{2} + (1/4\eta)(1 + 1/B - d^{1/2}),\ 1]. \end{cases}$$

Finally, if $\eta > \frac{1}{3}$ and $d \leqslant 0$, then

$$\mathbb{P}(X \geqslant t) \leqslant \frac{2\eta - t + 1/B}{t + 1/B}.$$

Proof. Proposition 4 tells us that $\mathbb{P}(X \geqslant t)$ must be at most the maximum of the optimal solutions to the following two optimization problems.

(a) Problem P:

$$\begin{aligned} &\text{maximize } b(s - Bt) + c \text{ in } a, b, c, s \\ &\text{subject to } a + (s-1)b + c = 1, \\ &\qquad\qquad s(s-1)b/2 + sc = B\eta, \\ &\qquad\qquad a > b \geqslant c \geqslant 0, \\ &\qquad\qquad s \in \{Bt, Bt + 1, \ldots, B\}. \end{aligned}$$

(b) Problem Q:

$$\begin{aligned} &\text{maximize } b \text{ in } a, b \\ &\text{subject to } Bta + b = 1, \\ &\qquad\qquad Bt(Bt - 1)a/2 + Btb = B\eta, \\ &\qquad\qquad b \geqslant a \geqslant 0. \end{aligned}$$

Problem P corresponds to case (b)(i) of proposition 4, and problem Q to case (b)(ii).

The solution to problem Q is determined entirely by the constraints, and we see that the optimal value is

$$\frac{2\eta - t + 1/B}{t + 1/B}. \tag{12}$$

To solve problem P, we break it into $B(1 - t) + 1$ subproblems: for $s \in \{Bt, Bt + 1, \ldots, B\}$, we define subproblem P(s) as follows.

$$\begin{aligned} &\text{Maximize } b\,(s - Bt) + c \text{ in } a, b, c \\ &\text{subject to } a + (s-1)b + c = 1, \\ &\qquad\qquad s(s-1)b/2 + sc = B\eta, \\ &\qquad\qquad b \geqslant c, \\ &\qquad\qquad a, b, c \geqslant 0. \end{aligned}$$

Note that we have not included the $a > b$ constraint. This is because proposition 4 ensures that this constraint is always satisfied at an optimal solution of problem P, so there exists s^* such that every optimal solution of problem P(s^*) corresponds to an optimal solution of problem P.

Now each subproblem is a standard linear programming problem, so we know that one of the basic feasible solutions must be optimal. Since $a > 0$, all basic feasible solutions must have either $c = 0$ or $b = c$. Thus we may replace the subproblems P(s) by problem P′(s).

Maximize $b\,(s - Bt + 1)$ in a, b
subject to $a + sb = 1$,
$s\,(s+1)b/2 = B\eta$,
$a, b \geqslant 0$.

The second constraint is enough to determine that the optimal value of problem P′(s) is

$$\frac{2B\eta(s - Bt + 1)}{s(s+1)} =: \gamma(s). \tag{13}$$

Now we can proceed to find an s^* which maximizes γ over $\{Bt, Bt+1, \ldots, B\}$. The sign of $\gamma'(s)$ is the sign of

$$-s^2 + 2(Bt - 1)s + Bt - 1.$$

This quadratic in s has roots

$$Bt - 1 \pm \sqrt{\{(Bt-1)^2 + Bt - 1\}}.$$

So $\gamma(s)$ is increasing for all $s \in \{Bt, Bt+1, \ldots, B\}$ with

$$s \leqslant Bt - 1 + \sqrt{\{(Bt - \tfrac{1}{2})^2 - \tfrac{1}{4}\}} =: s_0. \tag{14}$$

When $s_0 < B$, we must have $s^* \in \{2Bt - 2, 2Bt - 1\}$. In fact, by examining expression (13), we see that $\gamma(2Bt - 2) = \gamma(2Bt - 1)$. Also, from expression (14), we see that, when $t > \frac{1}{2}$, we have that $s_0 \geqslant B$, so $s^* = B$. So far, we have shown that

$$\mathbb{P}(X \geqslant t) \leqslant \max(b_1, b_2, b_3),$$

where bounds b_1, b_2 and b_3 are given by

$$b_1 := b_1(t, \eta, B) = \frac{2\eta - t + 1/B}{t + 1/B} \mathbb{1}_{\{\eta < t \leqslant \min(2\eta, 1)\}},$$

$$b_2 := b_2(t, \eta, B) = \frac{\eta}{2t - 1/B} \mathbb{1}_{\{\eta < t \leqslant 1/2\}},$$

$$b_3 := b_3(t, \eta, B) = \frac{2\eta(1 - t + 1/B)}{1 + 1/B} \mathbb{1}_{\{\max(\eta, 1/2) \leqslant t \leqslant 1\}}.$$

All that remains now is to determine which of b_1, b_2 and b_3 have the largest value. We first consider the case when $\eta \leqslant \frac{1}{3}$. When $t \leqslant \min(\frac{1}{2}, 2\eta)$,

$$\operatorname{sgn}(b_2 - b_1) = \operatorname{sgn}[\{t - \tfrac{3}{2}\eta - 1/(2B)\}(t - 1/B)].$$

Now, for $\frac{1}{2} < t \leqslant 2\eta$,

$$\frac{\partial b_3}{\partial t} = -\frac{2\eta}{1 + 1/B} \geqslant -\frac{(2\eta + 2/B)}{(t + 1/B)^2} = \frac{\partial b_1}{\partial t}.$$

Furthermore,

$$b_3\{\tfrac{1}{2} + 1/(2B), \eta, B\} = \eta \geqslant \frac{2\eta - \frac{1}{2} + 1/(2B)}{\frac{1}{2} + 3/(2B)} = b_1\{\tfrac{1}{2} + 1/(2B), \eta, B\}.$$

Putting this together gives the required bound for $\eta \leqslant \frac{1}{3}$.

When $\eta > \frac{1}{3}$, we can ignore b_2 as it is dominated by b_1. Comparing b_1 and b_3, we obtain the final cases of the bound.

We now prove theorem 2 as follows.

Recalling that $\mathbb{E}\{\tilde{\Pi}_B(k)\} = p^2_{k,\lfloor n/2 \rfloor}$, we follow the proof of lemma 1, but apply theorem 3 at the last step of expression (10) with $t = 2\tau - 1$ to deduce that, if the distribution of $\tilde{\Pi}_B(k)$ is unimodal, then

$$\mathbb{P}(k \in \hat{S}^{\mathrm{CPSS}}_{n,\tau}) \leqslant \mathbb{P}\{\tilde{\Pi}_B(k) \geqslant 2\tau - 1\} \leqslant C(\tau, B) p^2_{k,\lfloor n/2 \rfloor},$$

where $C(\tau, B)$ is given in the statement of theorem 2. The bound for $\mathbb{E}|\hat{S}^{\mathrm{CPSS}}_{n,\tau} \cap L_\theta|$ then follows in the same way that theorem 1 follows from lemma 1.

A.3. Proofs of results on r-concavity

A.3.1. Proof of proposition 1

Suppose that f is log-concave, so we may write $f = \exp(-\phi)$ where ϕ is a convex function. If $r < 0$, then $-r\phi$ is convex, and, as the exponential function is increasing and convex, $f^r = \exp(-r\phi)$ is convex.

Conversely, suppose that f is not log-concave, so there exist x, y and $\lambda \in (0, 1)$ with $f\{\lambda x + (1-\lambda)y\} < f(x)^\lambda f(y)^{1-\lambda}$. Then, as $M_r\{f(x), f(y); \lambda\} \to f(x)^\lambda f(y)^{1-\lambda}$ as $r \to 0$, we must have $f\{\lambda x + (1-\lambda)y\} < M_r\{f(x), f(y); \lambda\}$ for some $r < 0$, and so f cannot be r concave.

A.3.2. Proof of proposition 2

Let $I = \{1, \ldots, l\} \cup \{u, \ldots, B-1\}$. The conditions on f imply that

$$f_i > \min\{f_{i-1}, f_{i+1}\}, \qquad i \in I.$$

Then, as $M_r(f_{i-1}, f_{i+1}, \frac{1}{2}) \to \min\{f_{i-1}, f_{i+1}\}$ as $r \to -\infty$, for each $i \in I$, we may choose an $r_i < 0$ with

$$f_i > M_{r_i}(f_{i-1}, f_{i+1}; \tfrac{1}{2}). \tag{15}$$

Set $r = \min_{i \in I}(r_i)$. Observe that, as $M_r(a, b; \frac{1}{2})$ is increasing in r for all fixed a and b, the inequalities (15) are all satisfied when $r_i = r$. Thus $f_i^r \leqslant \frac{1}{2}(f_{i-1}^r + f_{i+1}^r)$ for all $i \in \{1, \ldots, B-1\}$, so f is r concave. □

By analogy with the unimodal case, let $\mathcal{F}_{r,\eta} = \{f \in \mathcal{F}_r : \mathcal{E}(f) \leqslant \eta\}$. In maximizing $\mathcal{T}_t$ over $\mathcal{F}_{r,\eta}$, there is again no loss of generality in assuming $0 < \eta < t$.

Lemma 3. For each $r < 0$, there is a maximizer of $\mathcal{T}_t$ in $\mathcal{F}_{r,\eta}$.

Proof. This proof is almost identical to that of lemma 2, except here we let $O = \{f \in \mathbb{R}^{B+1} : f^r \text{ is not convex}\}$. If $f \in O$, then there must exist $i_1 < i_2 < i_3$ such that

$$(i_3 - i_2) f_{i_1}^r + (i_2 - i_1) f_{i_3}^r < (i_3 - i_1) f_{i_2}^r$$

and it is clear that the above inequality must hold for all g in a sufficiently small open ball about f. Thus O is open, and the rest of the proof is clear.

Proposition 5. Any maximizer $f^* \in \mathcal{F}_{r,\eta}$ of $\mathcal{T}_t$ satisfies

(a) $\mathcal{E}(f^*) = \eta$ and
(b) f^{*r} is linear between f_0^{*r} and $f_{i_{M-1}}^{*r}$, where $i_M = B \max\{\mathrm{supp}(f^*)\}$.

Proof.

(a) Suppose that $\mathcal{E}(f^*) < \eta$. Define $i_m := B \min\{\mathrm{supp}(f^*)\}$. Let $\phi = f^{*r}$ and define a new sequence $\psi := (\psi_i : i = 0, \ldots, B)$ by

$$\psi_i = \begin{cases} \infty & \text{if } i < i_m, \\ \phi_i + \varepsilon_1 & \text{if } i = i_m, \\ \phi_i - \varepsilon_2 & \text{if } i > i_m \end{cases}$$

where $\varepsilon_1, \varepsilon_2 > 0$ are chosen such that $\Sigma_{i=0}^B \psi_i^{1/r} = 1$, but are sufficiently small that $\mathcal{E}(\psi^{1/r}) \leqslant \eta$. Then ψ is convex, so $\psi^{1/r} \in \mathcal{F}_{r,\eta}$. Since $\eta > 0$, we must have $\mathcal{T}_t(f^*) > 0$ so $\max\{\mathrm{supp}(f^*)\} \geqslant t$. Also, as we are assuming $\eta < \tau$, we must have $i_m < t$. Therefore $\mathcal{T}_t(\psi^{1/r}) > \mathcal{T}_t(f^*)$, which is a contradiction.

(b) Set $\phi = f^{*r}$, so ϕ is convex and $\phi^{1/r} = f^*$. Define $\psi' = (\psi'_0, \ldots, \psi'_B) \in \mathbb{R}^{B+1}$ as follows. Take $\psi'_i = \phi_i$ for $i \geqslant Bt$, but make ψ' linear between ψ'_0 and ψ'_{Bt} such that $g := \psi'^{1/r}$ has $\Sigma_{i=0}^B g_i = 1$ and $g_0 > 0$. This is possible since $\mathcal{E}(f^*) \leqslant \eta < t$, so $\min\{\mathrm{supp}(f^*)\} < t$. Note that ψ' is still convex since we must have $\psi'_{Bt} - \psi'_{Bt-1} \leqslant \phi_{Bt} - \phi_{Bt-1}$. Also $\mathcal{T}_t(g) = \mathcal{T}_t(f^*)$. Applying proposition 3, we see that $\mathcal{E}(g) \leqslant \mathcal{E}(f^*)$. Optimality of f^* means that equality must hold, so $f^* = g$ and also $\phi = \psi'$.

Now if ϕ is in fact linear between ϕ_0 and ϕ_B, condition (b) of the theorem is satisfied and we are done. Otherwise we may assume that ϕ is not a linear function between ϕ_{Bt-1} and ϕ_B and we can define ψ such that $\psi_i = \phi_i$ for $i \leqslant Bt$, that ψ is linear between ψ_{Bt-1} and ψ_{k-1} and $\psi_i = \infty$ for $i > k$. Here, k is chosen such that $g := \psi^{1/r}$ has $\Sigma_{i=0}^B g_i = 1$, and the convexity of ϕ ensures that such a $k \leqslant B$ exists. Applying proposition 3, we see that $\mathcal{E}(g) \leqslant \mathcal{E}(f^*)$. Since $\mathcal{T}_t(g) = \mathcal{T}_t(f^*)$, as before, optimality of f^* allows us to conclude that $f^* = g$.

A.4. Computing the r-concave tail probability bound

Here we describe a numerical algorithm that computes the function D that was defined in Section 3.3. Note that this is the maximum of $\mathcal{T}_t(f)$ over $f \in \mathcal{F}_{r,\eta}$. We shall only discuss the case where f^* is decreasing, as is always the case when $t > 2\eta$. The increasing case is very similar and less important for our application. We first note that we may parameterize the r-concave probability mass functions whose rth powers are linear as follows:

$$f_{a,k;i} = (a+i)^{1/r} \Big/ \sum_{j=0}^{k} (a+j)^{1/r}, \qquad i = 0, 1, \ldots, k, \tag{16}$$

where $k \leqslant B$. As $\mathcal{E}(f_{a,k})$ is strictly increasing in a, for each k, there is a unique a_k for which $\mathcal{E}(f_{a_k,k}) = \eta$. We also note here that a_k decreases with k. This is easily seen by observing that, regardless of the value of k, the parameter a in expression (16) determines the ratio of $f_{a,k;i}$ to $f_{a,k;j}$, for each i, j.

According to proposition 5, if $f^* \in \mathcal{F}_{r,\eta}$ maximizes $\mathcal{T}_t$, then f^{*r} is linear up to its penultimate support point. We can parameterize these in the following way. Write

$$\frac{\sum_{i=1}^{k} i(a+i)^{1/r} + (k+1)c}{\sum_{j=0}^{k} (a+j)^{1/r} + c} = B\eta,$$

and then solve for c:

$$c = c(a,k) = \frac{B\eta \sum_{j=0}^{k} (a+j)^{1/r} - \sum_{i=1}^{k} i(a+i)^{1/r}}{k+1-B\eta}.$$

We see that, as a ranges through $[a_{k+1}, a_k]$, we obtain all the relevant probability mass functions supported on $0, 1, \ldots, k+1$ via

$$g_{a,k;i} = \frac{(a+i)^{1/r}}{\sum_{j=0}^{k} (a+j)^{1/r} + c(a,k)}, \qquad i = 0, 1, \ldots, k,$$

$$g_{a,k;k+1} = \frac{c(a,k)}{\sum_{j=0}^{k} (a+j)^{1/r} + c(a,k)}.$$

The tail probability of $g_{a,k}$, when the threshold is t, is

$$\mathcal{T}_t(g_{a,k}) = 1 - \frac{(k+1-B\eta) \sum_{i=0}^{Bt-1} (a+i)^{1/r}}{\sum_{i=0}^{k} (k+1-i)(a+i)^{1/r}} \tag{17}$$

and we may maximize this over $a \in [a_{k+1}, a_k]$ to obtain an optimal a_k^* for each k. This is easily accomplished

by using a general purpose optimizer such as `optimize` in R. To summarize, we have the following simple procedure for computing $\mathcal{T}_t(f^*)$.

(a) For each $k \in \{t, \ldots, B\}$, determine (numerically) the solution in a_k to $\mathcal{E}(f_{a,k}) = \eta$.
(b) Find $a_k^* := \arg\max_{a \in [a_{k+1}, a_k]} \mathcal{T}_t(g_{a,k})$, for each k.
(c) Let $k^*(t) := \arg\max_k \mathcal{T}_t(g_{a_k^*,k})$.

Then $\mathcal{T}_t(f^*) = \mathcal{T}_t(g_{a^*_{k^*(t)},k^*(t)})$. When we wish to evaluate $\mathcal{T}_t(f^*)$ for a range of values of t, the process is simplified by the observation that $k^*(t)$ is increasing in t, and thus in step (b) we need only to consider those k which are at least $k^*(t - 1/B)$.

Using the algorithm described above, we have computed

$$\min\{D(\theta^2, 2\tau - 1, 50, -\tfrac{1}{2}), D(\theta, \tau, 100, -\tfrac{1}{4})\}$$

over a grid of θ- and τ-values (see Tables 1 and 2). An R implementation of the algorithm is available from both authors' Web sites.

References

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D. and Levine, A. J. (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natn. Acad. Sci. USA*, **96**, 6745–6750.

Amaratunga, D. and Cabrera, J. (2004) *Exploration and Analysis of DNA Microarray and Protein Array Data*. New York: Wiley-Interscience.

Bach, F. (2008) Bolasso: model consistent lasso estimation through the bootstrap. In *Proc. 25th Int. Conf. Machine Learning*, pp. 33–40. New York: Association for Computing Machinery.

Biau, G., Cérou, F. and Guyader, A. (2010) On the rate of convergence of the bagged nearest neighbor estimate. *J. Mach. Learn. Res.*, **11**, 687–712.

Breiman, L. (1996) Bagging predictors. *Mach. Learn.*, **24**, 123–140.

Breiman, L. (1999) Using adaptive bagging to debias regressions. *Technical Report*. Department of Statistics, University of California, Berkeley.

Bühlmann, P. and Yu, B. (2002) Analyzing bagging. *Ann. Statist.*, **30**, 927–961.

Cule, M., Samworth, R. and Stewart, M. (2010) Maximum likelihood estimation of a multi-dimensional log-concave density (with discussion). *J. R. Statist. Soc.* B, **72**, 545–607.

Dharmadhikari, S. and Joag-Dev, K. (1988) *Unimodality, Convexity and Applications*. Boston: Academic Press.

Dudoit, S., Fridlyand, J. and Speed, T. P. (2002) Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Am. Statist. Ass.*, **97**, 77–87.

Dümbgen, L. and Rufibach, K. (2009) Maximum likelihood estimation of a log-concave density and its distribution function: basic properties and uniform consistency. *Bernoulli*, **15**, 40–68.

Dümbgen, L., Samworth, R. J. and Schuhmacher, D. (2012) Stochastic search for semiparametric linear regression models. In *From Probability to Statistics and Back: High-dimensional Models and Processes; a Festschrift in Honor of Jon Wellner* (eds M. Banerjee and F. Bunea). Beachwood: Institute of Mathematical Statistics. To be published.

Fan, J. and Lv, J. (2010) A selective overview of variable selection in high dimensional feature space. *Statist. Sin.*, **20**, 101–148.

Fan, J., Samworth, R. and Wu, Y. (2009) Ultrahigh dimensional feature selection: beyond the linear model. *J. Mach. Learn. Res.*, **10**, 2013–2038.

Friedman, J., Hastie, T. and Tibshirani, R. (2010) Regularization paths for generalized linear models via coordinate descent. *J. Statist. Softwr.*, **33**, 1–22.

Hall, P. and Samworth, R. J. (2005) Properties of bagged nearest neighbour classifiers. *J. R. Statist. Soc.* B, **67**, 363–379.

Han, Y. and Yu, L. (2010) A variance reduction framework for stable feature selection. In *Proc. 10th Int. Conf. Data Mining*, pp. 206–215. Sydney: Institute of Electrical and Electronics Engineers Computer Society.

Kalousis, A., Prados, J. and Hilario, M. (2007) Stability of feature selection algorithms: a study on high-dimensional spaces. *Knowl. Inform. Syst.*, **12**, 95–116.

Koenker, R. and Mizera, I. (2010) Quasi-concave density estimation. *Ann. Statist.*, **38**, 2998–3027.

Kuncheva, L. (2007) A stability index for feature selection. In *Proc. 25th Int. Multi-conf. Artificial Intelligence and Applications*, pp. 390–395. Calgary: ACTA.

Lange, T., Braun, M., Roth, V. and Buhmann, J. (2003) Stability-based model selection. In *Advances in Neural Information Processing Systems*, vol. 15 (eds S. Becker, S. Thrun and K. Obermayer), pp. 617–624. Cambridge: MIT Press.

Loscalzo, S., Yu, L. and Ding, C. (2009) Consensus group based stable feature selection. In *Proc. 15th Int. Conf. Knowledge Discovery and Data Mining*, pp. 567–576. New York: Association for Computing Machinery.

Meinshausen, N. and Bühlmann, P. (2010) Stability selection (with discussion). *J. R. Statist. Soc.* B, **72**, 417–473.
R Development Core Team (2010) *R: a Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.
Saeys, Y., Abeel, T. and Peer, Y. V. (2008) Robust feature selection using ensemble feature selection techniques. In *Proc. Eur. Conf. Machine Learning.*, pp. 313–325. Berlin: Springer.
Samworth, R. J. (2011) Optimal weighted nearest neighbour classifiers. *Arxiv Preprint*. (Available from `http://arxiv.org/pdf/1101.5783`.)
Seregin, A. and Wellner, J. A. (2010) Nonparametric estimation of convex-transformed densities. *Ann. Statist.*, **38**, 3751–3781.
Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *J. R. Statist. Soc.* B, **58**, 267–288.
Walther, G. (2002) Detecting the presence of mixing with multiscale maximum likelihood. *J. Am. Statist. Ass.*, **97**, 508–513.