Integrated path stability selection

Omar Melikechi

Department of Biostatistics, Harvard T.H. Chan School of Public Health
and
Jeffrey W. Miller
Department of Biostatistics, Harvard T.H. Chan School of Public Health

August 27, 2024

Abstract

Stability selection is a popular method for improving feature selection algorithms. One of its key attributes is that it provides theoretical upper bounds on the expected number of false positives, E(FP), enabling control of false positives in practice. However, stability selection often selects very few features, resulting in low sensitivity. This is because existing bounds on E(FP) are relatively loose, causing stability selection to overestimate the number of false positives. In this paper, we introduce a novel approach to stability selection based on integrating stability paths rather than maximizing over them. This yields upper bounds on E(FP) that are orders of magnitude stronger than previous bounds, leading to significantly more true positives in practice for the same target E(FP). Furthermore, our method takes the same amount of computation as the original stability selection algorithm, and only requires one user-specified parameter, which can be either the target E(FP) or target false discovery rate. We demonstrate the method on simulations and real data from prostate and colon cancer studies.

Keywords: Bootstrap, error control, feature selection, resampling, variable selection

1 Introduction

Stability selection is a widely used method that uses subsampling to improve feature selection algorithms (Meinshausen and Bühlmann, 2010). It is attractive due to its generality, simplicity, and theoretical control on the expected number of false positives, E(FP), sometimes denoted EV and called the *per-family error rate* in the literature. Despite these favorable qualities, existing theory for stability selection—which heavily informs its implementation—provides relatively weak bounds on E(FP), resulting in a diminished number of true positives (Alexander and Lange, 2011; Hofner et al., 2015; Li et al., 2013; Wang et al., 2020). Furthermore, stability selection requires users to specify two of three parameters: the target E(FP), a selection threshold, and the expected number of selected features. Many works have shown that stability selection is sensitive to these choices, making it difficult to tune for good performance (Haury et al., 2012; Li et al., 2013; Wang et al., 2020; Werner, 2023; Zhou et al., 2013).

The limitations of stability selection are illustrated in Figure 1. Here, we consider the canonical example of a linear model $Y_i = \beta^\intercal X_i + \epsilon_i$ for $i \in \{1, \dots, n\}$, where $X_i \in \mathbb{R}^p$, $n = 200$, and $p = 1000$. The features and noise are generated as $X_{ij} \sim \mathcal{N}(0, 1)$ independently, and $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ independently given $X_1, \dots, X_n$, where $\sigma^2 = \frac{1}{2n} \sum_{i=1}^n (\beta^\intercal X_i)^2$ so that the signal-to-noise ratio is 2. The coefficient vector β has $s = 20$ nonzero entries $\beta_j \sim \text{Uniform}(-1, 1)$, located at randomly selected $j \in \{1, \dots, p\}$, and all other entries are 0. The results in Figure 1 are obtained by simulating 100 random data sets as above and running stability selection using lasso as the baseline selection algorithm (Tibshirani, 1996). A selected feature is a *true positive* if its corresponding β entry is nonzero, and is a *false positive* otherwise.

The horizontal axis in Figure 1 is the target E(FP). The vertical axes show the actual

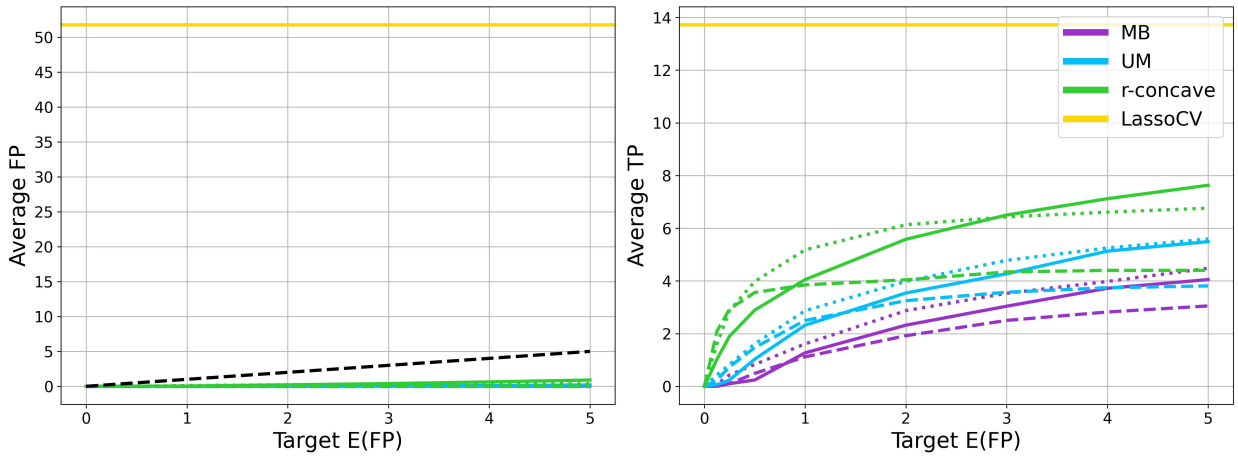

Figure 1: *Tradeoff between FP and TP.* Stability selection is overly conservative, yielding a small number of false positives (FP) at the expense of a small number of true positives (TP). Meanwhile, lasso has a high TP, but also a very high FP. (*Left*) Average FP versus target E(FP) for the original stability selection method of Meinshausen and Bühlmann (2010), denoted MB, and the unimodal (UM) and r-concave methods of Shah and Samworth (2013) at thresholds $\tau = 0.6$, 0.75, and 0.9 (solid, dotted, and dashed lines, respectively), as well as lasso with cross-validation, which does not depend on the horizontal axis. The black dashed line is the target E(FP). (*Right*) Average TP versus target E(FP) for the same methods.

numbers of false positives (FP) and true positives (TP), averaged over the 100 data sets. The stability selection methods have relatively few true positives on average, and usually have 0 false positives, undershooting the target E(FP). On the other hand, lasso with regularization parameter chosen by cross-validation (LassoCV) produces around 14 true positives, but over 50 false positives. This is a known trade-off: Typically, stability selection is too conservative, while lasso with cross-validation is not conservative enough (Alexander and Lange, 2011; Leng et al., 2006; Zou, 2006). Figure 1 also shows that the performance of stability selection, especially the r-concave version, depends on the threshold parameter, τ. Furthermore, while the UM and r-concave versions outperform MB, both rely on additional assumptions and have less transparent bounds on E(FP); in particular, r-concave requires an additional algorithm and its E(FP) bound does not admit a closed form (Shah and Samworth, 2013).

In this article, we introduce *integrated path stability selection* (IPSS) to address these limitations. We prove that IPSS satisfies bounds on E(FP) that are orders of magnitude stronger than existing stability selection bounds, thus yielding many more true positives for the same target E(FP). This is a key advantage since the actual number of false positives is unknown in practice, so stronger bounds enable one to increase the true positive rate while maintaining guaranteed control over E(FP). Another advantage of IPSS is that it requires just one user-specified parameter, either the target E(FP) or the target false discovery rate. This makes IPSS easier to use than stability selection which, in addition to the target E(FP), requires specification of either the selection threshold or the expected number of selected features, which can be difficult to tune. Finally, IPSS is simple to implement and requires no more computation than stability selection.

The rest of the article is organized as follows. In Section 2, we introduce our proposed

methodology. Section 3 provides a brief review of previous work. In Section 4, we present our theoretical results. Section 5 contains an extensive simulation study, and in Section 6, we apply IPSS to identify proteins/genes related to prostate cancer and colon cancer. We conclude in Section 7 with a brief discussion.

2 Integrated path stability selection

In this section, we define the setup of the problem (Section 2.1), describe the stability selection algorithm (Section 2.2), and introduce IPSS (Section 2.3).

2.1 Setup

Suppose $S \subseteq \{1, \ldots, p\}$ is an unknown subset to be estimated from independent and identically distributed data $Z_1, \ldots, Z_n$. Let $\hat{S}_\lambda(Z_{1:n}) \subseteq \{1, \ldots, p\}$ be an estimator of S, where $Z_{1:n} = (Z_1, \ldots, Z_n)$, $\lambda > 0$ is a parameter, and $\hat{S}_\lambda$ can be a random function such as a stochastic optimization algorithm. In regression, we have $Z_i = (X_i, Y_i)$ where $X_i \in \mathbb{R}^p$ is a vector of features and $Y_i \in \mathbb{R}$ is a response variable for each $i \in \{1, \ldots, n\}$. As illustrated in Section 1, a canonical estimator in this setting is the lasso algorithm (Tibshirani, 1996), in which case $\hat{S}_\lambda(Z_{1:n}) = \{j : \hat{\beta}_j \neq 0\}$ where

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \ \frac{1}{2} \sum_{i=1}^{n} (Y_i - \beta^{\mathtt{T}} X_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j|.$$

In an unsupervised learning setting such as graphical lasso, we have $Z_i = X_i \in \mathbb{R}^p$, without a response variable (Friedman et al., 2008). See Section 2 of Meinshausen and Bühlmann (2010) for details on applying stability selection to graphical lasso as well as other examples of feature selection algorithms amenable to stability selection.

We will frequently need to compute the estimator using a subset of the data, say, $Z_A = (Z_i : i \in A)$ for a given $A \subseteq \{1, \ldots, n\}$. In this case, $\hat{S}_\lambda(Z_A)$ denotes the estimator computed using only the data in Z_A. A key quantity is the probability that feature j is selected when using half of the data. We denote this *selection probability* by

$$\pi_j(\lambda) = \mathbb{P}\big(j \in \hat{S}_\lambda(Z_{1:\lfloor n/2 \rfloor})\big). \tag{2.1}$$

Stability selection and IPSS employ resampling-based estimators of $\pi_j(\lambda)$, called the *estimated selection probabilities* $\hat{\pi}_j(\lambda)$, computed using Algorithm 1. The *stability paths* $\lambda \mapsto \hat{\pi}_j(\lambda)$, which will be important in what follows, are shown in Figure 2. Both $\pi_j(\lambda)$ and $\hat{\pi}_j(\lambda)$ depend on n, but n is suppressed from notation since we always treat it as an arbitrary fixed value.

Algorithm 1 (Estimated selection probabilities)

Input: Data $Z_1, \ldots, Z_n$, selection algorithm $\hat{S}_\lambda$, parameter grid Λ, number of iterations B.

1: **for** $b = 1, \ldots, B$ **do**
2: Randomly select disjoint $A_{2b-1}, A_{2b} \subseteq \{1, \ldots, n\}$ with $|A_{2b-1}| = |A_{2b}| = \lfloor n/2 \rfloor$.
3: **for** $\lambda \in \Lambda$ **do**
4: Evaluate $\hat{S}_\lambda(Z_{A_{2b-1}})$ and $\hat{S}_\lambda(Z_{A_{2b}})$.
5: **end for**
6: **end for**

Output: Estimated selection probabilities $\hat{\pi}_j(\lambda) = \frac{1}{2B} \sum_{b=1}^{2B} \mathbb{1}\big(j \in \hat{S}_\lambda(Z_{A_b})\big)$.

In Algorithm 1 and throughout this work, $\mathbb{1}(\cdot)$ denotes the indicator function: $\mathbb{1}(E) = 1$ if E is true and $\mathbb{1}(E) = 0$ otherwise. Note that $\hat{S}_\lambda$ is evaluated on both disjoint subsets, A_{2b-1} and A_{2b}, at each iteration of Algorithm 1. This technique of using complementary pairs of subsets was introduced by Shah and Samworth (2013). By contrast, in the original stability

selection algorithm of Meinshausen and Bühlmann (2010), $\hat{S}_\lambda$ is applied to only one subset of size $\lfloor n/2 \rfloor$ at each iteration. This slight modification simplifies the assumptions needed for the theory (Shah and Samworth, 2013). The choice of $\lfloor n/2 \rfloor$ samples is required for the theory of stability selection to hold, both in this paper and in previous works.

2.2 Stability selection

Once the $\hat{\pi}_j$ values are computed, the set of features selected by stability selection is

$$\hat{S}_{\text{SS}} = \left\{ j : \max_{\lambda \in \Lambda} \hat{\pi}_j(\lambda) \geq \tau \right\} \tag{2.2}$$

where $\Lambda = [\lambda_{\min}, \lambda_{\max}] \subseteq (0, \infty)$ is an interval defined below and $\tau \in (0, 1)$ is a user-specified threshold. The upper endpoint $\lambda_{\max}$ is inconsequential provided it is large enough that all features have small selection probability, which is easy to determine empirically. Choosing $\lambda_{\min}$ is considerably more subtle, since many or even all features satisfy $\hat{\pi}_j(\lambda) \geq \tau$ as $\lambda \to 0$. While there is no consensus on how to choose $\lambda_{\min}$ (Li and Zhang, 2017; Zhou et al., 2013), a standard approach is to use theoretical upper bounds on E(FP) as follows.

The MB, UM, and r-concave versions of stability selection all satisfy theoretical upper bounds of the form $\text{E(FP)} \leq \mathcal{B}(q, \tau)$, which depend on the method (MB, UM, or r-concave), the average number of features selected over Λ, $q = \text{E}\big|\bigcup_{\lambda \in \Lambda} \hat{S}_\lambda(Z_{1:\lfloor n/2 \rfloor})\big|$, and the threshold, τ; see Section 4.3. To determine $\lambda_{\min}$, two of the following three quantities must be specified: (i) the target E(FP), denoted E(FP)_*, (ii) the threshold, τ, and (iii) the target number of features selected, q_*. The third quantity is then obtained by setting $\text{E(FP)}_* = \mathcal{B}(q_*, \tau)$ and solving. Once q_* is determined, $\lambda_{\min} = \sup\left\{\lambda \in (0, \lambda_{\max}) : \text{E}\big|\bigcup_{\lambda' \in [\lambda, \lambda_{\max}]} \hat{S}_{\lambda'}(Z_{1:\lfloor n/2 \rfloor})\big| \geq q_*\right\}$ is empirically estimated and $\Lambda = [\lambda_{\min}, \lambda_{\max}]$ is used in Equation 2.2.

The above construction elucidates some of the shortcomings of stability selection and motivates our formulation of IPSS. First, the inequalities $\mathrm{E(FP)} \leq \mathcal{B}(q, \tau)$ are replaced by equalities in order to determine $\lambda_{\min}$ in a way that controls FP. Thus, while the recommended procedure does typically keep the actual FP smaller than $\mathrm{E(FP)}_*$, it may be much smaller, as shown in Figures 1 and 3. This overconservative tendency leads to a lower TP than necessary. More precisely, weak bounds on E(FP) lead to large values of $\lambda_{\min}$, which prevent true features from being selected because their stability paths have not yet distinguished themselves from the noise (Figure 2). Second, $\mathrm{E(FP)}_*$, q_*, and τ are interdependent, making it difficult to select these parameters in practice. Meinshausen and Bühlmann (2010) recommended taking $\tau \in [0.6, 0.9]$, but stability selection is sensitive to τ even when restricted to this interval (Li et al., 2013; Wang et al., 2020). Nevertheless, τ must be specified in most cases because one usually has little *a priori* knowledge to inform the choice of q_*. Finally, while one can in principle use a smaller $\lambda_{\min}$, it is unclear what value to choose and doing so would invalidate the E(FP) control guarantee, making it hard to interpret the results.

2.3 Integrated path stability selection

For IPSS, the $\hat{\pi}_j$ values are computed using Algorithm 1 and the set of features selected is

$$\hat{S}_{\mathrm{IPSS},f} = \left\{ j : \textstyle\int_\Lambda f(\hat{\pi}_j(\lambda)) \mu(d\lambda) \geq \tau \right\} \tag{2.3}$$

where the probability measure μ, interval $\Lambda = [\lambda_{\min}, \lambda_{\max}] \subseteq (0, \infty)$, function $f : [0, 1] \to \mathbb{R}$, and threshold τ are defined in Section 2.3.1 below. Unlike the relatively coarse maximum criterion used by stability selection, the integral in Equation 2.3 incorporates information about the stability paths over a wide range of λ values. For example, Figure 2 compares

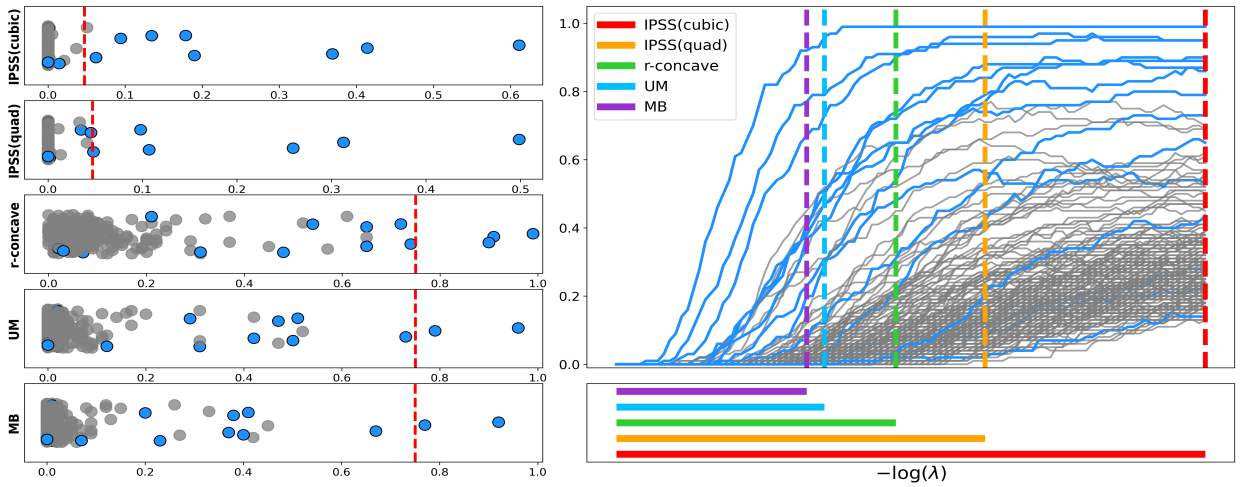

Figure 2: Linear regression with independent design as in Section 1 with $n = 150$, $p = 200$, SNR $= 1$, and $s = 15$ true features. (*Left*) The horizontal axis is the score for each feature j, that is, $\int_\Lambda f(\hat{\pi}_j(\lambda))\mu(d\lambda)$ for IPSS and $\max_{\lambda\in\Lambda} \hat{\pi}_j(\lambda)$ for the others. The vertical axis is random jitter for visualization. True features are shown in blue, and the red vertical lines show the threshold τ separating selected and unselected features for each method. We use $\mathrm{E(FP)}_* = 1$ for all methods, and for stability selection, we use $\tau = 0.75$. MB, UM, r-concave, IPSS(quad), and IPSS(cubic) identify 2, 2, 3, 6, and 8 true positives, respectively. All methods have 0 false positives except IPSS(cubic), which has 1, agreeing with the target E(FP). (*Right*) Estimated stability paths $\hat{\pi}_j(\lambda)$. Vertical dashed lines show $-\log(\lambda_{\min})$ and horizontal lines below the plot show the intervals $[-\log(\lambda_{\max}), -\log(\lambda_{\min})]$ for each method.

methods on the linear regression simulation from Section 1. IPSS captures the fact that the stability paths of the true features rise at different rates, with some overtaking many of the false features more gradually than others, a point missed by MB, UM, and r-concave.

2.3.1 Parameters

The role of f in Equation 2.3 is to transform the selection probabilities $\hat{\pi}_j$ to improve performance. Different functions f yield different bounds on E(FP), and tighter bounds very often yield more true positives at the same target E(FP). Thus, we emphasize that f is determined by theory and is not a free parameter to be selected by the user; indeed, an f without a corresponding bound on E(FP) is useless from the point of view of IPSS. In Theorem 4.1 we establish upper bounds on E(FP) for the class of functions

$$h_m(x) = (2x-1)^m \mathbb{1}(x \geq 0.5), \tag{2.4}$$

where $m \in \mathbb{N}$ (the h stands for "half", since h_m is positive on half the unit interval). We focus on IPSS with $f = h_2$ and $f = h_3$, denoted by IPSS(quad) and IPSS(cubic), respectively.

For IPSS, the interval $\Lambda = [\lambda_{\min}, \lambda_{\max}]$ is defined as follows. The upper endpoint $\lambda_{\max}$ is the same as in stability selection and equally inconsequential. The lower endpoint $\lambda_{\min}$ is based on a bound of the form $\mathrm{E(FP)} \leq \mathcal{I}(\lambda, \lambda_{\max})/\tau$ that depends on f, where $\mathcal{I}(\lambda, \lambda_{\max})$ is an integral over $[\lambda, \lambda_{\max}]$ such as in Equations 4.4 and 4.5. Specifically, we define

$$\lambda_{\min} = \inf\left\{\lambda \in (0, \lambda_{\max}) : \mathcal{I}(\lambda, \lambda_{\max}) \leq C\right\} \tag{2.5}$$

for a fixed cutoff, C. We always use $C = 0.05$, but extensive empirical results in Section S4.2 show that IPSS depends little on C over a wide range of settings. Further details about the construction of Λ and the evaluation of Equation 2.5 are in Section S1.1.

The probability measure μ in Equation 2.3 weights the different values of λ. Like C, experiments in Section S4.2 show that results depend little on μ. We always use $\mu(d\lambda) = z^{-1}\lambda^{-1}d\lambda$, where $z = \int_\Lambda \lambda^{-1}d\lambda$ is a normalizing constant. This corresponds to averaging over Λ on a log scale, as is common when working with regularization values.

The threshold τ is determined by specifying a target E(FP), denoted $\mathrm{E(FP)}_*$, and replacing the inequality in Equation 4.2 with an equality, representing a worst-case scenario under the assumptions of Theorem 4.1. This gives $\tau = \mathcal{I}(\Lambda)/\mathrm{E(FP)}_*$ and Equation 2.3 becomes

$$\hat{S}_{\mathrm{IPSS}} = \left\{ j : \textstyle\int_\Lambda f(\hat{\pi}_j(\lambda))\mu(d\lambda) \geq \dfrac{\mathcal{I}(\Lambda)}{\mathrm{E(FP)}_*} \right\} = \left\{ j : \mathrm{E(FP)}_* \geq \texttt{efp}(j) \right\} \tag{2.6}$$

where, for each $j \in \{1, \ldots, p\}$, the *efp score* of j is defined as

$$\texttt{efp}(j) = \min \left\{ \frac{\mathcal{I}(\Lambda)}{\int_\Lambda f(\hat{\pi}_j(\lambda))\mu(d\lambda)},\ p+1 \right\}. \tag{2.7}$$

Under the assumptions of Theorem 4.1, $\texttt{efp}(j)$ is the smallest bound on E(FP) if j is to be selected. The minimum in Equation 2.7 accounts for $\int_\Lambda f(\hat{\pi}_j(\lambda))\mu(d\lambda)$ being 0, in which case j is never selected since E(FP) is at most the number of features, p.

Equations 2.6 and 2.7 help to understand why IPSS does not depend strongly on μ and C (which determines Λ). The key quantity is $\mathcal{I}(\Lambda)/\int_\Lambda f(\hat{\pi}_j(\lambda))\mu(d\lambda)$, in which the numerator and denominator are expectations with respect to μ over Λ. IPSS depends primarily on the integrands of these functions, which are determined by the form of the upper bound and the function f, rather than the probability measure μ and its support Λ.

In summary, the choices of f and $n/2$ are determined by theory, μ and C (which determines Λ) are always fixed as above for all forms of IPSS, and results are robust to the choice of μ and C. The threshold τ is determined by the preceding parameters and the target E(FP). Finally, numerous works have noted that B is inconsequential provided it is sufficiently large;

$B = 50$ is a common choice, and the one we use throughout this work (Shah and Samworth, 2013). Thus, when implementing IPSS, the user only needs to specify E(FP).

2.3.2 Computation

Algorithm 2 is a step-by-step description of IPSS. The grid in Step 2 is used to accurately and efficiently approximate all integrals in the algorithm by simple Riemann sums (Proposition S1.1). We always use $r = 100$ grid points. Like many of the other parameters, r is inconsequential provided it is sufficiently large; in our experience, values greater than 25 suffice. This is because the functions h_m, the stability paths, the integrand in the upper bound $\mathcal{I}(\lambda_{\min}, \lambda_{\max})$, and the measure μ are all very numerically stable. The bounds referred to in Step 3 are in Theorem 4.2 for IPSS with h_2 and h_3. We find no discernible difference in computation time between IPSS and MB. This is because IPSS and MB both estimate the selection probabilities via Algorithm 1, which is much more expensive than evaluation of either Equation 2.2 or Equation 2.3. For more on the computational requirements of Algorithm 1, see Meinshausen and Bühlmann (2010, Section 2.6).

2.3.3 False discovery rate

IPSS and other forms of stability selection focus on E(FP) because it is interpretable and theoretically tractable. Another quantity of interest, the *false discovery rate* (FDR), is the expected ratio between the number of false positives and the total number of features selected, $\text{FDR} = \text{E}(\text{FP}/(\text{TP} + \text{FP}))$. When p is large, the FDR is well-approximated by $\text{E}(\text{FP})/\text{E}(\text{TP} + \text{FP})$ (Storey and Tibshirani, 2003). Thus, making the additional approximation $|\hat{S}_{\text{IPSS}}| \approx \text{E}(\text{TP} + \text{FP})$, we have $\text{FDR} \approx \text{E}(\text{FP})/|\hat{S}_{\text{IPSS}}|$. Relabeling features by their

Algorithm 2 (Integrated path stability selection)

Input: Data $Z_1, \ldots, Z_n$, selection algorithm $\hat{S}$, number of iterations B, function f, probability measure μ, target $\mathrm{E(FP)}_*$, integral cutoff value C, and number of grid points r.

1: Compute $\lambda_{\min}$ and $\lambda_{\max}$ as described above and in Section S1.1.

2: Partition $\Lambda = [\lambda_{\min}, \lambda_{\max}]$ into r grid points, typically on a log scale.

3: Compute $\mathcal{I}(\lambda_{\min}, \lambda_{\max})$ using the relevant upper bound on E(FP) and Proposition S1.1.

4: Set $\tau = \mathcal{I}(\lambda_{\min}, \lambda_{\max})/\mathrm{E(FP)}_*$.

5: Estimate selection probabilites via Algorithm 1 with $Z_1, \ldots, Z_n$, $\hat{S}$, Λ, and B.

6: Approximate $\hat{s}_j = \int_\Lambda f(\hat{\pi}_j(\lambda))\mu(d\lambda)$ using Proposition S1.1.

Output: Selected features $\hat{S}_{\mathrm{IPSS},f} = \{j : \hat{s}_j \geq \tau\}$.

efp scores so that $\texttt{efp}(1) \leq \texttt{efp}(2) \leq \cdots \leq \texttt{efp}(p)$, the set of features $\{1, \ldots, j\}$ has an approximate FDR that is bounded above by $\texttt{efp}(j)/j$ for each $j \in \{1, \ldots, p\}$. Hence, instead of specifying $\mathrm{E(FP)}_*$, one can either (i) specify a target FDR, say FDR_*, and choose the largest j such that $\texttt{efp}(j)/j \leq \mathrm{FDR}_*$, or (ii) choose j to minimize $\texttt{efp}(j)/j$. The resulting set of selected features is then $\{1, \ldots, j\}$, that is, the features with the j smallest efp scores. Various FDR results from the simulations in Section 5 are reported in Section S4.

3 Previous work

Stability selection was introduced by Meinshausen and Bühlmann (2010) and refined by Shah and Samworth (2013). These remain the preeminent works on stability selection and are the most commonly implemented versions to date. Zhou et al. (2013) provide the only other work we are aware of that aims to improve upon the selection criterion in Equation 2.2

and the E(FP) bound. They propose *top-k stability selection*, which averages over the k largest selection probabilities for each feature. The special case of $k = 1$ is stability selection. While Zhou et al. (2013) provide theory for top-k stability selection, their improvement upon the E(FP) upper bound of Meinshausen and Bühlmann (2010) is considerably weaker than the improved bound provided by IPSS; compare Zhou et al. (2013, Theorem 3.1) to our Theorem 4.1. Moreover, introducing k increases the number of parameters, and it is shown that results are sensitive to this choice (Zhou et al., 2013).

It is far more common for stability selection to be modified on an *ad hoc* basis to mitigate its sensitivity to parameters and overly conservative results. An example is the TIGRESS method of Haury et al. (2012), which uses stability selection to infer gene regulatory networks. To reduce sensitivity to the stability selection parameters, they use a selection criterion that averages over selection probabilities. It turns out that this is the special case of IPSS with $f(x) = x$ (the function w_1 in Section S3), whose analysis is relegated to the supplement because its bound on E(FP) is not nearly as strong as those in Theorems 4.1 and 4.2. Another example is in the work of Maddu et al. (2022), where stability selection is used to learn differential equations. There the authors use a selection criterion based only on selection probabilities at the smallest regularization parameter. Finally, a common approach is to use stability selection in conjunction with other methods. Examples include stability selection with boosting (Hofner et al., 2015) and grouping features prior to applying stability selection, which has been done in genome-wide association studies (Alexander and Lange, 2011). IPSS can be used instead of stability selection in such methods at no additional computational cost and with fewer required parameters.

4 Theory

In this section, we present our theoretical results (Section 4.2) and compare them to those of Meinshausen and Bühlmann (2010) and Shah and Samworth (2013) (Section 4.3). Our main result, Theorem 4.1, establishes a bound on E(FP) for IPSS with the functions h_m defined in Equation 2.4. Theorem 4.2 gives simplified formulas for this bound that we use in practice. The proofs of all results in this section are in Section S2. Additional results for other choices of f and their proofs are in Section S3.

4.1 Preliminaries

It is assumed that the random variables $Z_1, \ldots, Z_n$, the random subsets $A_1, \ldots, A_{2B}$, and any randomness in the feature selection algorithm $\hat{S}$ are defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Furthermore, E always denotes expectation with respect to $\mathbb{P}$. Let Λ be a Borel measurable subset of $(0, \infty)$ equipped with the Borel sigma-algebra, let μ be a probability measure on Λ, and assume $\hat{S}_\lambda(Z_A)$ is measurable as a function on $\Lambda \times \Omega$ for all $A \subseteq \{1, \ldots, n\}$.

4.2 Main results

The following condition is used in Theorem 4.1. Recall that S is the unknown subset of true features, and $S^c = \{1, \ldots, p\} \setminus S$ is its complement. Let $q(\lambda) = \mathrm{E}|\hat{S}_\lambda(Z_{1:\lfloor n/2 \rfloor})|$ denote the expected number of variables selected by $\hat{S}_\lambda$ on half the data.

Condition 1. *We say* Condition 1 holds for m *if for all* $\lambda \in \Lambda$,

$$\max_{j \in S^c} \mathbb{P}\left(j \in \bigcap_{b=1}^{m} \left(\hat{S}_\lambda(Z_{A_{2b-1}}) \cap \hat{S}_\lambda(Z_{A_{2b}}) \right) \right) \leq (q(\lambda)/p)^{2m}. \tag{4.1}$$

Equation 4.1, discussed in greater detail below, says the probability that any non-true feature j is selected by both $\hat{S}_\lambda(Z_{A_{2b-1}})$ and $\hat{S}_\lambda(Z_{A_{2b}})$ in m resampling iterations is no greater than the $2m$th power of the expected proportion of features selected by $\hat{S}_\lambda$ using half the data.

Theorem 4.1. *Let $\tau \in (0, 1]$ and $m \in \mathbb{N}$. Define $\hat{S}_{\text{IPSS},h_m}$ as in Equations 2.3 and 2.4. If Condition 1 holds for all $m' \in \{1, \ldots, m\}$, then*

$$\mathrm{E}(\text{FP}) = \mathrm{E}|\hat{S}_{\text{IPSS},h_m} \cap S^c| \leq \frac{p}{\tau B^m} \sum_{k_1+\cdots+k_B=m} \frac{m!}{k_1!k_2!\cdots k_B!} \int_\Lambda (q(\lambda)/p)^{2\sum_b \mathbb{1}(k_b \neq 0)} \mu(d\lambda) \quad (4.2)$$

where B is the number of subsampling steps in Algorithm 1 and the sum is over all nonnegative integers $k_1, \ldots, k_B$ such that $k_1 + \cdots + k_B = m$.

Equation 4.2 bounds the expected number of false positives when using IPSS with h_m. The following theorem shows that the bound simplifies considerably for certain choices of m.

Theorem 4.2. *Let $\tau \in (0, 1]$. If Condition 1 holds for $m = 1$, then IPSS with h_1 satisfies*

$$\mathrm{E}(\text{FP}) \leq \frac{1}{\tau} \int_\Lambda \frac{q(\lambda)^2}{p} \mu(d\lambda); \quad (4.3)$$

if Condition 1 holds for $m \in \{1, 2\}$, then IPSS with h_2 satisfies

$$\mathrm{E}(\text{FP}) \leq \frac{1}{\tau} \int_\Lambda \left(\frac{q(\lambda)^2}{Bp} + \frac{(B-1)q(\lambda)^4}{Bp^3} \right) \mu(d\lambda); \quad (4.4)$$

and if Condition 1 holds for $m \in \{1, 2, 3\}$, then IPSS with h_3 satisfies

$$\mathrm{E}(\text{FP}) \leq \frac{1}{\tau} \int_\Lambda \left(\frac{q(\lambda)^2}{B^2p} + \frac{3(B-1)q(\lambda)^4}{B^2p^3} + \frac{(B-1)(B-2)q(\lambda)^6}{B^2p^5} \right) \mu(d\lambda). \quad (4.5)$$

Taking the limit as $B \to \infty$, Equations 4.4 and 4.5 become

$$\limsup_{B\to\infty} \mathrm{E}(\text{FP}) \leq \frac{1}{\tau p^3} \int_\Lambda q(\lambda)^4 \mu(d\lambda) \quad \text{and} \quad \limsup_{B\to\infty} \mathrm{E}(\text{FP}) \leq \frac{1}{\tau p^5} \int_\Lambda q(\lambda)^6 \mu(d\lambda). \quad (4.6)$$

Although we do not use these asymptotic bounds, the pattern from h_1 (Equation 4.3) to h_2 to h_3 (Equation 4.6) provides insight into the relationships between E(FP), p, and h_m.

Condition 1 holds for $m = 1$ whenever $\max_{j \in S^c} \pi_j(\lambda) \leq q(\lambda)/p$ for all $\lambda \in \Lambda$ since $Z_1, \ldots, Z_n$ are i.i.d. and independent of $A_1, \ldots, A_{2B}$, and thus for any $j \in S^c$,

$$\begin{aligned} \mathbb{P}\big(j \in \hat{S}_\lambda(Z_{A_{2b-1}}) \cap \hat{S}_\lambda(Z_{A_{2b}})\big) &= \mathrm{E}\Big(\mathrm{E}\big(\mathbb{1}(j \in \hat{S}_\lambda(Z_{A_{2b-1}}))\,\mathbb{1}(j \in \hat{S}_\lambda(Z_{A_{2b}})) \;\big|\; A_{2b}, A_{2b-1}\big)\Big) \\ &= \mathrm{E}\big(\pi_j(\lambda)\pi_j(\lambda)\big) = \pi_j(\lambda)^2 \leq (q(\lambda)/p)^2. \end{aligned} \tag{4.7}$$

In turn, the $\max_{j \in S^c} \pi_j(\lambda) \leq q(\lambda)/p$ condition is implied by the exchangeability and not-worse-than-random-guessing conditions used by Meinshausen and Bühlmann (2010) and Shah and Samworth (2013) in the stability selection analogues of Theorem 4.1, detailed in Section 4.3. To be precise, Shah and Samworth (2013) do not require these conditions in their theory, but they are always assumed when implementing their versions of stability selection in practice. An empirical study and further details about Condition 1 for the practically relevant cases of $m \in \{1, 2, 3\}$ are provided in Section S4.3.

4.3 Comparison to stability selection

The analogue of Equation 4.2 for stability selection (Equation 2.2) under the exchangeability and not-worse-than-random-guessing conditions of Meinshausen and Bühlmann (2010) is

$$\mathrm{E}(\mathrm{FP}) \leq \frac{q^2}{(2\tau - 1)p}, \tag{4.8}$$

where $q = \mathrm{E}\big|\bigcup_{\lambda \in \Lambda_{\mathrm{MB}}} \hat{S}_\lambda(Z_{1:\lfloor n/2 \rfloor})\big|$. Under the additional assumptions that (a) $q(\lambda)^2/p \leq 1/\sqrt{3}$ for all $\lambda \in \Lambda$ and (b) the distributions of the simultaneous selection probabilities (defined in

Section S2) are unimodal, Shah and Samworth (2013) establish the stronger bound

$$\mathrm{E(FP)} \leq \frac{C(\tau, B)\, q(\lambda)^2}{p} \tag{4.9}$$

for stability selection where, for $\tau \in \left\{\frac{1}{2} + 1/B,\ \frac{1}{2} + 3/(2B),\ \frac{1}{2} + 2/B, \ldots, 1\right\}$,

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

There are several reasons the IPSS bounds in Theorem 4.2 are significantly tighter than the Meinshausen and Bühlmann (MB) and unimodal (UM) bounds in Equations 4.8 and 4.9. First, the IPSS bounds hold for all $\tau \in (0, 1]$, whereas the MB and UM bounds are restricted to $\tau \in (0.5, 1]$ since they go to ∞ as $\tau \to 0.5$. Second, all of the terms in the integrands of Equations 4.4 and 4.5 are typically orders of magnitude smaller than q^2/p in both Equations 4.8 and 4.9. Indeed, since $q(\lambda)$ is much smaller than p over a wide range of λ values in sparse or even moderately sparse settings, $q(\lambda)^4/p^3$ and $q(\lambda)^6/p^5$ are typically much smaller than $q(\lambda)^2/p$, and this difference becomes more pronounced as p grows. Additionally, the lower order terms in Equations 4.4 and 4.5 are $O(1/B)$ or $O(1/B^2)$, so with our typical choice of $B = 50$, the contribution of these terms is reduced even further, tending to 0 as $B \to \infty$ (Equation 4.6). By contrast, the MB bound has no B dependence, and the UM bound depends only weakly on B.

Shah and Samworth (2013) also derive an upper bound on E(FP) for stability selection based on assumptions of r-concavity. While tighter than the MB and UM bounds, this bound does not have a closed form and must be approximated with another algorithm, making the r-concave bound difficult to compare to Equations 4.2, 4.8, and 4.9 analytically. However,

the empirical results shown in Figure 3 indicate that the bounds in Theorem 4.2 for IPSS with h_2 and h_3 are tighter than all of the stability selection bounds, especially for target E(FP) values less than 5, which is the most practically relevant range.

5 Simulations

We present empirical results from linear and logistic regression simulations for a variety of feature distributions. The performance of IPSS is compared to the stability selection methods of Meinshausen and Bühlmann (2010) and Shah and Samworth (2013), as well as lasso.

Setup. Data are simulated from a linear regression model with normal residuals:

$$Y_i = \beta^\intercal X_i + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2),$$

from a linear model with residuals from a Student's t distribution with 2 degrees of freedom:

$$Y_i = \beta^\intercal X_i + \epsilon_i, \quad \epsilon_i \sim t(2),$$

and from a binary logistic regression model:

$$Y_i \sim \text{Bernoulli}(p_i), \quad p_i = \frac{\exp(\gamma \beta^\intercal X_i)}{1 + \exp(\gamma \beta^\intercal X_i)},$$

for $i \in \{1, \ldots, n\}$. For each simulated data set, the coefficient vector β has s nonzero entries $\beta_j \sim \text{Uniform}(-1, 1)$ located at randomly chosen coordinates $j \in \{1, \ldots, p\}$, and the remaining $p - s$ entries are set to 0. We simulate features from the following designs, which are similar to some of those used in Meinshausen and Bühlmann (2010). In all cases, X_i is independent of ϵ_i, and $(X_1, Y_1), \ldots, (X_n, Y_n)$ are independent and identically distributed.

- *Independent*: $X_i \sim \mathcal{N}(0, I_p)$ for all i, where I_p is the $p \times p$ identity matrix.

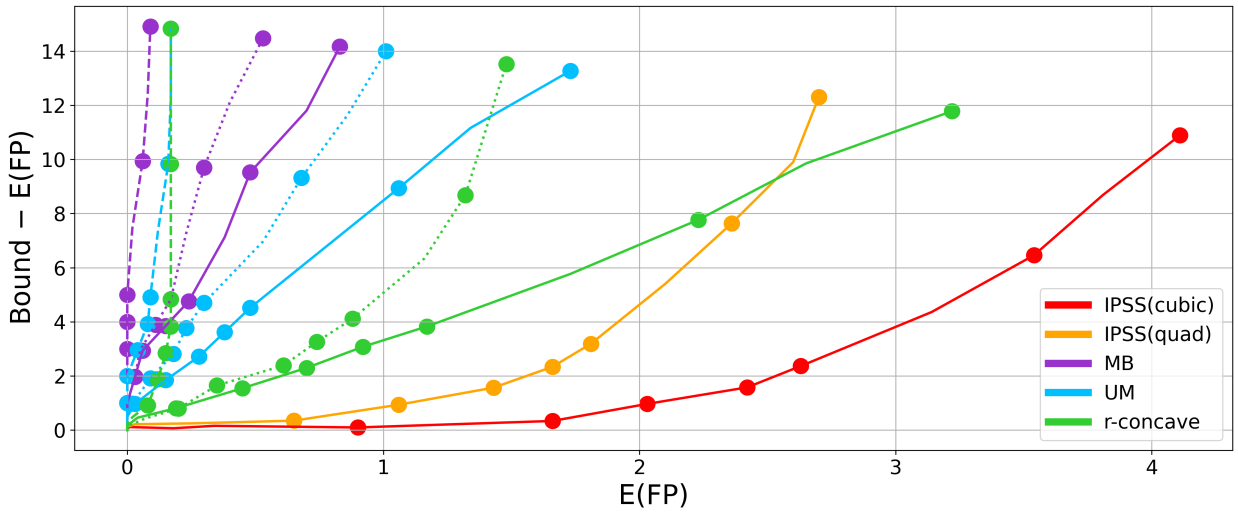

Figure 3: *Tightness of the E(FP) bounds.* Actual E(FP), averaged over 100 data sets simulated from an independent linear regression model with normal residuals and $p = 200$ features, as described in Section 5, versus the difference between the theoretical E(FP) bound for each method and its actual E(FP). Since all methods replace the inequalities in their E(FP) bounds with equalities, a perfectly calibrated method would generate a horizontal line at 0, that is, E(FP) = Bound. Dots show the results for each method when E(FP)_* equals, from left to right, 1, 2, 3, 4, 5, 10, and 15. IPSS(quad) and especially IPSS(cubic) are much closer to their theoretical bounds, particularly when $\text{E(FP)}_* \leq 5$. Solid, dotted, and dashed lines for the stability selection methods correspond to $\tau = 0.6$, 0.75, and 0.9, respectively.

- *Toeplitz:* $X_i \sim \mathcal{N}(0, \Sigma)$ where $\Sigma_{jk} = \rho^{|j-k|}$. We consider two cases: $\rho = 0.5$ and $\rho = 0.9$.

- *Factor model:* $X_i = \sum_{k=1}^{K} f_{ik}\varphi_k + \eta_i$ where $f_{ik}, \varphi_k, \eta_i \sim \mathcal{N}(0, 1)$ independently.

For each $j \in \{1, \ldots, p\}$, we standardize $(X_{1j}, \ldots, X_{nj})$ to have sample mean 0 and sample variance 1 before applying $\hat{S}_\lambda$. In the case of linear regression, the responses $Y_1, \ldots, Y_n$ are centered to have sample mean 0. For linear regression with normal residuals, σ^2 is chosen to satisfy a specified signal-to-noise ratio (SNR), defined as $\text{SNR} = \text{Var}(\beta^\intercal X_i)/\sigma^2$ and empirically estimated by $\text{SNR} \approx \sum_{i=1}^{n}(\beta^\intercal X_i)^2/(n\sigma^2)$ where β is generated as above and the X_i are drawn according to the specified designs and standardized as above. For logistic regression, $\gamma > 0$ determines the strength of the signal.

The 3 models (linear regression with normal and Student's t residuals, and logistic regression), 4 feature designs (independent, Toeplitz with $\rho = 0.5$ and 0.9, and factor model), and 2 feature dimensions ($p = 200$ and 1000) yield a total of 24 experiments. Each experiment consists of 100 trials, where 1 trial consists of generating data as above with n, s, and signal strength chosen according to Table 1, estimating the selection probabilities via Algorithm 1 with $B = 50$ subsamples, and choosing features according to each criterion. For linear regression, the baseline feature selection algorithm $\hat{S}_\lambda$ is lasso. For logistic regression, $\hat{S}_\lambda$ is L^1-regularized logistic regression (Lee et al., 2006). The factor model experiments use $K = 2$ factors when $p = 200$ and $K = 5$ factors when $p = 1000$.

Results. We quantify performance in terms of true positives (TP) and false positives (FP), where a selected feature is a *true positive* if its corresponding β entry is nonzero, and is a *false positive* otherwise. The dashed black line in the FP plots in Figures 4 and 5 shows the target value of E(FP). A tight bound on E(FP) should lead to curves lying close to this line,

p	n	s	SNR (regression)	γ (classification)
200	Uniform$\{50, \ldots, 200\}$	Uniform$\{10, \ldots, 20\}$	Uniform$(1/3, 3)$	Uniform$(1/2, 2)$
1000	Uniform$\{100, \ldots, 500\}$	Uniform$\{20, \ldots, 40\}$	Uniform$(1/3, 3)$	Uniform$(1/2, 2)$

Table 1: *Simulation parameters.* The number of samples n, number of true features s, and the signal strength parameters, SNR and γ, are randomly selected prior to each trial according to the above distributions, ensuring that our experiments cover a wide range of settings.

since all of the methods involve replacing the inequalities in their respective E(FP) bounds with equalities in order to calibrate the parameters. For example, if the target E(FP) is 2, a perfectly calibrated algorithm would produce an average FP of 2.

Figures 4 and 5 show results for linear regression with normal residuals when $p = 200$ and 1000, respectively. Similar plots for the other experiments are in Section S4 of the Supplement (Figures S1 to S4). In nearly every experiment, IPSS(quad) and IPSS(cubic) have average FPs that are closer to the target E(FP) than MB, UM, or r-concave with any $\tau \in \{0.6, 0.75, 0.9\}$. This is due to the relative strength of the IPSS bounds, as described in Section 4.3. Moreover, the average TP is higher for both IPSS methods than for the stability selection methods, often substantially so. Lasso with cross-validation (LassoCV) tends to have a relatively high TP at the expense of an exceedingly high FP.

These and the many additional results in Section S4.1 indicate that IPSS provides a better balance between true and false positives in a wide range of settings than both stability selection and LassoCV. That is, while stability selection has low FP at the expense of low TP, and LassoCV has high TP at the expense of high FP, IPSS gets closer to the user-specified E(FP), while achieving moderate-to-high TP that can even approach the TP of LassoCV.

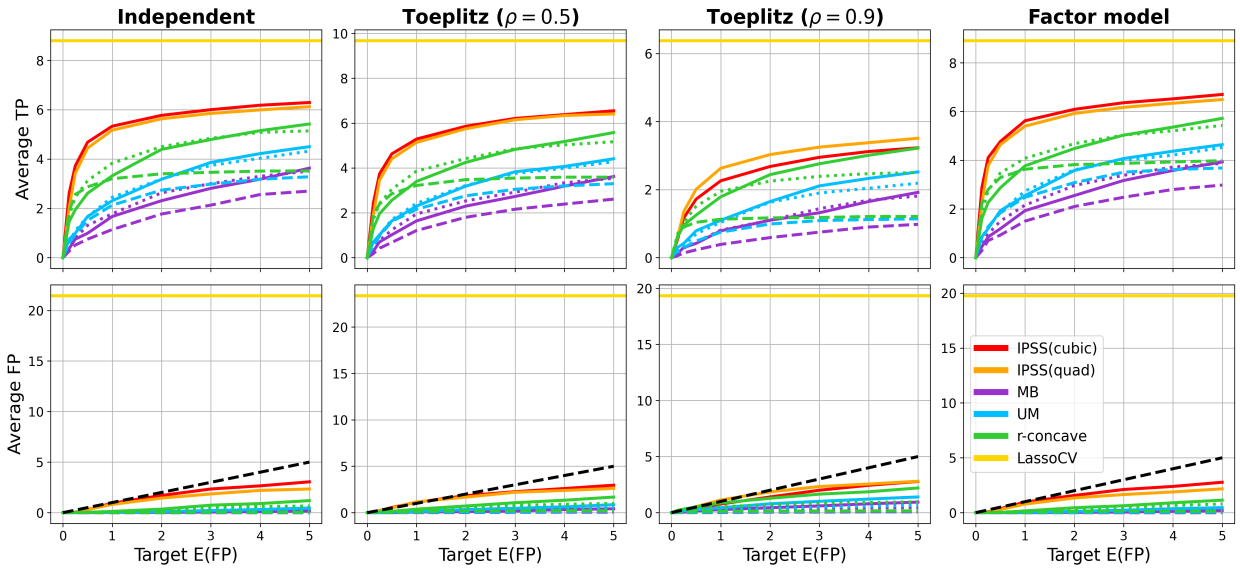

Figure 4: *Linear regression with normal residuals* ($p = 200$). The solid, dotted, and dashed lines for the stability selection methods correspond to $\tau = 0.6$, 0.75, and 0.9, respectively.

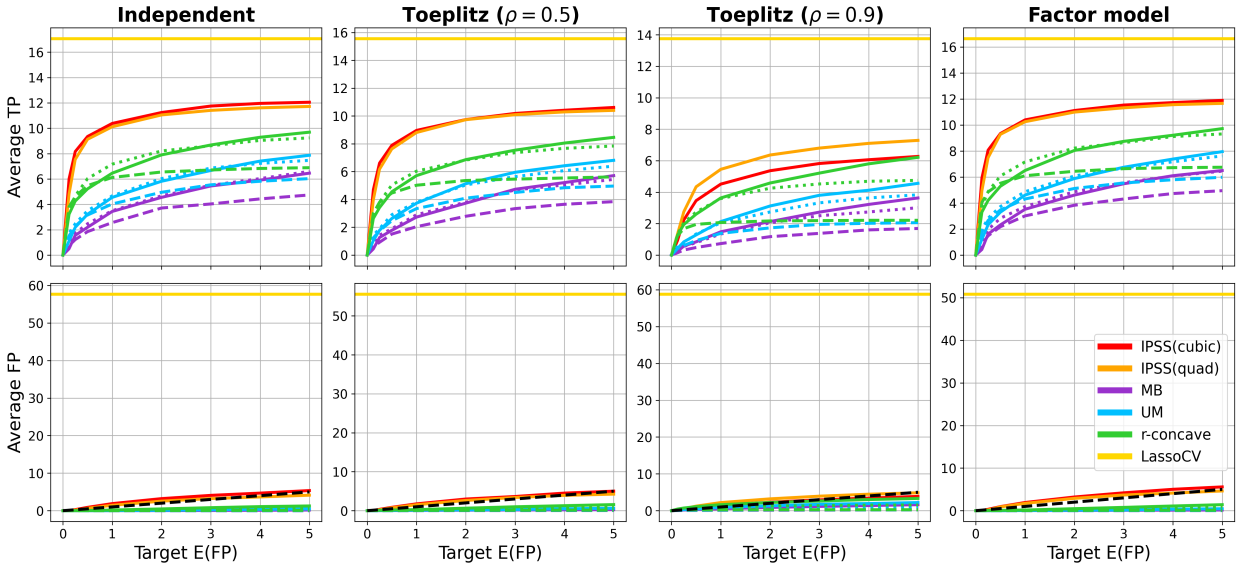

Figure 5: *Linear regression with normal residuals* ($p = 1000$). The solid, dotted, and dashed lines for the stability selection methods correspond to $\tau = 0.6$, 0.75, and 0.9, respectively.

6 Applications

6.1 Prostate cancer

We applied IPSS and the stability selection methods to reverse-phase protein array (RPPA) data consisting of expression levels of $p = 125$ proteins in $n = 351$ prostate cancer patients (Vasaikar et al., 2018). The response is *tumor purity*—the proportion of cancerous cells in a tissue sample—and the goal is to identify the genes that are most related to it. Since tumor purity takes values in $[0, 1]$, we use lasso as our baseline selection algorithm. The target E(FP) is 1 for all methods, and for MB, UM, and r-concave we set $\tau = 0.75$. Figure 6 shows the results. MB, UM, r-concave, IPSS(quad), and IPSS(cubic) select 4, 4, 5, 8, and 10 proteins, respectively. Although it is difficult to know which features should be selected on real data such as this, a literature search presented in Section S5 indicates that all 10 proteins identified by IPSS play a nontrivial role in prostate cancer. As in Figure 2, the stability selection methods miss important information about the stability paths that IPSS successfully captures. For example, the selection probabilities for PKC, BAK1, and PTEN are essentially 0 for many λ values before abruptly rising above many of the other paths.

6.2 Colon cancer

We applied IPSS and the stability selection methods to the expression levels of $p = 1908$ genes in $n = 62$ tissue samples, 40 cancerous and 22 normal (Alon et al., 1999). The goal is to identify genes whose expression levels differ between the cancerous and normal samples. Since the response is binary, we use L^1-regularized logistic regression as the underlying selection algorithm. We use a target E(FP) of 1/2 for all five methods, and the threshold for MB, UM,

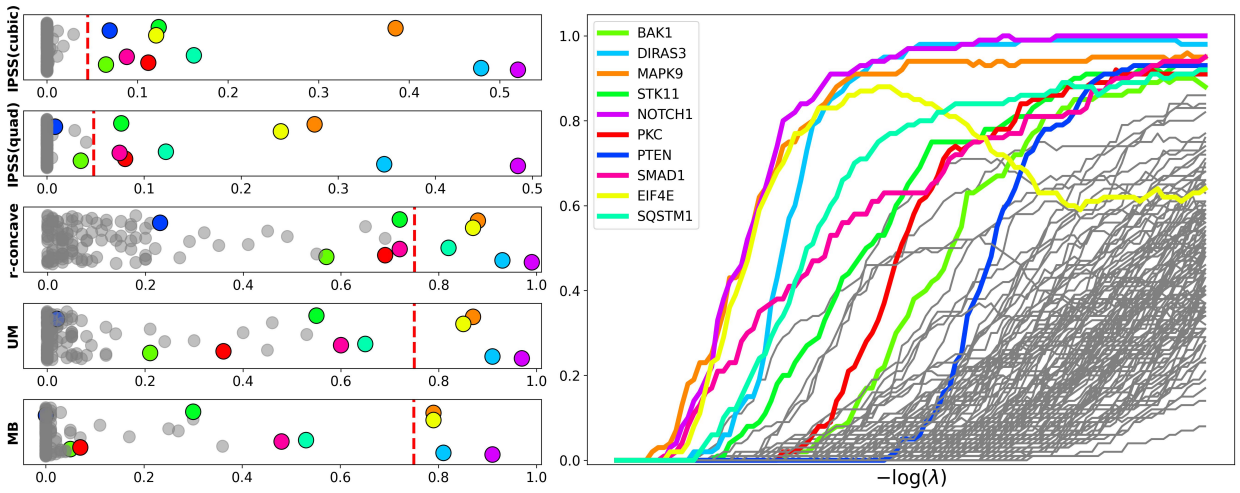

Figure 6: *Prostate cancer results.* (*Left*) Feature scores and thresholds (vertical red lines) separating selected and unselected genes for each method. Scores are one-dimensional and correspond to the horizontal axes. Every method's set of selected proteins is a subset of the proteins selected by IPSS(cubic), shown in color in all plots; the remaining proteins are in gray. (*Right*) Estimated stability paths for each protein. The horizontal axis is on a log scale.

and r-concave is $\tau = 0.75$. Expression levels are log-transformed and standardized as done by Shah and Samworth (2013), who also study these data in the context of stability selection. Figures 7 and 8 show the results. MB, UM, r-concave, IPSS(quad), and IPSS(cubic) select 1, 2, 7, 11, and 16 genes, respectively. In addition to Figure 8, a literature search reported in Section S5 supports the claim that all of the genes identified by IPSS are related to colon cancer.

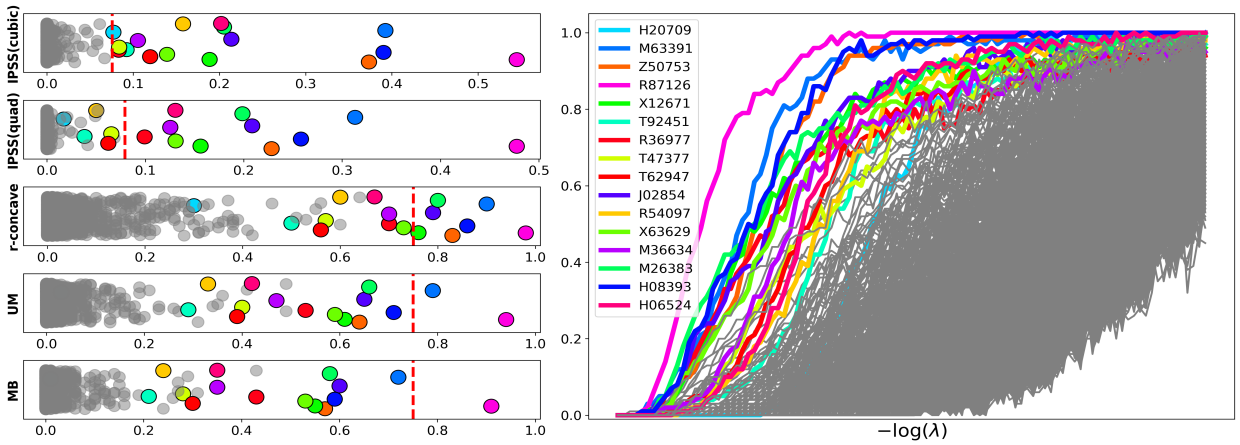

Figure 7: *Colon cancer results.* (*Left*) Feature scores and thresholds (vertical red lines) separating selected and unselected genes for each method. Every method's set of selected genes is a subset of the genes selected by IPSS(cubic), shown in color in all plots; the remaining genes are in gray. (*Right*) Estimated stability paths for each gene.

7 Discussion

IPSS has several attractive properties. It has stronger theoretical guarantees than stability selection, and yields significantly better performance on a wide range of simulated and real data. It has the same computational cost as stability selection and is easier to tune, requiring only the target E(FP) to be specified. In this work, we focused on the functions $f = h_2$ and $f = h_3$ due to their favorable theoretical properties and empirical performance. However, it is possible that other functions will lead to better results; investigating this point is an interesting line of future work.

Another interesting direction is to apply IPSS to other feature selection algorithms, such as graphical lasso, elastic net, and adaptive lasso (Friedman et al., 2008; Zou and Hastie, 2005; Zou, 2006). In the case of elastic net, there are two regularization parameters, and while

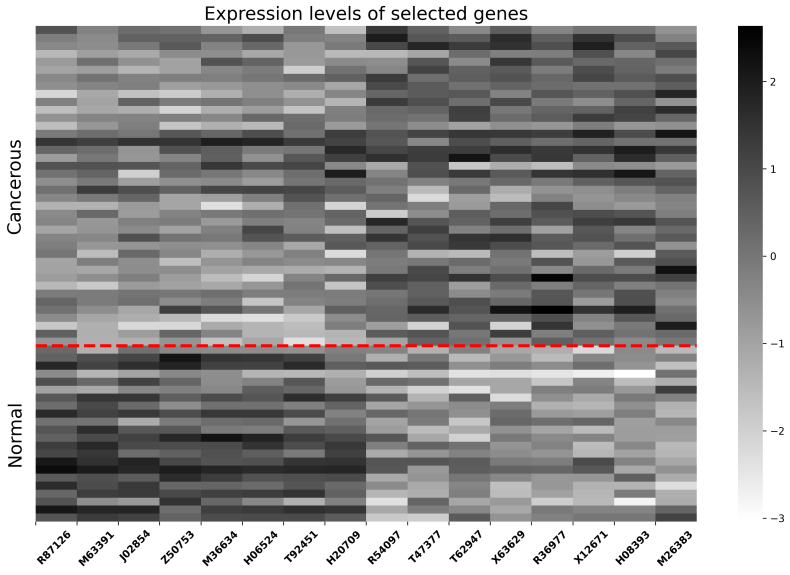

Figure 8: *Expression level heatmap for the 16 genes selected by IPSS(cubic).* Each of the 62 rows corresponds to one tissue sample. The first 40 rows are cancerous, and the latter 22 are normal; the dashed red line separates the two classes. Each column corresponds to a gene selected by IPSS(cubic), or equivalently, the union of genes selected by each method since each gene selected by the other methods was also selected by IPSS(cubic). For each gene, there is a clear distinction between expression levels for cancerous versus normal samples.

our methodology and theory appears to carry over to this setting (now with $\Lambda \subseteq (0, \infty)^2$ and $(\lambda_1, \lambda_2) \mapsto \hat{S}_{\lambda_1, \lambda_2}$), it remains to work out the details and investigate the performance of IPSS in the context of multiple regularization parameters. Finally, we noted in Section 3 that stability selection is often used in conjunction with other statistical methods. Given that IPSS yields better results than stability selection at no additional cost and with less tuning, it would be interesting to study these joint methods with IPSS in place of stability selection.

Data availability and code

The prostate cancer data set used in this work is freely available at `https://www.linkedomics.org/data_download/TCGA-PRAD`. The colon cancer data set is freely available at `http://genomics-pubs.princeton.edu/oncology/affydata/index.html`. Both data sets, along with code from this work, are also freely available at `https://github.com/omelikechi/ipss`.

Acknowledgments

O.M. would like to thank David Dunson and Steven Winter for initial discussions that took place under funding from Merck & Co. and the National Institutes of Health (NIH) grant R01ES035625. J.W.M. was supported in part by the National Institutes of Health (NIH) grant R01CA240299.

References

David H Alexander and Kenneth Lange. Stability selection for genome-wide association. *Genetic Epidemiology*, 35(7):722–728, 2011.

Uri Alon, Naama Barkai, Daniel A Notterman, Kurt Gish, Suzanne Ybarra, Daniel Mack, and Arnold J Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences*, 96(12):6745–6750, 1999.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

Anne-Claire Haury, Fantine Mordelet, Paola Vera-Licona, and Jean-Philippe Vert. Tigress: trustful inference of gene regulation using stability selection. *BMC Systems Biology*, 6(1): 1–17, 2012.

Benjamin Hofner, Luigi Boccuto, and Markus Göker. Controlling false discoveries in high-dimensional situations: boosting with stability selection. *BMC Bioinformatics*, 16:1–17, 2015.

Su-In Lee, Honglak Lee, Pieter Abbeel, and Andrew Y Ng. Efficient l_1-regularized logistic regression. In *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI)*, volume 6, pages 401–408, 2006.

Chenlei Leng, Yi Lin, and Grace Wahba. A note on the lasso and related procedures in model selection. *Statistica Sinica*, pages 1273–1284, 2006.

Jun-Li Li and Chun-Xia Zhang. Ensembling variable selectors by stability selection for the cox model. In *2017 International Conference on Machine Learning and Cybernetics (ICMLC)*, volume 1, pages 35–41. IEEE, 2017.

Shuang Li, Li Hsu, Jie Peng, and Pei Wang. Bootstrap inference for network construction with an application to a breast cancer microarray study. *The Annals of Applied Statistics*, 7(1):391, 2013.

Suryanarayana Maddu, Bevan L Cheeseman, Ivo F Sbalzarini, and Christian L Müller.

Stability selection enables robust learning of differential equations from limited noisy data. *Proceedings of the Royal Society A*, 478(2262):20210916, 2022.

Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 72(4):417–473, 2010.

Rajen D Shah and Richard J Samworth. Variable selection with error control: another look at stability selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 75(1):55–80, 2013.

John D Storey and Robert Tibshirani. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16):9440–9445, 2003.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.

Suhas V Vasaikar, Peter Straub, Jing Wang, and Bing Zhang. Linkedomics: analyzing multi-omics data within and across 32 cancer types. *Nucleic Acids Research*, 46(D1): D956–D963, 2018.

Fan Wang, Sach Mukherjee, Sylvia Richardson, and Steven M Hill. High-dimensional regression in practice: an empirical study of finite-sample prediction, variable selection and ranking. *Statistics and Computing*, 30:697–719, 2020.

Tino Werner. Loss-guided stability selection. *Advances in Data Analysis and Classification*, pages 1–26, 2023.

Jiayu Zhou, Jimeng Sun, Yashu Liu, Jianying Hu, and Jieping Ye. Patient risk prediction model via top-k stability selection. In *Proceedings of the 2013 SIAM International Conference on Data Mining*, pages 55–63. SIAM, 2013.

Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320, 2005.