As with before, we are grateful to the reviewers for the careful reading and very helpful editing and suggestions. We have fixed all the typos found by the reviewers. We also have improved the paper in other minor aspects in response to the review – these changes are described below.

Review 4

1. *On page 30, I did not understand : we set $K = 7$ for all the experiments, and then We set $s = 5$ and the true variables are X_j for $j = 5, 6, 7, 8, 9, 10$.*

   Here, the true function is a soft-max of linear forms $\beta_k^T x$, i.e., $f_0(x) = \log\left(\sum_{k=1}^K \exp(\beta_k^T x)\right) - \mu$ where the β_k's are randomly generated unit vectors. K is the number of linear pieces β_k's and s is the number of relevant variables. We have clarified this a bit more in section 6 of the paper.

Reviewer 5

1. *In Section 2.3, the authors preview their main results concerning variable selection consistency in a finite-sample setting. However, at this point, the true regression function f_0 has not yet been defined (it is defined later in Section 5.2). It would be helpful if the authors were to state the model generating the data in this section.*

   We have stated the model in the beginning of Section 2.3.

2. *In Example 3.3, the authors describe a method for approximating any bounded density over a hypercube arbitrarily well using boundary flat densities. The second paragraph mentions truncating the density to the hypercube $[\epsilon, 1\epsilon]^p$since the truncated function isnt a density, it should technically also be rescaled.*

   This has been fixed.

3. *In Assumption A4 at the bottom of page 25, the moment generating function should read $\mathbb{E}e^{tW}$ rather than $\mathbb{E}e^{t\epsilon}$. There is some change in notation in Section 6.1, where W is replaced by ϵ; perhaps it would be better to keep the notation consistent.*

   We have changed ϵ to W across the paper to keep the notations consistent.

4. *The last sentence of the first paragraph on page 26 is a bit hard to follow. What does if X is independent mean (does this mean $p(x)$ is a product distribution)? I did not understand where the simplification came from.*

   "X is independent" does mean that $p(x)$ is a product distribution; we have clarified this in the paper. We have also added a new section to the supplement (Section 11, Proposition 11.1) that explains and derives the simplification.

5. *The $\nu = 0.9$ curve in Figure 5(d) is rather ill-behaved compared to the other curves. However, if my understanding is correct, the probability of successful screening should still tend to 1 in this case. Perhaps the authors could consider plotting more samples to show eventual consistency for this curve, as well (and also take an average over more trials, since the behavior of the curve is somewhat noisy).*

   The probability of successful screening does tend to 1. That was not reflected in the old plot because the old plot describes the probability both that the number of false positives is 0 (successful screening) and that the number of false positives is less than 20. It was because of this latter criterion that the $\nu = 0.9$ curve did not tend to 1 in the old plot. The threshold of 20 was arbitrarily chosen and we have revised the plots to remove it. The updated figures 5(d) and 5(e) show only the probability that the number of false negatives is 0; the number of false positives is reflected in the newly added figures 6(f) and 6(g), which show that total number of variables selected in the experiments that correspond to figures 5(d) and 5(e). We have also updated the discussion in section 6.1.3 and 6.1.4.

Sincerely,

Min Xu, Minhua Chen, and John Lafferty
December 1, 2015