Title: Variable selection and model building via likelihood basis pursuit
JASA Ms.# 02-205

Note to Referee 1

Thank you for the review and helpful comments. This letter is in response to the questions and concerns raised in your review report. The labels correspond to those in the report.

1. We agree that the interpretation of complicated interactions between continuous variables can be challenging. In the paper, we intended to say that adding categorical variables to a model that already has continuous variables can increase the complexity of the model. It is certainly not our intention to claim that "working with categorical covariates is more difficult than continuous ones". We believe the misunderstanding was caused by the sentence "The two-factor interaction model which incorporates categorical variables is much more complicated than in the continuous case" [in old Section 2.3, the first sentence of the fourth paragraph]. To avoid confusion, we modified it into "Adding two factor interactions with categorical variables to a model that already includes parametric and smooth terms adds a number of additional terms to the general model" [in new Section 2.3].

   In this paper we mainly consider bivariate interactions among the variables. There exist a few methods of exploring two-way interactions effectively, such as contour plots and cross section plots. Some examples can be found in Hastie & Tibshirani (1990) and Lin et al.(2000). For higher order interactions, it is more difficult to display the estimated surfaces. We added the following sentence "Two factor interactions arise in many practical problems. See Hastie and Tibshirani (1990), Section 9.5.5 or Lin et al. (2000) page 1570-1700, Figures 9 and 10, for interpretable plots of two factor interactions with continuous variates" [in new Section 2.2.2, the first sentence].

   We replaced the two-way interaction example [in old Section 6.2] with a new example [in new section 6.2]. The true logit function in the new example has an interaction term $\cos(2\pi(X_1 - X_2))$. This interaction is non-monotonic with respect to either X_1 or X_2, and thus more complicated than the product term $X_1 * X_2$ in the old example. We used the cross section plots [in new Figure 6] to illustrate how our approach has effectively captured the estimated surface.

2. In the proposed sequential Monte Carlo bootstrap method, the ordering of covariates is determined by the importance of individual covariates, which can be measured by a certain norm of the function estimates such as the L_1 norm. When the tuning parameters are chosen properly and the RKHS is rich enough, the LBP model provides

accurate estimates for all the components. Under that circumstance, the ordering based on the importance measure of the estimated components should reflect the ordering of the importance of the covariates if true functions were known. We note that different importance measures could give different orderings of the variables. However, both L_1 and L_2 importance criteria gave excellent results in selecting the correct terms in our simulations, and the unimportant terms would have zero or tiny measures of importance in either L_1 or L_2 sense.

We replaced the sentence "An alternative measure based on the functional L_2 norm worked equally well in our simulation studies" [in old Section 4.1, the last sentence] with "The functional L_2 norm gave almost identical results in numerous simulation studies, (not reproduced here)." After the first two sentences in Section 4.2, we added "We will use a sequential procedure for selecting important terms".

3. The following corrections are made corresponding to the suggestions.

   1). "RKHS" was spelled out as "reproducing kernel Hilbert space" on page 4.

   2). The word "subject to" was changed to "where" [In old equations (2.13) and (2.14), which have become new equations (2.9) and (2.10)].

   3). The KL distance is always positive without taking absolute value. The equation number of old equation (3.2) was omitted for shortening the paper, and correspondingly we adjusted the other equation numbers.

   4). Cross validation is a common way for selecting tuning parameters. One argument that shows that CV is roughly an unbiased estimate of the CKL in density estimation can be found in Silverman (1986) page 53. The following is the argument adapted to the exponential family regression situation. For any tuning parameter λ, let f_λ be the corresponding estimate, and $f_\lambda^{[-i]}$ be the estimate with the i-th data point omitted. Denote the density function of the input vector x by $d(x)$. Then we have

   $$CKL(\lambda) = \int [-\mu(x) f_\lambda(x) + b(f_\lambda(x))] d(x).$$

   Let us consider the following CV score:

   $$CV(\lambda) = \frac{1}{n} \sum [-y_i f_\lambda^{[-i]}(x_i) + b(f_\lambda^{[-i]}(x_i))].$$

   Since (x_i, y_i), $i = 1, ..., n$, are i.i.d pairs, and since $f_\lambda^{[-i]}$ only depends on observations other than (x_i, y_i), we have

$$\begin{aligned}
E(CV) &= E[-y_n f_\lambda^{[-n]}(x_n) + b(f_\lambda^{[-n]}(x_n))] \\
&= E\{E[-y_n f_\lambda^{[-n]}(x_n)|x_n]\} + E[b(f_\lambda^{[-n]}(x_n))] \\
&= E\{E[-y_n|x_n]E[f_\lambda^{[-n]}(x_n)|x_n]\} + E[b(f_\lambda^{[-n]}(x_n))] \\
&= E\{-\mu(x_n)E[f_\lambda^{[-n]}(x_n)|x_n]\} + E[b(f_\lambda^{[-n]}(x_n))] \\
&= E\{E[-\mu(x_n)f_\lambda^{[-n]}(x_n)|x_n]\} + E[b(f_\lambda^{[-n]}(x_n))] \\
&= E[-\mu(x_n)f_\lambda^{[-n]}(x_n) + b(f_\lambda^{[-n]}(x_n))] \\
&= E\{E[-\mu(x_n)f_\lambda^{[-n]}(x_n) + b(f_\lambda^{[-n]}(x_n))|(x_i, y_i)_{i=1,\cdots,n-1}]\} \\
&= E\int[-\mu(x)f_\lambda^{[-n]}(x) + b(f_\lambda^{[-n]}(x))]d(x) \\
&\approx E\int[-\mu(x)f_\lambda(x) + b(f_\lambda(x))]d(x) \\
&= E(CKL).
\end{aligned}$$

The approximation comes from that on average the estimate based on n observations should be close to the estimate based on $(n-1)$ observations when n is decently large.

Thus $CV(\lambda)$ is commonly expected to be at least roughly unbiased for $CKL(\lambda)$. In practice, Xiang & Wahba (1996) and Gu (2002) have found the slightly modified cross validation score $CV = \frac{1}{n}\sum[-y_i f^{[-i]}(x_i) + b(f(x_i))]$ works a little better. This version of CV is used in this paper. It was first introduced in Xiang & Wahba (1996) and has been shown to perform well in many studies since then. See, for example, Gu (2002) for an overview. We do not think it is necessary to give the above heuristic argument in the paper.

5). Adding the detailed derivation of $ACV(\lambda)$ on page 11 to the paper would lengthen it at a point where we are trying hard to shorten it. However we can provide a copy of the relevant parts in the PhD thesis of the first author to the derivation, with the assurance that it is permanently available. If the outline of proof is absolutely required, we can add it.

6). On page 11, the first paragraph in Section 3.2, we added "based on the following theorem, which has been exploited by numerous authors, see e.g. Girard (1998)."

7). In the second paragraph of Section 4.2, we added more details about bootstrap sampling as "Conditional on the original covariates $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, we generate $\{y_1^{*(\eta)}, \ldots, y_n^{*(\eta)}\}$ (responses 0 or 1) using the logit function $f = \hat{b}_0 + \hat{f}_{(1)} + \cdots + \hat{f}_{(\eta)}$. In total we sample T independent sets of data $(\mathbf{x}_1, y_{1,t}^{*(\eta)}), \ldots, (\mathbf{x}_n, y_{n,t}^{*(\eta)}), t =$

$1, \ldots, T$, from the null model f, fit the main effects model for each set, and compute $\hat{L}_t^{*(\eta+1)}, t = 1, \ldots, T$."

8). In step 1 [on old page 14, which has become new page 13], q is defined as "any number slightly larger than $\hat{L}_{(1)}$", rather than being equal to $\hat{L}_{(1)}$.

9). There are many numerical methods for solving linear and quadratic programming problems. The challenge in this paper is to minimize a non-differentiable objective function with a general likelihood. According to the referee's comment, we changed the second half of the first sentence in Section 5 into "some numerical methods for optimization fail to solve this kind of problem."

10). Please see our response to question 1 and the new example in Section 6.2.

11). Please see our response to question 1 and the new example in Section 6.2.

12). In response to the request to shorten the manuscript, we removed Figures 8 and 10 in the old version and made the corresponding changes.

Reference

1. Girard, D. (1998), 'Asymptotic comparison of (partial) cross-validation, GCV and randomized GCV in nonparametric regression', *The Annals of Statistics* **26**, 315-334.

2. C. Gu. (2002), *Smoothing spline ANOVA models.* Springer-Verlag.

3. Hastie, T. J. & Tibshirani, R. J. (1990), *Generalized Additive Models.* Chapman & Hall.

4. X. Lin, G. Wahba, D. Xiang, F. Gao, R. Klein & B. Klein. (2000), 'Smoothing spline ANOVA models for large data sets with Bernoulli observations and the randomized GACV', *The Annals of Statistics* **28**, 1570-1600.

5. Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis.* Chapman and Hall.