Journal of Multivariate Analysis 150 (2016) 14–26

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Lack of fit tests for linear regression models with many predictor variables using minimal weighted maximal matchings

Forrest R. Miller [a], James W. Neill [b,*]

[a] Department of Mathematics, Cardwell Hall, Kansas State University, USA
[b] Department of Statistics, Dickens Hall, Kansas State University, USA

ARTICLE INFO

Article history:
Received 25 August 2015
Available online 19 May 2016

AMS subject classifications:
62J05
62F03

Keywords:
Linear regression
Lack of fit
Many predictors
Matchings

ABSTRACT

We develop lack of fit tests for linear regression models with many predictor variables. General alternatives for model comparison are constructed using minimal weighted maximal matchings consistent with graphs on the predictor vectors. The weighted graphs we employ have edges based on model-driven distance thresholds in predictor space, thereby making our testing procedure implementable and computationally efficient in higher dimensional settings. In addition, it is shown that the testing procedure adapts to efficacious maximal matchings. An asymptotic analysis, along with simulation results, demonstrate that our tests are effective against a broad class of lack of fit.

1. Introduction

We consider the problem of testing the adequacy of a linear regression model

$$\mathbf{y} = [\mathbf{1}_n, X]\beta + \mathbf{e}$$

where the rows of X consist of the predictor vectors $x_i = (x_{i1}, \ldots, x_{ip}) \in \mathbb{R}^p$, $1 \le i \le n$, which are possibly higher dimensional. In addition, $\mathbf{e}^\top = (e_1, \ldots, e_n)$ is a Gaussian distributed random vector with independent components having $\mathrm{E}(e_i) = 0$ and $\mathrm{E}(e_i^2) = \sigma^2$, and $\mathbf{1}_n \in \mathbb{R}^n$ denotes the vector consisting entirely of ones to accommodate an intercept in the model. In the following we let $W = [\mathbf{1}_n, X]$. Since parametric regression models with many predictors are frequently used in the natural and social sciences, it is important to first check the adequacy of a proposed model to avoid misleading inferences. In this paper we present regression lack of fit tests for the case of many predictors which are effective against a large class of lack of fit and are computationally efficient.

Our testing procedure involves the development of a supremum-type multiple test based on a collection of Fisher statistics, each constructed from a matching on the predictor vectors x_i. General alternatives for defining the Fisher statistics use minimal weighted maximal matchings consistent with graphs on the x_i. These graphs have edges weighted according to model-driven distance thresholds in predictor space. The matchings are constructed edge by edge, at each stage choosing the edge with smallest possible weight. It is the use of matchings which makes feasible the construction and implementation of our lack of fit tests when p is large, although as seen in Section 2 it is necessary to have $n > 2p + 2$.

* Corresponding author.
E-mail address: jwneill@k-state.edu (J.W. Neill).

http://dx.doi.org/10.1016/j.jmva.2016.05.005
0047-259X/

In previous work, Miller et al. [20,21] developed a graph theoretic representation of near replicate clusterings of statistical units to obtain lack of fit tests for linear regression models. This work helps provide a framework for generalizing the classical test presented by Fisher [11], an approach for testing linear regression lack of fit investigated by several authors as referenced by Miller et al. [20,21]. Although presented in generality, implementation of our previous tests was focused on models with lower dimensional predictor vectors. In particular, a graph on the predictor vectors was used to determine a special collection of clusterings (the atoms consistent with the graph), and then an optimization procedure (a maximin method or restricted least squares approach) was applied to choose an optimal clustering to test $E(\mathbf{y}) \in W$ (by which we mean $E(\mathbf{y})$ is an element of the column space of W). In the current work, we use a very different special collection of clusterings consistent with the graph. These are the clusterings that group at most two vertices together, provided these two vertices form an edge of the graph, which is why we call them edge clusterings. They are called matchings in the field of combinatorial optimization. Edge clusterings possess special advantages for testing regression lack of fit, some of which were discussed in more recent work by the authors (Miller and Neill [19]). However, they also allow efficient implementation in higher dimensional models with many predictor variables, which is the emphasis of the current work.

This paper is organized as follows. In Section 2 we first determine weighted graphs with vertices given by the $\{x_i\}$ and edges based on distance thresholds. Matchings on such graphs are then used to determine subspaces of the lack of fit space, and subsequently to construct lack of fit tests. The asymptotic behavior of such tests for a broad class of underlying true data generators is given, as well as the large sample behavior of matching sequences on a hypercube in $\mathbb{R}^p$. In addition, minimal weighted maximal matchings useful for detecting misspecification associated with the model $E(\mathbf{y}) \in W$ are defined in Section 2. A computationally efficient algorithm to determine such matchings is presented in Section 3, along with a multiple testing procedure which follows Baraud et al. [2]. Given the unknown nature of any underlying lack of fit, this procedure is implemented with a model-driven set of matchings to enhance detection of model inadequacy associated with the specified model $E(\mathbf{y}) \in W$. Section 4 provides the results of a simulation study for the cases $p = 10$ and $p = 20$. These results demonstrate efficient implementation of our testing procedure to effectively detect general lack of fit in linear regression models with many predictors. Proofs of theorems are given in the Appendix.

An extensive literature exists which addresses the problem of testing the adequacy of a specified parametric regression model. This work includes not only generalizations of Fisher's test as mentioned previously, but also the use of nonparametric regression methods to test the fit of a parametric model. Hart [14] provides a thorough review of the development and use of smoothing methodology to construct such lack of fit tests, with a focus on the $p = 1$ case. More recently, see also Eubank et al. [8]. Nonparametric regression techniques have also been employed for the multivariate predictor case (e.g. Staniswalis and Severini [23], Härdle and Mammen [13], Zheng [28], Dette [7], Fan et al. [10], Koul and Ni [17], Guerre and Lavergne [12], Song and Du [22]). However, even for smaller values of $p > 1$, the use of smoothing methods is problematic due to the curse of dimensionality. In many of these references, (y_i, x_i), $1 \leq i \leq n$, are considered to be independent and identically distributed $\mathbb{R}^{p+1}$ random vectors from a population, and interest involves testing $E(y|X = x) = m(x)$ where $m(x)$ is a specified parametric regression function. In this setting, Lavergne and Patilea [18] presented a test for $m(x)$ with many regressors involving estimation of conditional expectations given a linear index for a class of single-index models. Combining the expectations as a single numerically estimated integral provides a test against nonparametric alternatives, which reduces the dimension of the problem yet preserves consistency.

For another approach, Khmaladze and Koul [15] presented regression model adequacy tests based on innovation martingale transforms. These tests are asymptotically distribution free for fitting a parametric model to the regression function i.e. the asymptotic null distribution is free of the specified parametric model and the error distribution but depends on the design distribution when $p > 1$. Christensen and Lin [5] also presented tests based on partial sum processes of the residuals and determined the asymptotic null distributions of the maximized partial sums in order to check for lack of fit. This work involves modifications of the test proposed by Su and Wei [27], whose test is based on a partial ordering of the residuals and the asymptotic null distribution is approximated by simulation. As the power of tests based on partial sums of residuals can be greatly influenced by the ordering chosen, Christensen and Lin [5] suggested a total ordering of the data based on a modified Mahalanobis distance and empirically demonstrated their tests were effective for certain types of model inadequacy involving multivariate predictors. Following work by Stute [24] and Stute et al. [26], Stute et al. [25] also presented regression model adequacy tests based on empirical processes of the regressors marked by the residuals, and used a wild bootstrap approximation for the process distribution.

In addition to the preceding work, Aerts et al. [1] and Fan and Huang [9] considered lack of fit tests for multiple regression where the $\{x_i\}$ are considered to be fixed. To circumvent the curse of dimensionality, these authors placed restrictions on the alternative models. For example, Aerts et al. developed tests based on functions of score statistics but require specification of a path in the additive alternative models space, which quickly becomes complex with increasing predictor dimensionality. Fan and Huang constructed tests based on the adaptive Neyman test for the multivariate case but the ability of these tests to detect various types of model inadequacy depends on the ordering of residuals, which can be challenging in higher dimensions. Christensen and Sun [6] followed Fan and Huang with Fourier transforms in the multivariate linear model context, and suggested modifications to the normalizing constants for improved small sample size while maintaining the same asymptotic distributions as Fan and Huang. As noted by Christensen and Sun, their tests also depend on the ordering of the observations to ensure that large Fourier coefficients are concentrated on the lower frequencies. As in these works, we construct lack of fit tests based on fixed predictors in $\mathbb{R}^p$. Unlike these works, we consider the case of moderate to large

numbers of predictor variables and show our testing procedure is effective and implementable for such values of p. This is demonstrated in our simulations for the cases $p = 10$ and $p = 20$.

Some notation that is used in the paper is now introduced. In general, if $\mathcal{V}$ denotes a real inner product space, we let $\langle u, v\rangle$ represent the inner product for $u, v \in \mathcal{V}$ and let $\dim \mathcal{V}$ represent the dimension of $\mathcal{V}$. If $\mathcal{U} \subset \mathcal{V}$ is a subspace then $\mathcal{U}^{\perp}$ is the orthogonal complement of $\mathcal{U}$ and $P_{\mathcal{U}}$ denotes the orthogonal projection operator of $\mathcal{V}$ onto $\mathcal{U}$. Also, $\angle(u, \mathcal{U})$ denotes the angle in $[0, \pi/2]$ between a vector u and a subspace $\mathcal{U}$. In particular, if $\mathcal{V} = \mathbb{R}^n$ then $\langle u, v\rangle$ is the usual Euclidean inner product $u \cdot v = u_1v_1 + \cdots + u_nv_n$ and $\|v\|^2 = v_1^2 + \cdots + v_n^2$ denotes the squared Euclidean length of $v \in \mathbb{R}^n$. For a matrix A, $C(A)$ is the linear subspace of $\mathbb{R}^n$ generated by the columns of A and I_n is the $n \times n$ identity matrix. The cardinality of a set S will be denoted by $|S|$. In addition, we let $\mathcal{F}(r, s)$ denote the central $\mathcal{F}$ distribution with r numerator and s denominator degrees of freedom, and write $\mathcal{F}(\alpha; r, s)$ for the 100αth percentile of the $\mathcal{F}(r, s)$ distribution. Also, $\bar{\mathcal{F}}(u; r, s) = \Pr(F > u)$ where $F \sim \mathcal{F}(r, s)$ so that $\bar{\mathcal{F}}^{-1}(\alpha; r, s) = \mathcal{F}(1 - \alpha; r, s)$.

2. Matchings on a graph and lack of fit tests

To develop our lack of fit testing procedure, we first recall the definition of a matching associated with a specified graph $G = (V, E)$, where V is a set of points (vertices) and E is a collection of subsets of V, each having cardinality two. If $\{x, y\} \in E$, it is called the edge between x and y. As defined by Korte and Vygen [16], a matching for a graph G is a subset $M \subset E$ such that $\{x, y\} \in M$, $\{x_1, y_1\} \in M$, with $\{x, y\} \neq \{x_1, y_1\}$, implies $\{x, y\} \cap \{x_1, y_1\} = \emptyset$.

We define $\text{supp}M = \cup\{\{x, y\} | \{x, y\} \in M\} \subseteq V$, and say that a matching M is *maximal* if

$$\{x, y\} \in E \text{ implies } \{x, y\} \cap \text{supp}M \neq \emptyset.$$

The elements of $V \backslash \text{supp}M$ are called singleton vertices associated with M. We will denote the collection of maximal matchings associated with the graph G by $\mathcal{M}$.

In turn, subspaces determined by matchings are used to construct test statistics based on clusterings of the statistical units. With such in mind let $\mathbb{R}^V = \{f \,|\, f : V \to \mathbb{R}\}$, and with $(af + bg)(x) = af(x) + bg(x)$ and $\langle f, g\rangle = \sum_{x \in V} f(x)g(x)$ for $f, g \in \mathbb{R}^V$, $x \in V$ and $a, b \in \mathbb{R}$, it follows that $\mathbb{R}^V$ has the structure of a real inner product space. Next, letting M be a matching for G, we define the subspaces

$$V_M = \{f : V \to \mathbb{R} \,|\, \{x, y\} \in M \Longrightarrow f(x) = f(y)\}$$

and

$$V_M^{\perp} = \{f : V \to \mathbb{R} \,|\, \{x, y\} \in M \Longrightarrow f(x) = -f(y) \text{ and } x \notin \text{supp}M \Longrightarrow f(x) = 0\}.$$

Thus, $\mathbb{R}^V = V_M \oplus^{\perp} V_M^{\perp}$, and with $n = |V|$ and $m = |\text{supp}M|$,

$$\dim V_M = m/2 + (n - m) \quad \text{and} \quad \dim V_M^{\perp} = m/2.$$

Note that if we use the identification $\mathbb{R}^V \simeq \mathbb{R}^n$ then a matching M for the graph G corresponds to an edge clustering matrix Z, with V_M corresponding to $C(Z)$. The matrix Z contains only zeros and ones, and the nonzero values in each of $m/2$ columns of Z correspond to the vertices of a particular edge in M. The remaining $n - m$ columns of Z, if any, each contain a single nonzero value corresponding to any singleton vertices associated with M.

We remark that for most purposes in this paper, matchings may be based on the identification $\mathbb{R}^V \simeq \mathbb{R}^n$ as described above. However, in studying clustering on a weighted graph G, the vector space $\mathbb{R}^V$ is an often used construction, with subspaces, angles and operators naturally arising on this space. Such occurs in the hypotheses and proof of our Theorem 1, for example. Another example is the large field of spectral clustering where the graph Laplacian acts on this space, and the important functions $f : V \to \mathbb{R}$ are the eigenvectors of the graph Laplacian. To investigate the relationship between these different forms of clustering, the constructions given above find a natural utility. Furthermore, although our focus in this paper is on testing for between-cluster lack of fit, as defined by Christensen [3,4], the problem of testing for within-cluster lack of fit can also be usefully based on such constructions.

Statistics for testing lack of fit based on matchings can now be constructed. Let $V = \{x_1, \ldots, x_n\}$ consisting of the predictor vectors represent the graph vertices, and for a fixed $\delta > 0$ let an edge set be defined by

$$E_{\delta} = \{\{x_i, x_j\} \,|\, x_i \neq x_j \text{ and } \|x_i - x_j\|^2 < \delta\}.$$

Now for any matching M of the graph $G_{\delta} = (V, E_{\delta})$, we have $\mathbb{R}^n = W \oplus^{\perp} W^{\perp}$ where the lack of fit subspace $W^{\perp}$ follows a decomposition given by Christensen [3,4]. Specifically, using the correspondence $\mathbb{R}^V \simeq \mathbb{R}^n$, we use a special collection of clusterings consistent with the graph, namely the matchings (i.e. edge clusterings) to write

$$W^{\perp} = (V_M \cap W^{\perp}) \oplus^{\perp} (V_M^{\perp} \cap W^{\perp}) \oplus^{\perp} S_M$$

for a matching M. The first two subspaces in this decomposition are called the (orthogonal) between- and within-cluster lack of fit subspaces, respectively, corresponding to a particular matching. The third subspace S_M denotes the orthogonal complement of the sum of the first two subspaces with respect to $W^{\perp}$.

From the preceding, for every matching M, we have a lack of fit test based on the Fisher statistic

$$F_M(\mathbf{y}) = \frac{\dim E_M \, \|P_{B_M}\mathbf{y}\|^2}{\dim B_M \, \|P_{E_M}\mathbf{y}\|^2}$$

where $B_M = V_M \cap W^\perp$ and $E_M = (V_M^\perp \cap W^\perp) \oplus^\perp S_M$. For a particular matching, we may thus conclude lack of fit, that is $\mathrm{E}(\mathbf{y}) \notin W$, whenever $F_M(\mathbf{y}) > \mathcal{F}(1-\alpha; \dim B_M, \dim E_M)$. Recall $\dim V_M = m/2 + (n-m)$ where $m = |\mathrm{supp}M|$, so that (generically) $\dim B_M = \dim V_M - \dim P_{V_M} W = \{m/2 + (n-m)\} - (p+1)$ and $\dim E_M = \dim V_M^\perp - \dim P_{V_M^\perp} W + \dim S_M = (m/2) - (p) + (p)$. The Fisher test statistic is based on the constructed alternative model subspace B_M for a given matching M. With Gaussian errors, the Fisher statistic $F_M(\mathbf{y})$ provides a level test α test for any fixed n, and the test is UMPI against between-cluster lack of fit based on a specified matching M. However, our interest centers on testing $\mathrm{E}(\mathbf{y}) \in W$ against a large class of lack of fit. Such is investigated in Section 4 with simulated data generators, as well as an asymptotic analysis in Section 2.1. Test construction using alternative models is simply a method to develop the test statistic.

The choice of δ and a matching M are of course important in order to distinguish the model $\mathrm{E}(\mathbf{y}) \in W$ from a true data generator. In this regard, there are three points of consideration to be discussed next, including a matching selection criterion. Model-driven choices for δ are presented in Section 3.2.

2.1. Asymptotic behavior of tests against a broad class of alternatives

The first point discusses the role of δ and variance estimation. In particular, $\|P_{E_M}\mathbf{y}\|^2$ must be controlled so that $\|P_{E_M}\mathbf{y}\|^2/\dim E_M$ is a good estimate of the error variance. Such can be achieved by taking δ sufficiently small, as indicated in Theorem 1. Prediction points must accommodate a choice of small δ so that $E_\delta \neq \emptyset$, although $\mathrm{supp}M \neq V$ is permitted. If $E_\delta = \emptyset$ then δ must be readjusted. The choice of $\delta > 0$ restricts the support of our chosen matching M and thus prevents the use of edges that contaminate $\|P_{E_M}\mathbf{y}\|^2$. The large n, small δ behavior of $F_M(\mathbf{y})$ against a broad class of lack of fit is given in Theorem 1.

In Theorem 1, and Lemma 2 in the Appendix, X denotes the p-dimensional subspace of W generated by the columns of an $n \times p$ matrix with rows given by the predictor vectors x_i, $1 \le i \le n$. In the proof of Theorem 1, we assume without loss of generality that $W = \mathbf{1}_n \oplus^\perp X$. We define $T = P_W P_{V_M} P_W$, a self-adjoint operator with $TX \subset X$, and let X_k, $1 \le k \le p$, be an orthonormal basis for X consisting of eigenvectors for T with corresponding eigenvalues λ_k, $1 \le k \le p$. Also, we let $\theta_k = \angle(X_k, V_M)$ with $0 \le \theta_k \le \pi/2$ for $1 \le k \le p$. In the theorem, δ_n determines a graph G_{δ_n} as described above. Finally, we assume independent and identically distributed errors $\{e_i\}$ with finite fourth order moments in the proof of Theorem 1.

Theorem 1. *Suppose the true data generator is*

$$y_i = f(x_i) + e_i \quad \text{for } 1 \le i \le n$$

where $f : U \to \mathbb{R}$ is of class C^1 on an open set $U \subset \mathbb{R}^p$. Suppose $\delta_n \to 0$ as $n \to \infty$ with a corresponding sequence of matchings M_n such that $\theta_{nk} \to 0$, $1 \le k \le p$. Then with predictor vectors contained in any compact convex subset $B \subset U$, we have for large n that $F_{M_n}(\mathbf{y}) \ge F_n(\mathbf{y})$ where

$$F_n(\mathbf{y}) = \frac{\dim E_{M_n} \|P_{B_{M_n}}\mathbf{y}\|^2}{\dim B_{M_n} \sum (e_i - e_j)^2} + o_P(1)$$

where the sum is over all $\{i, j\}$ for which $\{x_i, x_j\} \in M_n$.

As shown in Lemma 2, the condition on the angles θ_{nk} implies that the third space S_{M_n} in the decomposition of $W^\perp$ is marginalized as far as lack of fit is concerned, and thus leads to better variance estimation. Our simulations show this tends to be true for large n. Indeed, Lemmas 1 and 2 given in the Appendix, and used to prove Theorem 1, show that $\|P_{E_{M_n}}\mathbf{y}\|^2/\dim E_{M_n}$ is asymptotically upper bounded by $\sum (e_i - e_j)^2/\dim E_{M_n}$, a model-independent function of the errors which is unbiased for σ^2 for a broad class of true data generators (C^1 functions with errors not necessarily Gaussian). In turn, Theorem 1 concludes the Fisher statistic is asymptotically lower bounded by a random variable with denominator equal to this model-independent unbiased function for σ^2. Since $F_{M_n}(\mathbf{y})$ is a variance-ratio test statistic, it is most effective (i.e., detects departures from the hypothesized model) when the denominator variance estimator well-approximates the error variance σ^2 under not only the hypothesized model but also under departures representing general lack of fit. Given the unknown nature of any underlying lack of fit, a multiple testing procedure is described in Section 3.2. The procedure rejects the adequacy of the model $\mathrm{E}(\mathbf{y}) \in W$ only if at least one of a collection of Fisher statistics is significant. To enhance detection of any model inadequacy, effective variance estimation is thus critical under a broad class of lack of fit.

2.2. A criterion for selecting matchings

The second point involves the specification of a matching selection criterion to choose a maximal matching in E_δ, along with properties of matching sequences for increasing n. Because of Theorem 1, we want to select a matching from

$\mathcal{M}_\delta = \{M \subset E_\delta \mid M \text{ is a maximal matching}\}$ with small weights $w(x_i, x_j) = \|x_i - x_j\|^2$ for $\{x_i, x_j\} \in E_\delta$, which by construction restricts to edges with weights less δ. We determine such a maximal matching edge by edge, at each stage choosing the edge with the smallest possible weight, and define the result to be a *minimal weighted maximal matching*. Note that consistent with smoothing residuals, we are choosing an $M \subset E_\delta$ such that $W_0 := P_{V_M} W$ approximates W. In this case $P_{B_M}\mathbf{y}$, that is $P_{V_M}\mathbf{y} - P_{W_0}\mathbf{y}$, approximates $\mathbf{y} - P_W\mathbf{y}$.

The existence of matching sequences on a hypercube in $\mathbb{R}^p$ with arbitrarily small edge differences for large n is given in Theorem 2. The importance of Theorem 2 stems from the need to consider matchings with small δ as established in Theorem 1. Also, the predictor vectors x_i are assumed to lie in a fixed hypercube, given by S in Theorem 2.

Theorem 2. *With p and $b \in \mathbb{R}$ fixed, let $S = [0, b]^p \subset \mathbb{R}^p$. Suppose we have a sequence of finite subsets, $P_n \subset S$, such that $\lim_{n\to\infty} |P_n| = \infty$, and let $\zeta > 0$. Then there exists n_0 and a sequence of matchings M_n for P_n such that $n \geq n_0$ gives*

$$\{v, w\} \in M_n \Longrightarrow \|v - w\| < \zeta \quad \text{and} \quad \lim_{n\to\infty} |M_n| = \infty.$$

In fact,

$$|M_n| > |P_n|/2 - 2^{pk} - 1$$

where k is chosen such that $\sqrt{p}(b/2^k) < \zeta$. Further, the number of singletons for each M_n is bounded by $2(2^{pk} + 1)$, independent of n for a fixed ζ.

2.3. Computational feasibility for large p

Lastly, a third point addresses the computational feasibility for determining a minimal weighted maximal matching in the case of large p. Since we only use $\mathbb{R}^p$ in determining the distances $\|x_i - x_j\|^2$ for each pair of predictor vertices in V, the calculations are reduced to a graph matching construction on $G_\delta = (V, E_\delta)$, as an abstract graph with weights $w(x_i, x_j) = \|x_i - x_j\|^2$ for $\{x_i, x_j\} \in E_\delta$, and are thus independent of the value of p. An efficient algorithm to compute minimal weighted maximal matchings for graphs G_δ is given in Section 3.1.

3. Implementation of minimal weighted maximal matchings for testing lack of fit

For a specified δ, we consider a graph $G_\delta = (V, E_\delta)$ as defined in Section 2, and present an algorithm to compute a minimal weighted maximal matching for G_δ. In addition, given the unknown nature of the underlying regression function, we implement a testing procedure which uses multiple matchings to enhance detection of general lack of fit associated with the specified model $\mathrm{E}(\mathbf{y}) \in W$. Thus, our lack of fit testing procedure actually involves the specification of a set Δ of δ values, along with a corresponding set of minimal weighted maximal matchings, each of which is determined by the following algorithm. The algorithm is described with $V = \{x_i\}$, followed by a multiple lack of fit testing procedure and a model-driven specification of the set Δ.

3.1. An algorithm for minimal weighted maximal matchings

For a specified value of δ, the following steps provide an algorithm to efficiently compute a minimal weighted maximal matching M_δ for a graph G_δ.

1. Create a list L_δ containing E_δ and corresponding $\|x_i - x_j\|^2$ (if length $L_\delta = 0$ then adjust δ for new L_δ; otherwise continue).
2. Determine the edge in L_δ with minimum $\|x_i - x_j\|^2$.
3. Store this edge in a list M_δ and remove this edge from L_δ, along with any edge connected to the chosen edge.
4. Return to steps 2 and 3 with updated M_δ and L_δ until length $L_\delta = 0$.
5. Identify any remaining singleton vertices associated with M_δ.

An R code was written to implement this algorithm for the simulations in Section 4 and is available from the authors on request.

3.2. A multiple testing procedure for matchings with model-driven choices for δ

The efficacy of a particular choice of δ and corresponding maximal matching depends on the unobservable lack of fit. Thus, we implement a testing procedure using a model-driven collection of maximal matchings. Specifically, given the predictor vectors $\{x_i\}$, a generally applicable choice of δ values, consistent with variance estimation as discussed in Section 2.1, is given by

$$\Delta = \{\eta \,\mathrm{MSD} \mid \mathrm{MSD} = \texttt{mean}\{\|x_i - x_j\|^2, i \neq j = 1, \ldots, n\} \text{ and } \eta = .1, \ldots, .5\},$$

where MSD denotes the mean-squared distance of the edge weights and is computed by the R function `mean`. That is, we use a model-driven choice based on the average of all edge weights to determine the set Δ, borrowing an idea that is sometimes

Table 1
True data generators have $E(\mathbf{y}) \in W + \gamma$LOF where LOF is one of the lack of fit functions given by A, . . . , I.

Key	Lack of fit functions for simulations
A	$\sum_{j=1}^{3} x_j^2$
B	$(\sum_{j=1}^{3} x_j)^2$
C	$\sum_{j=1}^{3} x_j^2 + x_1x_2 + x_1x_3 + x_2x_3 + x_1x_2x_3$
D	$\sum_{j=1}^{6} x_j^2$
E	$(\sum_{j=1}^{6} x_j)^2$
F	$\sum_{j=1}^{6} x_j^2 + x_1x_2 + x_1x_3 + x_2x_3 + x_1x_2x_3$
G	$\sum_{j=1}^{6} x_j^2 + x_1x_2 + x_3x_4 + x_5x_6 + x_1x_2x_3 + x_2x_3x_4 + x_4x_5x_6$
H	$\sum_{j=1}^{12} x_j^2$
I	$(\sum_{j=1}^{12} x_j)^2$

used in the definition of clustering. In particular, a cluster may be defined as a group of points whose intra-point distance is less than the average distance in the pattern as a whole.

The multiple testing approach of Baraud et al. [2] is then implemented based on the corresponding set of minimal weighted maximal matchings $\{M_\delta \mid \delta \in \Delta\}$ as determined by the algorithm in Section 3.1 to test the adequacy of the model $E(\mathbf{y}) \in W$. In particular, we simultaneously employ more than one alternative lack of fit subspace, each based on a maximal matching selected from $\{M_\delta \mid \delta \in \Delta\}$. Given the unknown nature of any potential lack of fit, the objective of such a multiple testing procedure is to increase our chance of detecting a discrepancy associated with the proposed model. Accordingly, we let

$$T_\alpha = \sup_{\delta \in \Delta} \{F_{M_\delta}(\mathbf{y}) - \tilde{\mathcal{F}}^{-1}(\alpha_\delta; \dim B_{M_\delta}, \dim E_{M_\delta})\}$$

and reject $E(\mathbf{y}) \in W$ whenever $T_\alpha > 0$. In the preceding, $\{\alpha_\delta \mid \delta \in \Delta\}$ is a collection of numbers in $(0, 1)$ such that $\Pr(T_\alpha > 0 \mid W) \leq \alpha$. Note that with $\mathbf{e} \sim \mathcal{N}(0, \sigma^2 I_n)$, this multiple testing procedure rejects the adequacy of the specified model if the Fisher statistic $F_{M_\delta}(\mathbf{y})$ for testing $E(\mathbf{y}) \in W$ is significant at level α_δ for some $\delta \in \Delta$. Further, if a_n denotes the α-quantile of the random variable

$$T_n = \inf_{\delta \in \Delta} \tilde{\mathcal{F}}(R_\delta; \dim B_{M_\delta}, \dim E_{M_\delta})$$

where $R_\delta = F_{M_\delta}(\mathbf{e})$, then the choice of $\alpha_\delta = a_n$ for all $\delta \in \Delta$ ensures that $\Pr(T_\alpha > 0 \mid W) = \alpha$. As in Baraud et al. [2], in this paper we use simulation (using Gaussian errors) to determine the value of a_n. This is discussed in Section 4 in connection with simulation studies to assess the ability of our testing procedure to detect lack of fit. Alternatively, a choice of α_δ such that $\sum_{\delta \in \Delta} \alpha_\delta = \alpha$ provides a conservative testing procedure with level at most equal to α according to the Bonferroni inequality.

4. Simulations

Simulations were performed to study the ability of the testing procedure based on T_α to detect lack of fit associated with the model $E(\mathbf{y}) \in W$ involving many predictors, and in particular for regression models with $p = 10$ and $p = 20$. The true data generators were of the form $E(\mathbf{y}) \in W + \gamma$LOF involving various functions of the predictor vectors $\{x_i\}$ as given in Table 1. Note that these functions are not related to the subspaces B_{M_δ} of the test statistics, which ensures realistically meaningful simulations. Tables 2–10 report minimal γ values leading to empirical power 1 for the testing procedure based on T_α, with $\alpha = .05$, corresponding to the various lack of fit functions. That is, the percentage of rejection of the model $E(\mathbf{y}) \in W$ was computed to be approximately 100% for the listed γ and corresponding true data generator representing model lack of fit. The simulated size of our test procedure is included in the captions for Tables 2–10. In addition, an analytical diagnostic, which is developed in Section 4.2, is also included in the tables for evaluating the simulation results. An R code was written to implement the test procedure for the simulations and is available from the authors on request.

4.1. Simulation parameters

For the case $p = 10$, we let $n = 100$ and the $\{x_i\}$ were randomly generated as multivariate normal vectors in $\mathbb{R}^{10}$. Once generated, these predictors were held fixed throughout all simulated datasets for various functions representing model inadequacy. Then the random error vectors $\mathbf{e} \sim \mathcal{N}(0, \sigma^2 I_{100})$ were generated for specified σ to provide 4000 simulated datasets, each with $n = 100$ observations. We used $\sigma = 1.5$ in the simulations for the case $p = 10$. Thus, empirical power is the frequency of rejections of the model $E(\mathbf{y}) \in W$ based on the values of $T_{.05}$ for the 4000 simulated datasets corresponding to a particular lack of fit function listed in Table 1. The same approach was followed for the case $p = 20$ except that we let $\sigma = 1.0$ and used $n = 200$, and generated predictor vectors in $\mathbb{R}^{20}$. We remark that in cases with many predictors, it would

Table 2
$p = 10$ and $n = 100$ with equal group predictor means for 4 groups; simulated size = .05325.

LOF	γ	Diagnostic	Critical point	$\lvert M_\delta \rvert$
A	.05525	13.51831	1.771288	47
B	.01875	14.03796	1.771288	47
C	.00625	12.06046	1.771288	47
D	.02875	14.86045	1.771288	47
E	.00485	15.16785	1.771288	47
F	.00575	13.22626	1.771288	47
G	.00275	12.20669	1.771288	47

Table 3
$p = 10$ and $n = 100$ with rowwise loaded group predictor means for 4 groups; simulated size = .05525.

LOF	γ	Diagnostic	Critical point	$\lvert M_\delta \rvert$
A	.007500	11.33978	1.669059	49
B	.002500	11.52123	1.669059	49
C	.000375	10.83961	1.669059	49
D	.000375	10.87604	1.669059	49
E	.000675	10.98682	1.669059	49
F	.000375	10.85550	1.669059	49
G	.000130	10.14780	1.669059	49

Table 4
$p = 10$ and $n = 100$ with columnwise loaded group predictor means for 4 groups; simulated size = .05200.

LOF	γ	Diagnostic	Critical point	$\lvert M_\delta \rvert$
A	.375000	7.316503	2.198433	25
B	.139750	19.26270	10.70237	4
C	.047250	28.73274	10.70237	4
D	.261250	7.026641	2.198433	25
E	.057500	10.42160	2.198433	25
F	.046875	25.86990	10.70237	4
G	.019500	35.95294	10.70237	4

Table 5
$p = 20$ and $n = 200$ with equal group predictor means for 4 groups; simulated size = .04775.

LOF	γ	Diagnostic	Critical point	$\lvert M_\delta \rvert$
D	.011725	42.41159	1.548174	70
E	.002075	40.90712	1.548174	70
F	.002450	26.79558	1.548174	70
G	.000975	40.46241	1.548174	70
H	.005950	59.25644	1.548174	70
I	.000510	60.30450	1.548174	70

be unrealistic to simulate each predictor according to a particular uniform distribution. To build structure into the predictor vectors, we generated them as multivariate normal vectors for our simulations as described next.

In particular, for $p = 10$, 4 groups of predictor vectors, each of size 25, were generated with the R function `mvnorm`. The mean and covariance parameters for the respective groups were specified as

$$(2\mathbf{1}_{10}, I_{10}), (6\mathbf{1}_{10}, 2I_{10}), (8\mathbf{1}_{10}, 2I_{10}), (12\mathbf{1}_{10}, 4I_{10}),$$

with simulation results presented in Table 2. In addition to this case, simulation results were obtained for other mean parameters, with covariance matrices unchanged. Specifically, simulation results in Table 3 correspond to mean parameters obtained as row vectors in a 4×10 matrix, which was obtained from the R function `matrix` where the integers 1 through 40 are loaded rowwise. Similarly, simulation results in Table 4 correspond to mean parameters obtained as row vectors in a 4×10 matrix obtained by columnwise loading of the integers 1 through 40. Thus, the simulations in Table 3 are based on mean vectors whose corresponding components are dissimilar, while the results in Table 4 have mean vectors with corresponding components which are not dissimilar.

Table 6
$p = 20$ and $n = 200$ with rowwise loaded group predictor means for 4 groups; simulated size = .05175.

LOF	γ	Diagnostic	Critical point	$\lvert M_\delta \rvert$
D	.000355000	46.85538	1.417452	100
E	.000061250	46.99042	1.417452	100
F	.000019250	51.59371	1.417452	100
G	.000006738	54.80082	1.417452	100
H	.000177500	48.61367	1.417452	100
I	.000014250	48.64822	1.417452	100

Table 7
$p = 20$ and $n = 200$ with columnwise loaded group predictor means for 4 groups; simulated size = .05425.

LOF	γ	Diagnostic	Critical point	$\lvert M_\delta \rvert$
D	.105750	3.912653	1.856307	79
E	.022150	4.137984	1.856307	79
F	.029550	4.534516	2.041126	50
G	.007625	3.804703	1.911481	66
H	.014250	4.611947	1.856307	79
I	.001625	4.693023	1.856307	79

Table 8
$p = 20$ and $n = 200$ with equal group predictor means for 8 groups; simulated size = .05600.

LOF	γ	Diagnostic	Critical point	$\lvert M_\delta \rvert$
D	.0033250	25.87497	1.461465	93
E	.0005400	26.46414	1.461465	93
F	.0004325	16.80317	1.461465	93
G	.0001540	23.76868	1.461465	93
H	.0016750	28.45851	1.461465	93
I	.0001350	29.64136	1.461465	93

Table 9
$p = 20$ and $n = 200$ with rowwise loaded group predictor means for 8 groups; simulated size = .05200.

LOF	γ	Diagnostic	Critical point	$\lvert M_\delta \rvert$
D	.000083310	19.44710	1.448027	100
E	.000014150	19.43938	1.448027	100
F	.000002211	19.35169	1.448027	100
G	.000000761	19.01732	1.448027	100
H	.000041750	19.05819	1.448027	100
I	.000003544	19.07397	1.448027	100

Table 10
$p = 20$ and $n = 200$ with columnwise loaded group predictor means for 8 groups; simulated size = .04575.

LOF	γ	Diagnostic	Critical point	$\lvert M_\delta \rvert$
D	.0354300	5.97736	2.047189	30
E	.0062920	6.142708	2.047189	30
F	.0043140	7.214581	1.676749	70
G	.0009167	6.715452	2.047189	30
H	.0067750	6.594721	2.047189	30
I	.0007275	6.336556	2.047189	30

For $p = 20$, simulations were obtained for groups of size 50 and 25. For the case of size 50, 4 groups of predictor vectors were again generated with the R function `mvnorm`. The mean and covariance parameters for the respective groups were specified as

$$(2\mathbf{1}_{20}, I_{20}), (6\mathbf{1}_{20}, 2I_{20}), (8\mathbf{1}_{20}, 2I_{20}), (12\mathbf{1}_{20}, 4I_{20}),$$

with simulation results presented in Table 5. Similar to the $p = 10$ case, simulation results were obtained for other mean parameters, with covariance matrices unchanged. In particular, simulations in Tables 6 and 7 correspond, respectively, to mean parameters obtained from 4×20 matrices obtained from the R function `matrix` by rowwise and columnwise loading

of the integers 1 through 80. Finally, for size 25, 8 groups of predictor vectors were generated with mean and covariance parameters for the respective groups specified as

$$(2\mathbf{1}_{20}, I_{20}), (6\mathbf{1}_{20}, 2I_{20}), (8\mathbf{1}_{20}, 2I_{20}), (12\mathbf{1}_{20}, 4I_{20})$$
$$(14\mathbf{1}_{20}, 4I_{20}), (18\mathbf{1}_{20}, 6I_{20}), (20\mathbf{1}_{20}, 6I_{20}), (24\mathbf{1}_{20}, 1I_{20}),$$

with simulation results presented in Table 8. As in the preceding, simulation results were obtained for other mean parameters, with covariance matrices unchanged. Specifically, simulations in Tables 9 and 10 correspond, respectively, to mean parameters obtained from 8×20 matrices obtained from the R function `matrix` by rowwise and columnwise loading of the integers 1 through 160.

The test statistic T_α was computed with $\alpha_\delta = a_n$ for all $\delta \in \Delta$ to ensure a level α testing procedure, with Δ as indicated in Section 3.2. The value of a_n was calculated by simulation as follows. First, 10,000 values of the random variable T_n were generated. To do so, 10,000 observations of the random vector $\mathbf{e}$, which appears in the random variables R_δ, $\delta \in \Delta$, were generated according to the $\mathcal{N}(0, I_n)$ distribution. Then the α-quantile of the corresponding values of T_n was computed to approximate the value of a_n. As noted above, we set $\alpha = .05$.

4.2. A diagnostic to evaluate lack of fit detection capability

A diagnostic that is useful for assessing detection capability of model lack of fit in our simulations is based on the following theorem. It will be seen that the diagnostic uses only the specified predictors, not the observable responses. As in the proof of Theorem 1, we assume independent and identically distributed errors $\{e_i\}$ with finite fourth order moments in the proof of Theorem 3.

Theorem 3. *Suppose the true data generator is given by* $\mathbf{y} = \mathbf{w}_0 + \gamma\xi_0 + \mathbf{e}$ *where* $\mathbf{w}_0 \in W$ *and* $\xi_0 \in W^\perp$ *with* $\|\xi_0\| = 1$ *where* $P_{B_{M_\delta}}\xi_0 \neq 0$ *and* $P_{E_{M_\delta}}\xi_0 \neq 0$, *and* $\xi_0 \in W^\perp$ *is constructed as* $\xi_0 = \xi_1 \oplus \xi_2 \in B_{M_\delta} \oplus^\perp E_{M_\delta}$. *Then for any* $\delta \in \Delta$,

$$\lim_{\gamma\to\infty} F_{M_\delta}(\mathbf{y}) = \lambda(M_\delta, \xi_0) + o_P(1)$$

where

$$\lambda(M_\delta, \xi_0) := \frac{\dim E_{M_\delta}\ \|\xi_1\|^2}{\dim B_{M_\delta}\ \|\xi_2\|^2}.$$

Thus, rejection of $\mathrm{E}(\mathbf{y}) \in W$ by the test based on T_α requires, at least to approximation, that

$$\lambda(M_\delta, \xi_0) > \bar{\mathcal{F}}^{-1}(\alpha_\delta; \dim B_{M_\delta}, \dim E_{M_\delta})$$

for some $\delta \in \Delta$.

For each of the lack of fit functions in Table 1 we evaluate the function at the n predictor settings to get a vector in $\mathbb{R}^n$, which we project into $W^\perp$ to get a vector ξ_0 and then obtain the diagnostic $\lambda(M_\delta, \xi_0)$. This diagnostic is reported in the tables below as an indicator of the ability of the test based on T_α to detect lack of fit when the true data generator is not necessarily constructed from the specified alternative model space B_{M_δ} for some $\delta \in \Delta$. In particular, we report the diagnostic and corresponding critical point in the preceding inequality with largest difference. The diagnostic being greater than the critical point (given in the tables) guarantees that we will obtain power 1 for sufficiently large γ. We also give the smallest value of γ which gives power 1. In addition, we report the number of matchings in the minimal weighted maximal matching for the corresponding $\delta \in \Delta$.

4.3. Tables and discussion of simulation results

As indicated in the captions of Tables 2–10, all simulated cases have empirical size nearly equal to the desired $\alpha = .05$. Also, the tables report minimal γ values for which the testing procedure based on T_α achieves empirical power 1 for the lack of fit functions specified in Table 1, and based on the predictor vectors $\{x_i\}$ discussed in Section 4.1. As can be observed in the simulation results, the diagnostic developed in Section 4.2 provides a means of assessing the effectiveness of the test statistic for a given data generator. In particular, diagnostic values and corresponding critical point with largest difference are reported in the tables, which demonstrates the effectiveness of T_α to detect general lack of fit. In addition, the number of matchings in the associated M_δ is reported, which indicates greater numbers of matchings corresponding to cases with predictor vectors whose corresponding components are dissimilar. In particular, Tables 2 and 3, corresponding to $p = 10$, $n = 100$ simulations, and Tables 5, 6, 8 and 9, corresponding to $p = 20$, $n = 200$ simulations, all have either equal group predictor means or rowwise loaded group predictor means as described in Section 4.1. In either case, predictor vectors between groups have corresponding components that are dissimilar, and thus provide vectors with a clustering structure in higher dimensional predictor space. In all corresponding table entries, the cardinality of the selected matchings is consistently near or at the possible maximal number across the lack of fit functions. On the other hand, Table 4, corresponding

to $p = 10$, $n = 100$ simulations, and Tables 7 and 10, corresponding to $p = 20$, $n = 200$ simulations, have group predictor means loaded columnwise as described in Section 4.1. Such predictor vectors have corresponding components that are not dissimilar, and thus do not display a clustered structure. As seen in these tables, the cardinality of the matchings is not consistently near possible maximal values, and for certain lack of fit functions the number of matchings chosen is relatively low. However, the ability of T_α to detect lack of fit in all simulated cases indicates our testing procedure effectively adapts to cases with smaller cardinality of M_δ as well.

Here we also note the usefulness of the multiple testing procedure i.e., there were cases when the diagnostic was less than the corresponding critical point indicating failure to detect model lack of fit. However, the use of T_α overcame this problem in all simulated cases. Furthermore, we point out that if the simulated response noise were to be increased (i.e., increased σ for the error vector $\mathbf{e}$), then we may have to take γ larger in order to achieve empirical power equal to 1. However, for our simulations, it is guaranteed (because of the diagnostic) that the test statistic will exceed the critical point for at least one $\delta \in \Delta$ for sufficiently large γ. Finally, we note that the test based on T_α was successful in detecting lack of fit when the true data generator was not necessarily constructed from the specified alternative model spaces B_{M_δ}, $\delta \in \Delta$. Thus, the simulation results, along with the asymptotic analysis in Section 2.1, support our interest in effectively testing $\mathrm{E}(\mathbf{y}) \in W$ against a large class of lack of fit.

4.4. Discussion and simulation results for non-smooth and irregular lack of fit

The purpose of this paper was to give computable lack of fit tests for linear regression models with many predictor variables which are effective against a large class of lack of fit functions, including piecewise continuous polynomial functions. This is seen *asymptotically* in Theorem 1, which is based on choosing matchings that effectively provide model-independent variance estimates by controlling the terms $\|P_{V_{M_n}^\perp}\mathbf{y}\|^2$ and $\|P_{S_{M_n}}\mathbf{y}\|^2$ as seen in Lemmas 1 and 2, respectively. For such control, the variation of the lack of fit function within the matchings must not be too large. For a *fixed finite n*, this has to do with the behavior of the lack of fit function in comparison with the design (predictor) points. To elucidate this point, for a lack of fit function $f(x)$ and a subset $\mathcal{A}$ of the predictor space, let

$$\mathbf{V}_f(\mathcal{A}) = \sup\{|f(u) - f(v)| \ \mid u, v \in \mathcal{A}\}.$$

If $\mathbf{V}_f(\mathcal{A})$ is small for most $\mathcal{A}$ which satisfy $(\text{diameter}\mathcal{A})^2 \leq \eta\text{MSD}$, where MSD is defined in Section 3.2 and $0 < \eta < 1$, we say that the lack of fit is *smooth* for this design. Although purposefully and necessarily somewhat vague, this idea of smoothness with respect to a design has diagnostic implications. For example, for the tests developed in this paper it is a condition for being able to detect lack of fit. In fact, for large p, we can expect to detect lack of fit only if the lack of fit is smooth with respect to the design. Even for smooth functions this can fail for *fixed n*. For example, functions which are rapidly oscillating, say $\cos(\nu x)$ for ν large. For discontinuous functions, if the sets of discontinuity are not too dense, the procedure (helped by multiple testing) may still succeed. The following simulations illustrate this.

To investigate the performance of the test procedure based on T_α when lack of fit associated with the model $\mathrm{E}(\mathbf{y}) \in W$ is irregular, we used the $p = 10$ dimensional predictor vectors which were generated for Table 3, again with $n = 100$. For simulations with non-smooth lack of fit, we used the `R` function `stepfun`. In particular, define a step function f_o by

$$f_o(u) = \begin{cases} 0, & \text{if } u < 5; \\ 5, & \text{if } 5 \leq u < 15; \\ 10, & \text{if } 15 \leq u < 25; \\ 5, & \text{if } 25 \leq u < 35; \\ 10, & \text{if } u \geq 35, \end{cases}$$

and let f_p be defined similarly but taking value 400 for $15 \leq u < 25$. The function f_o was used to represent non-smooth lack of fit at relatively low levels while f_p provides non-smooth lack of fit with a peak. Lack of fit denoted by J and K in Table 11 were used to simulate non-smooth lack of fit with one and two peaks, respectively. To simulate smooth but oscillating lack of fit, the cosine function was employed as with L and M in Table 11.

Simulations were carried out in the same manner as described in Section 4.1 for functions representing model inadequacy as given in Table 11. Table 12, which is structured the same as Table 3, gives the results of these simulations and shows that for the particular parameters and predictors used in these simulations, the test based on T_α was successful. However, we note that additional simulations were performed with lack of fit specified by only the function $25\cos(2x_1)$. In this case, the test procedure was not successful, the (asymptotic) diagnostic and corresponding critical point with largest difference given by 1.642214 and 1.669059, respectively, with the number of matchings in the corresponding minimal weighted maximal matching observed to be 49. These results illustrate the ideas concerning the detection of lack of fit in relation to smoothness with respect to a design for large p as discussed above.

5. Concluding remarks

A multiple testing procedure for assessing linear regression model adequacy with many predictors was presented, and shown to be computationally efficient and effective against a broad class of lack of fit. Our approach to constructing general

Table 11
True data generators have $E(\mathbf{y}) \in W + \gamma$LOF where LOF is one of the non-smooth or oscillating lack of fit functions given by J, . . . , M.

Key	lack of fit functions for simulations
J	$f_o(x_1) + f_p(x_2) + f_o(x_3)$
K	$f_p(x_1) + f_p(x_2) + f_o(x_3)$
L	$25\cos(x_1)$
M	$25\cos(2x_1) + x_2^2 + x_3^2$

Table 12
$p = 10$ and $n = 100$ with rowwise loaded group predictor means for 4 groups; simulated size = .05525.

LOF	γ	Diagnostic	Critical point	$\lvert M_\delta \rvert$
J	.0100000	15.412370	1.670107	48
K	.0050000	13.655720	1.670107	48
L	.5000000	2.014524	1.669059	49
M	.0106250	10.679910	1.669059	49

alternatives for model comparison is based on model-driven minimal weighted maximal matchings of the statistical units. These maximal matchings allow our testing procedure to be efficiently implemented in higher dimensional problems, unlike other tests (e.g. smoothing based tests) which are prohibited from implementation due to complexity involving the curse of dimensionality. Indeed, since we only use $\mathbb{R}^p$ in determining the distance between each pair of predictor vectors, the calculations are reduced to a graph matching construction, as an abstract graph with edges weighted by such distances, and are thus independent of the value of p.

Appendix

Lemma 1. *Under the assumptions for Theorem 1,*

$$\|P_{V_{M_n}^\perp}\mathbf{y}\|^2/\dim E_{M_n} \le \sum (e_i - e_j)^2/2\dim E_{M_n} + o_P(1).$$

Proof of Lemma 1. Since f is of class C^1 on U, $|f(u) - f(v)| \le L\|u - v\|\ \forall u, v \in B$ where L is a Lipschitz constant for f. Thus, with $L_* = L\delta_n$ and $m_n = |\text{supp}M_n|$,

$$\|P_{V_{M_n}^\perp}\mathbf{y}\|^2 = \sum |y_i - y_j|^2/2 \le m_n L_*^2/4 + L_* \sum |e_i - e_j| + \sum (e_i - e_j)^2/2$$

where the sum is over all $\{i, j\}$ for which $\{x_i, x_j\} \in M_n$. Since $\delta_n \to 0$ as $n \to \infty$, $m_n L_*^2/4\dim E_{M_n} = L^2\delta_n^2/2 \to 0$ as $n \to \infty$. Further, since $\sum |e_i - e_j|/\dim E_{M_n} = O_P(1)$, it follows that $L_* \sum |e_i - e_j|/\dim E_{M_n} = o_P(1)$. □

Lemma 2. *Under the assumptions for Theorem 1,*

$$\|P_{S_{M_n}}\mathbf{y}\|^2/\dim E_{M_n} \le \sum (e_i - e_j)^2/2\dim E_{M_n} + o_P(1).$$

Proof of Lemma 2. First note $\lambda_{ni}X_i = TX_i = P_W(P_{V_{M_n}}X_i)$, $1 \le i \le p$, so that

$$\lambda_{ni}X_i = P_{J_n}(P_{V_{M_n}}X_i) \oplus^\perp \sum_{j=1}^{p} (P_{V_{M_n}}X_i \cdot X_j)X_j$$

and thus $P_{V_{M_n}}X_i \cdot X_j = 0$ for $i \ne j$, and $\lambda_{ni} = \|P_{V_{M_n}}X_i\|^2 = \cos^2\theta_{ni}$. Next note $S_{M_n} = P_{W^\perp}P_{V_{M_n}}W$ and define $Q = P_{V_{M_n}} - T$, from which it follows $S_{M_n} = QW = QX$ and $QX_i \cdot QX_j = 0$ for $i \ne j$. Letting $\kappa = \{i \mid 1 \le i \le p \text{ and } 0 < \theta_{ni} < \pi/2\}$, it follows that $\{QX_i \mid i \in \kappa\}$ is an orthogonal basis for S_{M_n}. Since $QX_i \in W^\perp$, $\mathbf{y} \cdot QX_i = P_{W^\perp}\mathbf{y} \cdot QX_i = -P_{W^\perp}\mathbf{y} \cdot P_{V_{M_n}^\perp}X_i$ using $QX_i = (1 - \lambda_{ni})X_i - P_{V_{M_n}^\perp}X_i$. Thus,

$$\|P_{S_{M_n}}\mathbf{y}\|^2 = \sum_{i\in\kappa} (\mathbf{y} \cdot QX_i)^2/\|QX_i\|^2 = \sum_{i\in\kappa} (P_{V_{M_n}^\perp}P_{W^\perp}\mathbf{y} \cdot P_{V_{M_n}^\perp}X_i)^2/\|QX_i\|^2.$$

Letting $U_i = P_{V_{M_n}^\perp}X_i/\|P_{V_{M_n}^\perp}X_i\|$ and noting that $\|QX_i\|^2 = \lambda_{ni}(1 - \lambda_{ni})$ and $\|P_{V_{M_n}^\perp}X_i\|^2 = 1 - \lambda_{ni}$, we have

$$\|P_{S_{M_n}}\mathbf{y}\|^2 = \sum_{i\in\kappa} (P_{V_{M_n}^\perp}P_{W^\perp}\mathbf{y} \cdot U_i)^2/\lambda_{ni}.$$

Now $P_{V_{M_n}^\perp} P_{W^\perp}\mathbf{y} = P_{V_{M_n}^\perp}\mathbf{y} - \sum_{i=1}^p (\mathbf{y}\cdot X_i) P_{V_{M_n}^\perp} X_i$. Since $\mathrm{E}\|\mathbf{y}\|^4 = O(n^2)$, we have $\|\mathbf{y}\|^2 = O_p(n)$. Also, $\|P_{V_{M_n}^\perp} X_i\|^2 = 1 - \lambda_{ni} = \sin^2\theta_{ni}$. Thus

$$\left\| \sum_{i=1}^p (\mathbf{y}\cdot X_i) P_{V_{M_n}^\perp} X_i \right\|^2 \le O_p(n) \left(\sum_{i=1}^p \sin^2\theta_{ni} \right)^2 .$$

Since $\theta_{ni} \to 0$ for $1 \le i \le p$, and using Lemma 1, we have that Lemma 2 holds. □

Proof of Theorem 1. Since $\|P_{E_{M_n}}\mathbf{y}\|^2 = \|P_{V_{M_n}^\perp \cap W^\perp}\mathbf{y}\|^2 + \|P_{S_{M_n}}\mathbf{y}\|^2$ and

$$\|P_{V_{M_n}^\perp \cap W^\perp}\mathbf{y}\|^2 \le \|P_{V_{M_n}^\perp}\mathbf{y}\|^2,$$

it follows that

$$F_{M_n}(\mathbf{y}) \ge \frac{\dim E_{M_n} \|P_{B_{M_n}}\mathbf{y}\|^2}{\dim B_{M_n} (\|P_{V_{M_n}^\perp}\mathbf{y}\|^2 + \|P_{S_{M_n}}\mathbf{y}\|^2)}.$$

Thus, by Lemmas 1 and 2, $F_{M_n}(\mathbf{y}) \ge F_n(\mathbf{y})$ where

$$F_n(\mathbf{y}) = \frac{\dim E_{M_n} \|P_{B_{M_n}}\mathbf{y}\|^2}{\dim B_{M_n} \sum (e_i - e_j)^2} + o_P(1). \quad \square$$

Lemma 3. *With p and b fixed, let $P = \{v_1, \ldots, v_n\} \subset [0,b]^p \subset \mathbb{R}^p$ be n distinct points. For k fixed, let $n \ge 2^{pk+1}$. Then there exist distinct points v_i, v_j in P such that $\|v_i - v_j\| \le \sqrt{p}(b/2^k)$.*

Proof of Lemma 3. Let $S_0 = [0,b]^p \subset \mathbb{R}^p$ and divide each of the edge intervals into two equal parts $[0,b] = [0,b/2] \cup [b/2,b]$, thus dividing S_0 into 2^p hypercubes with edge length $b/2$. Choose S_1 to be one of these hypercubes such that $|P \cap S_1| \ge n/2^p$. Continue in this way to obtain $S_0 \supset S_1 \supset S_2 \supset \cdots \supset S_k$ where S_k is a $b/2^k \times \cdots \times b/2^k$ hypercube $\mathbb{R}^p$ and $|P \cap S_k| \ge n/2^{pk}$. Also note that diameter $S_k = \sqrt{p}(b/2^k)$. With $n \ge 2^{pk+1}$, $|P \cap S_k| \ge 2$ so that there exist two distinct points v_i, v_j in P and S_k. Thus, $\|v_i - v_j\| \le$ diameter $S_k = \sqrt{p}(b/2^k)$. □

Proof of Theorem 2. Suppose $\sqrt{p}(b/2^k) < \zeta$. Since $\lim_{n\to\infty} |P_n| = \infty$, there exists n_0 such that $n \ge n_0$ implies $|P_n|/2 - 2^{pk} \ge 2$. Next choose the integer t_n such that $t_n + 1 > |P_n|/2 - 2^{pk} \ge t_n$ so that $\lim_{n\to\infty} t_n = \infty$. Thus, $|P_n| \ge (2)(2^{pk}) + 2t_n$. This implies, using Lemma 3, that there exists a matching M_n of P_n such that $|M_n| = t_n$ and $\{v, w\} \in M_n$ implies $\|v - w\| \le \sqrt{p}(b/2^k) < \zeta$. Finally, $|M_n| = t_n > |P_n|/2 - 2^{pk} - 1$. Thus, letting the number of singletons for M_n be s_n, $t_n = |P_n|/2 - s_n/2 > |P_n|/2 - (2^{pk} + 1)$, which gives $s_n < 2(2^{pk} + 1)$. □

Proof of Theorem 3. Let $\delta \in \Delta$ and note

$$F_{M_\delta}(\mathbf{y}) = \frac{\dim E_{M_\delta} \|\gamma P_{B_{M_\delta}} \xi_o + P_{B_{M_\delta}}\mathbf{e}\|^2}{\dim B_{M_\delta} \|\gamma P_{E_{M_\delta}} \xi_o + P_{E_{M_\delta}}\mathbf{e}\|^2}.$$

Thus, it suffices to show $\|P_{B_{M_\delta}}\mathbf{e}/\gamma\|^2$ and $\|P_{E_{M_\delta}}\mathbf{e}/\gamma\|^2$ are $o_P(1)$ as $\gamma \to \infty$. Since both terms are no larger than $\|\mathbf{e}\|^2/\gamma^2$, which is $o_P(1)$ as $\gamma \to \infty$, we have

$$\lim_{\gamma\to\infty} F_{M_\delta}(\mathbf{y}) = \frac{\dim E_{M_\delta} \|\xi_1\|^2}{\dim B_{M_\delta} \|\xi_2\|^2} + o_P(1). \quad \square$$

References

[1] M. Aerts, G. Claeskens, J.D. Hart, Testing lack of fit in multiple regression, Biometrika 87 (2000) 405–424.
[2] Y. Baraud, S. Huet, B. Laurent, Adaptive tests of linear hypotheses by model selection, Ann. Statist. 31 (2003) 225–251.
[3] R. Christensen, Lack of fit based on near or exact replicates, Ann. Statist. 17 (1989) 673–683.
[4] R. Christensen, Small-sample characterizations of near replicate lack of fit tests, J. Amer. Statist. Assoc. 86 (1991) 752–756.
[5] R. Christensen, Y. Lin, Lack-of-fit tests based on partial sums of residuals, Commun. Stat. - Theory Methods 44 (2015) 2862–2880.
[6] R. Christensen, S.K. Sun, Alternative goodness-of-fit tests for linear models, J. Amer. Statist. Assoc. 105 (2010) 291–301.
[7] H. Dette, A consistent test for the functional form of a regression based on a difference of variance estimators, Ann. Statist. 27 (1999) 1012–1040.
[8] R. Eubank, C.-S. Li, S. Wang, Testing lack of fit of parametric regression models using nonparametric regression techniques, Statist. Sinica 15 (2005) 135–152.
[9] J. Fan, L.-S. Huang, Goodness-of-fit tests for parametric regression models, J. Amer. Statist. Assoc. 96 (2001) 640–652.
[10] J. Fan, C. Zhang, J. Zhang, Generalized likelihood ratio statistics and Wilks phenomenon, Ann. Statist. 29 (2001) 153–193.
[11] R.A. Fisher, The goodness of fit of regression formulae and the distribution of regression coefficients, J. R. Stat. Soc. 85 (1922) 579–612.
[12] E. Guerre, P. Lavergne, Data-driven rate-optimal specification testing in regression models, Ann. Statist. 33 (2005) 840–870.
[13] W. Härdle, E. Mammen, Comparing nonparametric versus parametric regression fits, Ann. Statist. 21 (1993) 1926–1947.

[14] J.D. Hart, Nonparametric Smoothing and Lack of Fit Tests, Springer, New York, 1997.
[15] E.V. Khmaladze, H.L. Koul, Martingale transforms goodness-of-fit tests in regression models, Ann. Statist. 32 (2004) 995–1034.
[16] J. Korte, J. Vygen, Combinatorial Optimization, Theory and Algorithms, fifth ed., Springer, New York, 2012.
[17] H.L. Koul, P. Ni, Minimum distance regression model checking, J. Statist. Plann. Inference 119 (2004) 109–141.
[18] P. Lavergne, V. Patilea, One for all and all for one: regression checks with many regressors, J. Amer. Statist. Assoc. 30 (2012) 41–52.
[19] F.R. Miller, J.W. Neill, General lack of fit tests based on families of groupings, J. Statist. Plann. Inference 138 (2008) 2433–2449.
[20] F.R. Miller, J.W. Neill, B.W. Sherfey, Maximin clusters for near replicate regression lack of fit tests, Ann. Statist. 26 (1998) 1411–1433.
[21] F.R. Miller, J.W. Neill, B.W. Sherfey, Implementation of a maximin power clustering criterion to select near replicates for regression lack of fit tests, J. Amer. Statist. Assoc. 94 (1999) 610–620.
[22] W. Song, J. Du, A note on testing the regression functions via nonparametric smoothing, Canad. J. Statist. 39 (2011) 108–125.
[23] J.G. Staniswalis, T.A. Severini, Diagnostics for assessing regression models, J. Amer. Statist. Assoc. 86 (1991) 684–692.
[24] W. Stute, Nonparametric model checks for regression, Ann. Statist. 25 (1997) 613–641.
[25] W. Stute, W. Gonzalez Manteiga, M. Presedo Quindimil, Bootstrap approximations in model checks for regression, J. Amer. Statist. Assoc. 93 (1998) 141–149.
[26] W. Stute, S. Thies, L.X. Zhu, Model checks for regression: an innovation process approach, Ann. Statist. 26 (1998) 1916–1934.
[27] J.Q. Su, L.J. Wei, A lack of fit test for the mean function in a generalized linear model, J. Amer. Statist. Assoc. 86 (1991) 420–426.
[28] J.X. Zheng, A consistent test of functional form via nonparametric estimation techniques, J. Econometrics 75 (1996) 263–289.