Assessing linearity in high dimensions[1]

Art B. Owen
Stanford University

Summary

Some standard numerical problems become intractable in high dimensions. Yet successes are often achieved in practice. This may be explained in terms of the underlying target function being somehow simpler than assumed in the intractability arguments. A prototypical form of simplicity is approximate linearity. In moderate dimensions linearity can be investigated by linear regression. In very high dimensions this becomes computationally inefficient and eventually infeasible, as the cost of regression for n observations in d dimensions grows as nd^2. This paper presents a quasi-regression method for determining the degree of linearity in a function, where the cost grows only as nd. A bias corrected version of quasi-regression is able to estimate the degree of linearity with a sample size of order $d^{2/3}$. An example is given of a function on $[0,1]^{1,000,000}$, for which the amount of linear variability is accurately estimated from only $100,000$ observations.

1 Introduction

Let f be a function of d variables that are free to vary independently over the unit cube $[0,1]^d$. As d increases, numerical problems like integration, approximation, and non-convex optimization can become computationally intractable, particularly in worst case analyses. Yet it is also common for specific algorithms to succeed on specific high dimensional problems.

One explanation is that the high dimensional functions appearing in these success stories do not have full d dimensional complexity. The function may have a lower effective dimension than d. See Paskov & Traub (1995) and Caflisch, Morokoff & Owen (1997) for examples. For mathematical descriptions, see Caflisch et al. (1997), Hickernell & Wozniakowski (2000) and Sloan & Wozniakowski (1998). Some special forms of simplicity involve

[1] *AMS 1991 Subject classifications.* Primary 65C05 65D15. Secondary 41A58 41A63.
Key words and phrases. Computer Experiment, Function Mining, Monte Carlo, Quasi-interpolation, Quasi-regression, Tractability.
This work was supported by NSF grant DMS-9704495.

the underlying function f being nearly additive, nearly quadratic, or nearly linear.

It is of interest in applications to detect the presence of such simplified structure, in order to know whether the problem at hand can be treated by simpler methods. A condition like additivity is comparatively easy to test: if replicated Latin hypercube samples (McKay, Beckman & Conover 1979) yield a much smaller variance than ordinary Monte Carlo sampling, then from the results in Stein (1987), the function is nearly additive.

This paper focuses on the specific problem of assessing the amount of linearity in a function, while An & Owen (1999) consider higher order response surface models. If a function is sufficiently linear, then it may pay to develop a linear approximation, in order to exploit special purpose algorithms. For an industrial problem that is nearly linear, linear programming becomes available for optimization. If the value of a portfolio of financial instruments is nearly linear in the returns of a large number of underlying assets, then that portfolio is comparatively simple to hedge. Finally, a linear function can be represented graphically by plotting or color coding the coefficient values.

Our interest in this problem was sparked by a finding in Caflisch et al. (1997). There, an integration problem motivated by valuation of collateralized mortgage obligations (CMOs) appears as an integral over the unit cube in 360 dimensions. The problem has one dimension for each monthly interest rate fluctuation, for a portfolio of thirty year mortgages. Interest rate fluctuations affect the value of the portfolio in two ways: they change the discount factor applied to the monthly payments and they change the rate of prepayment. Prepayments reduce all future payments. Given simple models for interest rate fluctuations and prepayment tendencies, the present value of the CMO is a 360 dimensional integral. The function appears to be highly nonlinear, and yet it could be very accurately integrated by some low discrepancy methods. In fact Latin hypercube sampling was also very effective, beating some low discrepancy methods. This made it clear that the function was nearly additive. Inspection of the additive parts showed that the function was in fact nearly linear.

In $d = 360$ dimensions it is not difficult to fit a linear model and inspect it. But this problem raised the following challenge: How would one assess the amount of linearity, in a very high dimensional function?

Section 2 defines the amount of linearity in a function through the variance of a best linear approximation. Section 3 describes direct methods based on estimating this linear approximation. These methods are ordinary least squares regression and the computationally less expensive method of

quasi-regression. Section 4 presents exact expressions for the bias and variance of the quasi-regression based estimate, without assuming $n > d$. This allows the construction of a bias corrected version of quasi-regression that can estimate the amount of linear variation in a function using $n = O(d^{2/3})$ observations. Section 5 finds expressions for the bias and variance of the least squares regression based estimate, in the case where $n \to \infty$ with d fixed. Section 6 presents a variance reduction for quasi-regression. Some examples including one with $d = 1,000,000$ are presented in Section 7, and conclusions are given in Section 8.

2 Examples, Notation and Definitions

This paper considers d dimensional functions in some limits as $d \to \infty$. To fix ideas, we first consider some settings with high dimensional functions on $[0,1]^d$, where it is natural to consider an increasing sequence of values for d. In other settings we may simply have a function over $[0,1]^d$ with large d and no natural interpretation for how that function would change if d were increased.

Example 2.1 (Extension) *A function may describe some aspect, such as the average waiting time, of the performance of a queue over t time steps. Each time step may require s input variables to describe. There are then $d = st$ input variables in the system, and d increases with t.*

Example 2.2 (Refinement) *Some functions considered in financial mathematics depend a Brownian motion W_t over $t \in [0,T]$. It takes an infinite number of random variables to describe W_t. In practice one might represent W_t by $\sum_{j=1}^{d} Z^j \psi_j(t)$ where $\psi_j(t)$ are given functions on $[0,T]$ and Z^j are independent $N(0,1)$ random variables. Increasing d corresponds to exploiting finer structure in W_t.*

Example 2.3 (Computer experiments) *An engineer may have a software model of a physical process, such as the performance of an aircraft wing or a semiconductor. The model may allow an extremely large number D of variables to be specified. The engineer may arrange these variables from those thought to be most important to those thought to be least important. Then it is natural to consider a function defined over the first d input variables with the last $D - d$ input variables fixed at default values.*

For computer experiments it is reasonable from Taylor's theorem, to expect f to be nearly linear over a small d dimensional region. The practical

difficulty is knowing whether a given region is small enough, and exceptions can arise if the region contains a stationary point of the function.

It will often be the case that the cost of a function is either constant in d or increases linearly in d. For example to follow a simulation through d steps would often take computation of order d. For a computer experiment with D variables of which d are being changed from their defaults, the computation may be constant in d. Some realistic functions will require $O(d^2)$ time to compute, but these are outside the scope of this paper.

2.1 Definitions

This section defines how linear a function is. The approach is through the best linear approximation in a least squares sense.

A typical point in $[0,1]^d$ is written $X = (X^1, X^2, \ldots, X^d)$. A data point is written $X_i = (X_i^1, X_i^2, \ldots, X_i^d)$. The underlying data are assumed to be sampled using $X_i^j \sim U(0,1)$ independently for $1 \le i \le n$ and $1 \le j \le d$. Let the function f be defined on the unit cube $[0,1]^d$, with $\int f(X)^2 dX < \infty$. Here and elsewhere an integral without a specific domain is understood to be over $[0,1]^d$.

A linear function takes the form

$$g(X) = \gamma_0 + \sum_{r=1}^{d} \gamma_r \phi(X^r) \tag{1}$$

where the functions ϕ is a fixed *known univariate* function with $\int_0^1 \phi(x)dx = 0$ and $\int_0^1 \phi(x)^2 dx = 1$, and the γ_r are scalars. This definition is motivated by the following two examples.

Example 2.4 (Ordinary linearity) *Taking $\phi(x) = \sqrt{12}(x - 0.5)$, for $r \ge 1$, recovers the usual definition of linearity on $[0,1]^d$.*

Example 2.5 (Gaussian linearity) *Let $\Phi(x) = \int_{-\infty}^{x} (2\pi)^{-1/2} \exp(-z^2/2)dz$. Then taking $\phi(x) = \Phi^{-1}(x)$, for $r \ge 1$, yields an ordinary linear function of independent normally distributed input variables on R^d.*

The linear function f_{L} that minimizes $\int (f(X) - f_{\mathrm{L}}(X))^2 dX$ is

$$f_{\mathrm{L}}(X) = \beta_0 + \sum_{r=1}^{d} \beta_r \phi(X^r) \tag{2}$$

where $\beta_0 = \int f(X)dX$ and for $r > 0$, $\beta_r = \int f(X)\phi(X^r)dX$. The residual function $\eta(X) = f(X) - f_{\mathrm{L}}(X)$ satisfies orthogonality relations $\int \eta(X)dX =$

$\int \eta(X)\phi(X^r)dX = 0$ for $r = 1, \ldots, d$. The linear variation in f is defined to be

$$\sigma_{\mathrm{L}}^2 = \sigma_{\mathrm{L}}^2(f) = \int (f_{\mathrm{L}}(X) - \beta_0)^2 dX = \sum_{r=1}^{d} \beta_r^2, \tag{3}$$

and the nonlinear variation in f is

$$\sigma_{\mathrm{NL}}^2 = \sigma_{\mathrm{NL}}^2(f) = \int (f(X) - f_{\mathrm{L}}(X))^2 dX = \int \eta(X)^2 dX. \tag{4}$$

To streamline notation, we introduce functions ϕ_r, with $\phi_0(X) = 1$ for all $X \in [0,1]^d$ and $\phi_r(X) = \phi(X^r)$ for $1 \leq r \leq d$.

The fourth moment of ϕ appears frequently below. It is conveniently expressed in terms of the kurtosis

$$\kappa = \int_0^1 \phi(x)^4 dx - 3. \tag{5}$$

For the Gaussian example, $\kappa = 0$, and for the linear example, $\kappa = -6/5$.

2.2 Normalization

The variance of f is $\sigma^2 = \int (f(X) - \beta_0)^2 dX = \sigma_{\mathrm{L}}^2 + \sigma_{\mathrm{NL}}^2$. The quantities that interest us the most are the fractions $\sigma_{\mathrm{L}}^2/\sigma^2$ and $\sigma_{\mathrm{NL}}^2/\sigma^2$, of variation in f that is due to linear and nonlinear structure respectively. These ratios are unaffected when f is replaced by a dimensionally dependent scaling $c_d f$ where $c_d \in (0, \infty)$. We assume that, perhaps after some scaling,

$$0 < m \leq \quad \sigma^2 \quad \leq M < \infty \tag{6}$$

uniformly in d, where m and M are two constants.

We further assume that

$$0 < m \leq \quad \sigma_{\mathrm{L}}^2/\sigma_{\mathrm{NL}}^2 \quad \leq M < \infty, \tag{7}$$

ruling out cases where $\sigma_{\mathrm{L}}^2/\sigma^2$ tends to 0 or 1 as $d \to \infty$. This seems reasonable for many applications. If there is some nonlinearity in the effect of the first d_0 input variables considered, it cannot be expected to disappear when $d_1 > d_0$ input variables are considered.

We also assume that the basis function satisfies

$$\int_0^1 \phi(x)^8 dx < \infty. \tag{8}$$

This is true for both the uniform and Gaussian examples. Finally, we assume that, after scaling to induce (6),

$$\int f(X)^8 dX < M, \tag{9}$$

holds uniformly in d. In many practical examples f itself is bounded, making (9) hold. Condition (9) fails for a function such as $\prod_{r=1}^{d} \phi_r(X)$ which has a large number of equally important variables, and an anova decomposition in which only the order d interaction is nonzero. This condition is used below to arrange terms in some expansions according to their asymptotic size. When the condition fails, it could affect the relative importance of some terms. For such cases different asymptotic expansions should be used.

The quantity

$$S^2 = S^2(X) = \sum_{r=1}^{d} \phi_r(X)^2 \tag{10}$$

appears frequently below. It has mean d and variance $(2+\kappa)d$, so that $V(S^2) = 2d$ in the Gaussian example, and $V(S^2) = 4d/5$ in the uniform example.

2.3 Approximations and limits

This paper makes use asymptotes as the sample size n and dimension d, tend to infinity separately or together. We employ a form of big-O notation that is stricter in d than in n. Given two quantities $a(n,d)$ and $b(n,d)$, we say that

$$a(n,d) = O(b(n,d))$$

if there exist constants $1 \le n_0 < \infty$ and $C < \infty$ such that

$$|a(n,d)| \le C b(n,d), \quad 1 \le d < \infty, \quad n_0 \le n < \infty.$$

The central limit theorem implies a concentration of measure result in which $P(1-\epsilon < S^2(X)/d < 1+\epsilon) \to 0$ as $d \to \infty$, for any $\epsilon > 0$. Thus while S^2 ranges from 0 to $d \sup_x \phi^2(x)$, it is almost equal to d on all but a vanishingly small part of $[0,1]^d$, when d is large.

Let $g(X)$ be a function defined on $[0,1]^d$ with $\int g(x)^2 dx < \infty$. Write $\mu_g = \int g(X) dX$ and $\sigma_g^2 = \int (g(X) - \mu_g)^2 dX$, and let $\rho = \rho_{S^2,g}$ be the correlation between $S^2(X)$ and $g(X)$ for $X \sim U[0,1]^d$. Then

$$\int g(X) S^2(X) dX = d \int g(X) dX + d^{1/2} \rho \sigma_g (\kappa+2)^{1/2} \tag{11}$$

Because $|\rho| \le 1$, we have

$$\left| d^{-1} \int g(X) S^2(X) dX - \mu_g \right| \le d^{-1/2} \sigma_g (\kappa + 2)^{1/2}, \tag{12}$$

so that

$$\int g(X) S^2(X) dX = d\mu_g + O(d^{1/2}),$$

when $\sigma_g^2 < \infty$.

We assume throughout that $\int f(X)^8 dX < \infty$ and that $\int_0^1 \phi(x)^8 dx < \infty$, allowing a Cauchy-Shwarz inequality to be applied to $E(f^4 S^4)$. The following moment approximations are used below,

$$\begin{aligned} E(f^2 S^2) &= dE(f^2) + O(d^{1/2}) & (13) \\ E(f^4 S^4) &= O(d^2). & (14) \end{aligned}$$

Equation (14) can also be obtained if either f or ϕ is bounded and the other has a finite fourth moment.

3 Regression and quasi-regression

The approach used here for assessing linearity uses n independent points $X_1, \ldots, X_n \sim U[0,1]^d$. A linear approximation

$$f_{\mathrm{L}}^*(X) = \beta_0^* + \sum_{r=1}^{d} \beta_r^* \phi_r(X^r)$$

is fit to f and then the linear variability is estimated by

$$\sigma_{\mathrm{L}}^{*2} = \sum_{r=1}^{d} \beta_r^{*2}.$$

The nonlinear variability is estimated by subtracting the estimated linear variability from $\hat{\sigma}^2 = 1/n \sum_{i=1}^{n} (f(X_i) - \bar{f})^2$.

Although we are primarily interested in the ratio $\sigma_{\mathrm{L}}^2/\sigma^2$, we focus on the accuracy of estimates for σ_{L}^2. The reason is that σ^2 is ordinarily easier to estimate than σ_{L}^2. Furthermore when two or more competing estimates of σ_{L}^2 are under consideration, they may be used with the same estimate of σ^2.

3.1 Definitions

We consider two estimators, the usual linear regression estimator $\hat{f}_{\mathrm{L}}$ and a quasi-linear regression estimator $\widetilde{f}_{\mathrm{L}}$ that avoids solving a least squares equation. The term "quasi-linear regression" or simply "quasi-regression" is adopted because the same idea when used for fast approximate interpolation is called quasi-interpolation (Chui & Diamond 1987).

For the regression methods, let $\mathcal{X}$ be the n by $d+1$ matrix with i'th row equal to $(1, \phi_1(X_i), \ldots, \phi_d(X_i))$ and $\mathcal{Y}$ be the column vector of length n with i'th entry $Y_i = f(X_i)$. These are related by

$$\mathcal{Y} = \mathcal{X}\beta + \mathcal{E},$$

where $\mathcal{E}$ is the vector with i'th entry $\eta(X_i)$ and $\beta = (\beta_0, \beta_1, \ldots, \beta_d)'$.

The true value β can be written

$$\begin{aligned} \beta &= \left[E(\mathcal{X}'\mathcal{X})\right]^{-1} E(\mathcal{X}'\mathcal{Y}) && (15)\\ &= E(\mathcal{X}'\mathcal{Y}), && (16) \end{aligned}$$

with (16) following by orthogonality of $\phi_r(X)$. The usual least squares estimate

$$\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_d)' = (\mathcal{X}'\mathcal{X})^{-1}\mathcal{X}'\mathcal{Y}, \tag{17}$$

is a sample analogue of equation (15). The quasi-regression estimate is

$$\widetilde{\beta} = (\widetilde{\beta}_0, \widetilde{\beta}_1, \ldots, \widetilde{\beta}_d)' = \frac{1}{n}\mathcal{X}'\mathcal{Y}, \tag{18}$$

a sample analogue of equation (16).

What we call quasi-regression, has been studied by Efromovich (1992) for orthogonal series smoothing on the unit interval. Owen (1992) proposes it for regression modeling of data from Latin hypercube samples, Koehler & Owen (1996) propose it for computer experiments, and An & Owen (1999) implement it for some computer experiments. The quasi-regression problem treated here is simpler than the general case. Not only are the basis functions $\phi_r(X)$ uncorrelated, they are also independent.

3.2 Statistical efficiency

To analyze the estimates $\hat{\beta}$ and $\widetilde{\beta}$, let

$$\frac{1}{n}\mathcal{X}'\mathcal{X} = I + A, \quad \frac{1}{n}\mathcal{X}'\mathcal{E} = \delta.$$

The matrix A and the vector δ both have mean zero and elements with variance proportional to $1/n$. We use indices running from 0 through d for the entries in A and δ, so that the subscripts correspond naturally with those of β.

The error in quasi-regression is

$$\begin{aligned}\widetilde{\beta} - \beta &= n^{-1}\mathcal{X}'\mathcal{Y} - \beta \\ &= n^{-1}(\mathcal{X}'\mathcal{X}\beta + \mathcal{X}'\mathcal{E}) - \beta \\ &= \delta + A\beta,\end{aligned}$$

and the error in least squares regression is

$$\begin{aligned}\hat{\beta} - \beta &= (\mathcal{X}'\mathcal{X})^{-1}\mathcal{X}'\mathcal{Y} - \beta \\ &= (\mathcal{X}'\mathcal{X})^{-1}\mathcal{X}'\mathcal{E} \\ &= (I + A)^{-1}\delta.\end{aligned}$$

Write $I + A = \sum_{k=1}^{d} \lambda_k e_k e_k'$, where λ_k and e_k are eigenvalue-eigenvector pairs. The matrix A has the same eigenvectors e_k with corresponding eigenvalues $\nu_k = \lambda_k - 1$. The regression error $(I + A)^{-1}\delta$ can be written

$$\hat{\beta} - \beta = (I - A + A^2 - A^3 \cdots)\delta \tag{19}$$

provided that $\max_{1 \le k \le d} |\nu_k| < 1$. If also $\max_{1 \le k \le d} |\nu_k| \to 0$ as $n \to \infty$, then the expansion is dominated by early terms allowing an approximation of the form

$$\hat{\beta} - \beta \doteq \delta - A\delta, \tag{20}$$

or even $\hat{\beta} - \beta \doteq \delta$.

For the rest of this subsection, consider the limit with d fixed and $n \to \infty$. In this case, all d eigenvalues of $I + A$ converge to 1 with a $\log(n)/\sqrt{n}$ rate (Anderson 1984). It follows that the eigenvalues of A converge to 0 at that rate, justifying (19) and (20).

The error δ contributes directly to both estimators $\widetilde{\beta}$ and $\hat{\beta}$. They differ in how the error A enters: through $A\beta$ for quasi-regression and, approximately as $A\delta$ for regression. In limits with d fixed and $n \to \infty$, the coefficients of η are $O_p(n^{-1/2})$ while those of β are constant in n. Thus in such limits we can expect that regression should have better performance than quasi-regression. Because they both share the error δ, which, like $A\beta$ is $O_p(n^{-1/2})$ the rates of convergence for the methods are the same, while the constant ordinarily favors regression.

In a limit with $\sigma_{\mathrm{NL}} \to 0$ it is possible for linear regression to achieve a better rate of convergence than $n^{-1/2}$. In a limit with $\sigma_{\mathrm{L}} \to 0$ it is possible for the $A\beta$ term to be negligible compared to $A\delta$, giving quasi-regression a more favorable constant. Both of these are ruled out by the condition (7).

That regression is more statistically efficient than quasi-regression is borne out by some more detailed calculations in Sections 4 and 5. This notion of statistical efficiency is taken with respect to a fixed n. Section 3.4 shows that for a fixed computing budget, quasi-regression can be more efficient, for large d.

3.3 Regression with $d \to \infty$

Here we describe the case of regression with $d \to \infty$. We draw on the recent survey paper Bai (1999).

Apart from the first row and column (corresponding to the constant ϕ_0), the matrix $I + A$ is a large sample covariance matrix of the type recently surveyed in Bai (1999). The limit of covariance matrices with $d, n \to \infty$ and $d/n \to \tau \in (0, \infty)$ has been thoroughly studied. The Marcenko-Pastur law describes the limiting empirical distribution of $\lambda_1, \ldots, \lambda_d$. For the case where $\tau < 1$ the limiting distribution has density function

$$p_\tau(\lambda) = \frac{(b_\tau - \lambda)(\lambda - a_\tau)}{2\pi\tau\lambda}, \quad \text{on } a_\tau \le \lambda \le b_\tau, \tag{21}$$

where $a_\tau = (1 - \tau^{1/2})^2$ and $b_\tau = (1 + \tau^{1/2})^2$. Finite fourth moments of $\phi_r(X_i)$ are sufficient for this and for the sharper results that

$$\begin{aligned} \lim_{d\to\infty} \min_{1\le k\le d} \lambda_k &= (1 - \tau^{1/2})^2, \quad \text{a.s.} \\ \lim_{d\to\infty} \max_{1\le k\le d} \lambda_k &= (1 + \tau^{1/2})^2, \quad \text{a.s.}. \end{aligned}$$

It follows that in the limit as $d \to \infty$ with $d/n \to \tau < 1$, expansion (19) can be made. The terms decrease in magnitude as the power of A increases, but they are not of a smaller order in n. The approximation of $(I + A)^{-1}$ by $I - A$ can be made very accurate by taking a small value for τ, but for any fixed $\tau > 0$ the approximation does not become increasingly accurate as $n \to \infty$.

For the case with $d \to \infty$ and $d/n \to 0$, it is known that the limiting density of the eigenvalues of $(n/d)^{1/2}A$ is given by the semicircular law

$$\frac{(4 - \nu^2)^{1/2}}{2\pi}, \quad |\nu| \le 2. \tag{22}$$

This corresponds to eigenvalues of A distributed between $\pm 2(d/n)^{1/2}$ for which successive powers of A are of smaller order in n. Less is known about the rate at which extreme eigenvalues approach their limits in this case.

The qualitative conclusion from the previous section should still hold in this case. Along a given sequence $d \to \infty$ with $d/n \to 0$, regression should be more accurate than quasi-regression, because $\hat{\beta} - \beta$ is approximately $\delta - A\delta$ and $A\delta$ should be smaller than $A\beta$.

3.4 Computational efficiency

A large dimension d will require a large sample size n for either method. Consider regression using $n = \hat{n} = \hat{a}d^q$ for some $q \geq 1$ and $\hat{a} > 0$. The computational cost of regression is at least $\hat{n}d^2/2 + d^3/6$ floating point operations (Golub & Van Loan 1983). For large d, this dominates the cost of $\hat{n}$ function evaluations, under our assumption that function evaluations cost at most $O(d)$.

The cost of estimation for quasi-regression, using $n = \widetilde{n}$ is of order $\widetilde{n}d$, and therefore by setting $\widetilde{n}d = \hat{n}d^2/2$, for the same computational effort it is possible to use at least $\widetilde{n} = \hat{a}d^{1+q}/2$ function evaluations in quasi-regression. If the cost of computing f grows proportionally to d then taking this into account gives $\widetilde{n} = \widetilde{a}d^{1+q}$ for some $\widetilde{a} \leq \hat{a}/2$. The bias and variance in $\widetilde{\sigma}^2_{\mathrm{L}}$ and $\hat{\sigma}^2_{\mathrm{L}}$ are shown below to be of the same order with leading constants that are likely to favor $\hat{\sigma}^2_{\mathrm{L}}$. But for the same computational effort a much larger sample size is possible for $\widetilde{\sigma}^2_{\mathrm{L}}$.

It follows that as dimension increases, quasi-regression will eventually be more accurate than regression, when given the same amount of computer time. Quasi-regression also requires only $O(d)$ storage compared to $O(d^2)$ for regression.

4 Accuracy of quasi-regression

This section considers the mean and variance of $\widetilde{\sigma}^2_L$ in quasi-regression. An exact expression is presented. Then it is simplified with respect to n, then simplified with respect to d.

Proposition 4.1 *The quasi-regression based estimator* $\widetilde{\sigma}^2_{\mathrm{L}} = \sum_{r=1}^{d} \widetilde{\beta}^2_r$ *has expectation*

$$E(\widetilde{\sigma}^2_{\mathrm{L}}) = \frac{n-1}{n}\sigma^2_{\mathrm{L}} + \frac{1}{n}E\left(S(X)^2 f(X)^2\right). \tag{23}$$

Proof: This follows immediately from short moment calculations. □

Variance in $\widetilde{\beta}_r$ induces a positive bias in the estimate $\widetilde{\sigma}^2_{\mathrm{L}}$. This bias is of order d/n:

$$E(\widetilde{\sigma}^2_{\mathrm{L}}) = \sigma^2_{\mathrm{L}} + \frac{d}{n} E\left(f(X)^2\right) + O\left(\frac{d^{1/2}}{n}\right)$$

with the error bound coming from equation (13). Next we consider the variance of $\widetilde{\sigma}^2_{\mathrm{L}}$.

Theorem 4.1 *For the quasi-regression estimator,*

$$\begin{aligned} V(\widetilde{\sigma}^2_{\mathrm{L}}) &= \frac{(n-1)(6-4n)}{n^3}\sigma^4_{\mathrm{L}} + 4\frac{(n-1)(n-2)}{n^3}E((f_{\mathrm{L}}-\beta_0)^2 f^2) \\ &\quad -4\frac{n-1}{n^3}\sigma^2_{\mathrm{L}}E(S^2f^2) + 2\frac{n-1}{n^3}\sum_r\sum_s E(\phi_r\phi_s f^2)^2 \\ &\quad -\frac{1}{n^3}E(S^2f^2)^2 + 4\frac{n-1}{n^3}E(S^2(f_{\mathrm{L}}-\beta_0)f^3) + \frac{1}{n^3}E(S^4f^4). \end{aligned} \tag{24}$$

Proof: The proof requires tedious but straightforward algebra. Details are available from the author. □

Equation (24) in Theorem 4.1 holds for any $n \geq 1$ and $d \geq 1$. In particular it does not require $n > d$. After simplifying and re-arranging terms we obtain

$$\begin{aligned} V(\widetilde{\sigma}^2_{\mathrm{L}}) &= \frac{4}{n}\Big(E((f_{\mathrm{L}}-\beta_0)^2f^2) - \sigma^4_{\mathrm{L}}\Big) + \frac{2}{n^2}\sum_r\sum_s E(\phi_r\phi_s f^2)^2 \\ &\quad +\frac{4}{n^2}\Big(E(S^2(f_{\mathrm{L}}-\beta_0)f^3) - \sigma^2_{\mathrm{L}}E(S^2f^2)\Big) \\ &\quad +\frac{1}{n^3}\Big(E(S^4f^4) - E(S^2f^2)^2\Big) + O\left(\frac{1}{n}\right). \end{aligned}$$

For large d, concentration of measure gives that S^2 is nearly d and this leads to the further simplification

$$\begin{aligned} V(\widetilde{\sigma}^2_{\mathrm{L}}) &= \frac{4}{n}\Big(E((f_{\mathrm{L}}-\beta_0)^2f^2) - \sigma^4_{\mathrm{L}}\Big) + \frac{2}{n^2}\sum_r\sum_s E(\phi_r\phi_s f^2)^2 \\ &\quad +\frac{4d}{n^2}\Big(E((f_{\mathrm{L}}-\beta_0)f^3) - \sigma^2_{\mathrm{L}}E(f^2)\Big) \\ &\quad +\frac{d^2}{n^3}\Big(E(f^4) - E(f^2)^2\Big) + O\left(\frac{d^2}{n^3}\right). \end{aligned} \tag{25}$$

The first, third and fourth terms in (25) are of orders $1/n$, d/n^2 and d^2/n^3 respectively. The dominant error is from applying concentration of measure to the fourth term, using (14), and it is of the same order of magnitude as that term.

The second term, while apparently of order d^2/n^2 is in fact only of order d/n^2 because the sum of the off diagonal terms $r \neq s$ cannot be larger than $2E(f^2)^2$. To see this note that

$$\sum_{r=1}^{d}\sum_{s=0}^{r-1} E(\phi_r\phi_s f^2)^2 \leq E(f^2)^2,$$

by projecting f^2 on the $d(d+1)/2$ mutually orthogonal basis functions $\phi_r(X^r)\phi_s(X^s)$. The $r = s$ terms contribute $O(d/n^2)$ which dominates the $O(1/n^2)$ contribution from the $r \neq s$ terms.

Thus quasi-regression produces an estimate $\widetilde{\sigma}^2_{\mathrm{L}}$ with a squared bias of order d^2/n^2 and a variance of order $1/n + d/n^2 + d^2/n^3$.

Consider the simple limit where $n \to \infty$ with d fixed, followed by $d \to \infty$. In this limit the mean squared error is dominated by the first term in the variance (25), and d does not appear:

$$\lim_{n\to\infty} nE\Big((\widetilde{\sigma}^2_{\mathrm{L}} - \sigma^2_{\mathrm{L}})^2\Big) = 4\Big(E((f_{\mathrm{L}} - \beta_0)^2 f^2) - \sigma^4_{\mathrm{L}}\Big).$$

The natural sample size $\widetilde{n}$ for quasi-regression is proportional to d. Taking $\widetilde{n} \doteq \alpha d^q$ for $q > 1$ has the unreasonable consequence that accuracy improves as the dimension of the problem increases and taking $q < 1$ implies that as d increases the method eventually fails.

Using the natural sample size $n \doteq \alpha d$, for $\alpha > 0$ and taking the limit as $d \to \infty$, we find

$$\lim_{\substack{d\to\infty \\ n=\lfloor \alpha d\rfloor}} E\Big((\widetilde{\sigma}^2_{\mathrm{L}} - \sigma^2_{\mathrm{L}})^2\Big) = \frac{E(f(X)^2)}{\alpha^2} + O\left(\frac{1}{n^{1/2}}\right),$$

where the error term contains a factor proportional to $\alpha^{-1/2}$. The resulting error is dominated by bias. This motivates bias corrected versions of quasi-regression considered next.

4.1 Bias corrected quasi-regression

Unless n is of order d^2 or larger, the bias dominates the error in $\widetilde{\sigma}^2_{\mathrm{L}}$. Since we have an expression for the bias, we consider bias corrected quasi-regression

estimators

$$\widetilde{\sigma}^2_{\mathrm{L},BC} = \frac{n}{n-1}\Big(\widetilde{\sigma}^2_{\mathrm{L}} - \frac{1}{n^2}\sum_{i=1}^{n} S(X_i)^2 f(X_i)^2\Big) \tag{26}$$

and

$$\widetilde{\sigma}^2_{\mathrm{L},BCCM} = \frac{n}{n-1}\Big(\widetilde{\sigma}^2_{\mathrm{L}} - \frac{d}{n^2}\sum_{i=1}^{n} f(X_i)^2\Big). \tag{27}$$

The estimate (26) is unbiased by Proposition 4.1 but requires extra computation on the order of nd to calculate the sample average of $S^2 f^2$. This is the same order of computation required to construct $\widetilde{\sigma}^2_{\mathrm{L}}$ itself. The estimate (27) is derived by employing the concentration of measure approximation, replacing every $S(X_i)^2$ by d. It only requires extra computation on the order of n, instead of nd, and hence may be preferable for large d.

The variance of the bias correction in (26) is $n^{-3}(E(S^4 f^4) - E(S^2 f^2)^2) = O(d^2/n^3)$, using (14). The variance already has a term of this order, and so the bias correction does not increase the asymptotic order of the variance. Because the variance of $\widetilde{\sigma}^2_{\mathrm{L}}$ is of smaller order than the original squared bias, bias correction should provide an asymptotic improvement.

For bias corrected quasi-regression estimates of σ^2_{L}, it is possible to have $n < d$ as $d \to \infty$. The natural sample size for bias corrected quasi-regression grows proportionally to $d^{2/3}$. If n grows faster than $d^{2/3}$ one gets better answers in higher dimensions than in lower ones, and if n grows slower than $d^{2/3}$ the method fails for large enough d. With $n \doteq a d^{2/3}$ the bias of $\widetilde{\sigma}^2_{\mathrm{L},BC}$ is zero and the variance tends to zero as $a \to \infty$.

5 Accuracy of regression

The regression method requires $n > d$. In limits with $n \to \infty$ and $d/n \to \tau < 1$, the extreme eigenvalues of $\mathcal{X}'\mathcal{X}$ are well behaved, and $\hat{\beta}$ exists. In this section, we develop expressions for the bias and varaince of $\hat{\sigma}^2_{\mathrm{L}}$ for limits with $n \to \infty$ and d fixed. These turn out to be similar in form to the expressions for quasi-regression, allowing a comparison of the two methods, although the formulas for quasi-regression hold for any n and d.

We consider approximations formed by truncating

$$\hat{\beta} - \beta = \delta - A\delta + A^2\delta - A^3\delta \ldots, \tag{28}$$

with $\hat{\beta} - \beta \doteq \delta$ being the first order approximation and $\hat{\beta} - \beta \doteq \delta - A\delta$ being the second order one. As is customary in delta method calculations, a first order approximation suffices for variance calculations, but a second order

one is required for bias calculations, because $E(\delta) = 0$. The approximations developed here may not apply to cases with $d \to \infty$ and $d/n \to 0$ because the correct number of terms to consider in such cases may not be the same as in the usual delta method. The delta method is discussed by Efron & Tibshirani (1993).

Proposition 5.1 *The r'th component of* $E(\delta - A\delta)$ *is* $-\frac{1}{n}E(\phi_r S^2 \eta)$.

Proof: The proof involves straightforward moment calculations. It is available from the author. □

Using concentration of measure

$$E((\delta - A\delta)_r) - \beta_r = -(d/n)E(\phi_r \eta) + O\left(\frac{d^{1/2}}{n}\right) = O\left(\frac{1}{n}\right).$$

Thus while quasi-regression gives unbiased estimates $\widetilde{\beta}$, regression estimates have a slight bias. For a first order delta approximation

$$V(\hat{\delta}_r) = \frac{1}{n}E(\phi_r^2 \eta^2).$$

This results in

$$\begin{aligned} E\left(\sum_{r=1}^{d} \hat{\beta}_r^2\right) &= \sigma_{\mathrm{L}}^2 + \frac{1}{n}E\left(S(X)^2\eta^2\right) + O(n^{-2}) \\ &= \sigma_{\mathrm{L}}^2 + \frac{d}{n}\sigma_{\mathrm{NL}}^2 + O(n^{-1}). \end{aligned}$$

The lead term in the bias of $\hat{\sigma}_{\mathrm{L}}^2$ is $(d/n)E(\eta^2)$ compared to $(d/n)E(f^2)$ under quasi-regression. Thus the regression estimate typically has smaller bias and can have much smaller bias for nearly linear functions f.

For regression,

$$V(\hat{\sigma}_{\mathrm{L}}^2) = \sum_{r=1}^{d}\sum_{s=1}^{d}\Big(E(\hat{\beta}_r^2\hat{\beta}_s^2) - E(\hat{\beta}_r^2)E(\hat{\beta}_s^2)\Big), \tag{29}$$

and,

Lemma 5.1

$$\begin{aligned} E(\hat{\beta}_r^2) &= \beta_r^2 - \frac{2\beta_r}{n}E(S^2\phi_r^2\eta^2) + \frac{1}{n}E(\phi_r^2\eta^2) + o\left(\frac{1}{n}\right), \\ E(\hat{\beta}_r^2\hat{\beta}_s^2) - E(\hat{\beta}_r^2)E(\hat{\beta}_s^2) &= \frac{4\beta_r\beta_s}{n}E(\phi_r\phi_s\eta^2) + o\left(\frac{1}{n}\right). \end{aligned}$$

Proof: In the series

$$\hat{\beta}_r = \beta_r + \delta_r - (A\delta)_r + (A^2\delta)_r - \cdots$$

the consecutive terms are of order $n^{-k/2}$ in probability, for integer $k \geq 0$. The proof requires straightforward but lengthy expansion and evaluation of terms. Details of the proof are available from the author. □

Theorem 5.1

$$V(\hat{\sigma}_{\mathrm{L}}^2) \;=\; \frac{4}{n} E\Big((f_{\mathrm{L}} - \beta_0)^2 \eta^2\Big) + o\left(\frac{1}{n}\right).$$

Proof: This follows directly from Lemma 5.1. □

To compare the variance of regression and quasi-regression, write

$$\begin{aligned} \lim_{n\to\infty} nV(\hat{\sigma}_{\mathrm{L}}^2) &= 4E\Big((f_{\mathrm{L}} - \beta_0)^2 \eta^2\Big), \\ \lim_{n\to\infty} nV(\widetilde{\sigma}_{\mathrm{L}}^2) &= 4E\Big((f_{\mathrm{L}} - \beta_0)^2 (f^2 - \sigma_{\mathrm{L}}^2)\Big). \end{aligned}$$

For large n, the variance comparison becomes one of comparing a weighted mean of η^2 to a similarly weighted mean of $f^2 - \sigma_{\mathrm{L}}^2$. Expanding $f = \beta_0 + (f_{\mathrm{L}} - \beta_0) + \eta$ we find that f^2 contains a term η^2, a term $(f_{\mathrm{L}} - \beta_0)^2$ that more than compensates for the $-\sigma_{\mathrm{L}}^2$, a nonnegative and possibly large term β_0^2, as well as some cross terms. As in Section 3.2 we conclude that the accuracy of $\hat{\sigma}_{\mathrm{L}}^2$ should generally be superior to that of $\widetilde{\sigma}_{\mathrm{L}}^2$ on the same sample size n, though for some f the cross terms may be large enough and of right sign to make $V(\widetilde{\sigma}_{\mathrm{L}}^2) < V(\hat{\sigma}_{\mathrm{L}}^2)$.

6 Adjusted quasi-regression

The regression estimator $\hat{\sigma}_{\mathrm{L}}^2$ has a bias and variance that both depend only on $\eta(X)$ and not on β. The quasi-regression estimator is not invariant in this way. The performance of the quasi-regression estimator even depends on the constant β_0; adding a constant to f can change the accuracy of $\widetilde{\sigma}_{\mathrm{L}}^2$.

This suggests the use of

$$\widetilde{\beta}_r^{(c)} = \frac{1}{n} \sum_{i=1}^{n} \phi_r(X_i)(f(X_i) - c)$$

with some constant c for $r \geq 1$. A natural choice for c is β_0, resulting in the estimate $\widetilde{\sigma}^{(\beta_0)^2} = \sum_{r=1}^{d} \widetilde{\beta}_r^{(\beta_0)^2}$ with

$$E(\widetilde{\sigma}_{\mathrm{L}}^{(\beta_0)^2}) = \sigma_{\mathrm{L}}^2 + \frac{d\sigma^2}{n} + O\left(\frac{d^{1/2}}{n}\right). \tag{30}$$

and

$$\lim_{n\to\infty} nV(\widetilde{\sigma}_{\mathrm{L}}^{(\beta_0)^2}) = 4E\Big((f_{\mathrm{L}} - \beta_0)^2((f-\beta_0)^2 - \sigma_{\mathrm{L}}^2)\Big). \tag{31}$$

In practice one would usually have to estimate β_0 by $\bar{f} = 1/n\sum_{i=1}^{n} f(X_i)$, so that the estimate of β_r becomes

$$\widetilde{\beta}_{r\cdot 0} = \frac{1}{n}\sum_{i=1}^{n} \phi_r(X_i)(f(X_i) - \bar{f}). \tag{32}$$

This can be done in one pass over the data in a numerically stable way, as described in Chan, Golub & LeVeque (1983). Using an estimated value of β_0 will cause a small increase in the variance of the estimate of σ_{L}^2.

More generally one can use least squares regression for $1 + d_0$ of the components of β and quasi-regression on the others. Efromovich (1992) studies this approach for orthogonal series expansions over $[0, 1]$.

7 Numerical results

7.1 Dimension one million

This subsection considers a single example in the $d = 10^6$ dimensional unit cube. Linear regression requires $n \geq d$. The linear algebra for this problem would therefore require $nd^2/2 \geq 5 \times 10^{17}$ floating point evaluations and storage for $d^2/2 = 5 \times 10^{11}$ floating point numbers.

The function f was generated randomly with $\beta_r \sim U[0, (3/d)^{1/2}]$, independently, for $1 \leq r \leq d$, $\beta_0 = 0$ and $\eta(X) = 10^{-1/2}(\phi_1(X)^2 - 1.0)$. For $r \geq 1$, the uniform ϕ_r were used. The function had $\sigma_{\mathrm{L}}^2 = 0.999549$, varying only slightly from the expected value 1, because d is so large. The true value of σ_{NL}^2 is 0.08.

The quasi-regression estimate using $n = 10^5$ was $\widetilde{\sigma}_{\mathrm{L}}^2 = 11.80$. The theory in Section 4 predicts a bias of $(d/n)E(f(X)^2) = (10^6/10^5)(1.08) \doteq 10.8$ closely matching the observed bias. Bias corrected quasi-regression gives $\widetilde{\sigma}_{\mathrm{L},BC}^2 = 0.985429$ with which the concentration of measure version $\widetilde{\sigma}_{\mathrm{L},BCCM}^2 = 0.985433$, agrees to five significant digits.

The true proportion of linearity in this function is $1/1.08 \approx 0.926$. The sample estimate of variance in this example is 1.0817. The two bias corrected quasi-regression methods give a very accurate estimate $.9854/1.0817 \approx 0.911$ of this proportion. They also give a good estimate $1.0817 - .9854 = .0963$ of the difference $\sigma^2_{\mathrm{NL}} = .08$.

This function requires roughly 50 hours of cpu time on an R10000 processor running at 196Mhz. Most of that time is spent generating random numbers and applying the ϕ function $\sqrt{12}(X^r - 0.5)$. But roughly 5.7 hours is spent looping over the data computing $\widetilde{\beta}_r$ and the estimate of bias. If we suppose conservatively that just one hour was spent on the nd floating point operations to form $\widetilde{\beta}_r$, then to do the $d^3/2 + d^3/6$ calculations required for the smallest possible least squares regression, having $n = d + 1$ should take about $(2^{-1} + 6^{-1})d^3/(nd) \approx 6.7 \times 10^6$ hours, or over 760 years.

This example establishes that quasi-regression can get useful answers in some large problems where regression is infeasible.

7.2 Experiment

The function f in the previous subsection may have been special: The d dimensions were all roughly equally important, the function η only depended on X^1, it had $\sigma^2_{\mathrm{L}}/\sigma^2 \approx 0.926$ which might possibly be much easier than other ratios, it had $\beta_0 = 0$, which is likely to have made the problem easier, and it was only run once, because of its expense. This section explores the effects of the issues raised above, using a small designed experiment.

The functions ϕ_r for $r > 0$ were taken to be either Gaussian or uniform. Two different intercepts β_0 are used: 0 and 4. Two different functions $f_{\mathrm{L}} - \beta_0$ are used: one has $\beta_r \propto 2^{-r/2}$ for $r \geq 1$ so that the importance of the variables drops off quickly. The other has $\beta_r \propto r^{-1/2}$, so that many variables are important. Both are scaled so that $\sigma^2_{\mathrm{L}} = \sum_{r=1}^{d} \beta_r^2 = 1$. The noise function η can depend on one of the X^r or on roughly half of them: $\eta(X) \propto \phi_1^2(X) - 1.0$ or $\eta(X) \propto \sum_{1 \leq r \leq \lfloor d/2 \rfloor} \prod_{s=0}^{2} \phi_{r+s}(X)$. This function η is orthogonal to all of the ϕ_r when $d \geq 5$. The constant of proportionality in η was taken to obtain either $R^2 = \sigma^2_{\mathrm{L}}/\sigma^2 = 0.5$ or 0.95. In the functions above the important variables for f_{L} tend to be the same as those for η. This might make the functions atypical. Therefore, a direction reversal sending $\beta_r \to \beta_{d+1-r}$ for $r > 0$ was also considered. The factors in this experiment are set out in Table 1. All together 64 different functions were considered, and four replicated data sets were generated for each of the 64 functions.

For each function the following estimates of σ^2_{L} were computed: $\widetilde{\sigma}^2_{\mathrm{L}}$,

Factor	Level 0	Level 1	Label
$\phi_r, r > 0$	Uniform	Gaussian	G
Intercept β_0	0.0	4.0	B
β_r^2 decay	Linear	Exponential	E
$\eta(X)$	1 Dimensional	$\lfloor d/2 \rfloor + 2$ Dimensional	D
$R^2 = \sigma_{\mathrm{L}}^2/\sigma^2$	0.50	0.95	R
β_r Direction	Decreasing	Increasing	I

Table 1: Factors in experiment: 64 different functions $f(X)$ were considered in the experiment. The table lists the six binary factors varied, gives their levels and a mnemonic label for each.

$\widetilde{\sigma}^2_{\mathrm{L},BC}$, $\widetilde{\sigma}^2_{\mathrm{L},BCCM}$, a centered estimate

$$\widetilde{\sigma}^2_{\mathrm{L},C} = \sum_{r=1}^{d} \widetilde{\beta}^2_{r\cdot 0}$$

where $\widetilde{\beta}_{r\cdot 0}$ is given by equation (32), and a centered estimate with an approximate bias correction derived from the concentration of measure

$$\widetilde{\sigma}^2_{\mathrm{L},CCM} = \widetilde{\sigma}^2_{\mathrm{L},C} - \frac{d}{n^2}\sum_{i=1}^{n}(f(X_i) - \bar{f})^2$$

where $\bar{f} = (1/n)\sum_{i=1}^{n} f(X_i)$. The factor $n/(n-1)$ in $\widetilde{\sigma}^2_{\mathrm{L},BC}$, and $\widetilde{\sigma}^2_{\mathrm{L},BCCM}$ was not used, but the effect is negligible.

The experiment was done with $n = 10,000$ and $d = 1,000$. The results are presented in Figure 1. The top histogram shows the estimates from $\widetilde{\sigma}^2_{\mathrm{L}}$. It is clearly bimodal. The rightmost mode consists of those cases with $\beta_0 = 4$. As can be expected from the theory, a moderately large non-zero intercept can inflate the bias in $\widetilde{\sigma}^2_{\mathrm{L}}$.

The other four methods all reduce the bias from $\beta_0 \neq 0$, and all perform better than the original quasi-regression. The simple mean adjustment method $\widetilde{\sigma}^2_{\mathrm{L},C}$ still shows some bias as predicted by equation (30). The concentration of measure approximation is very accurate here, with the relation $\widetilde{\sigma}^2_{\mathrm{L},BC} \approx 0.0002 + 0.9984\widetilde{\sigma}^2_{\mathrm{L},BCCM}$ explaining 99.987% of the variation in $\widetilde{\sigma}^2_{\mathrm{L},BC}$ over the 256 generated functions. These bias corrections do not reduce that variance, and so the histograms for these methods is roughly a superposition of the two histograms from the top histogram, after shifting to remove bias. The mean adjusted quasi-regression estimate $\widetilde{\sigma}^2_{\mathrm{L},CCM}$, at the bottom of Figure 1 reduces both variance and bias compared to the original quasi-regression.

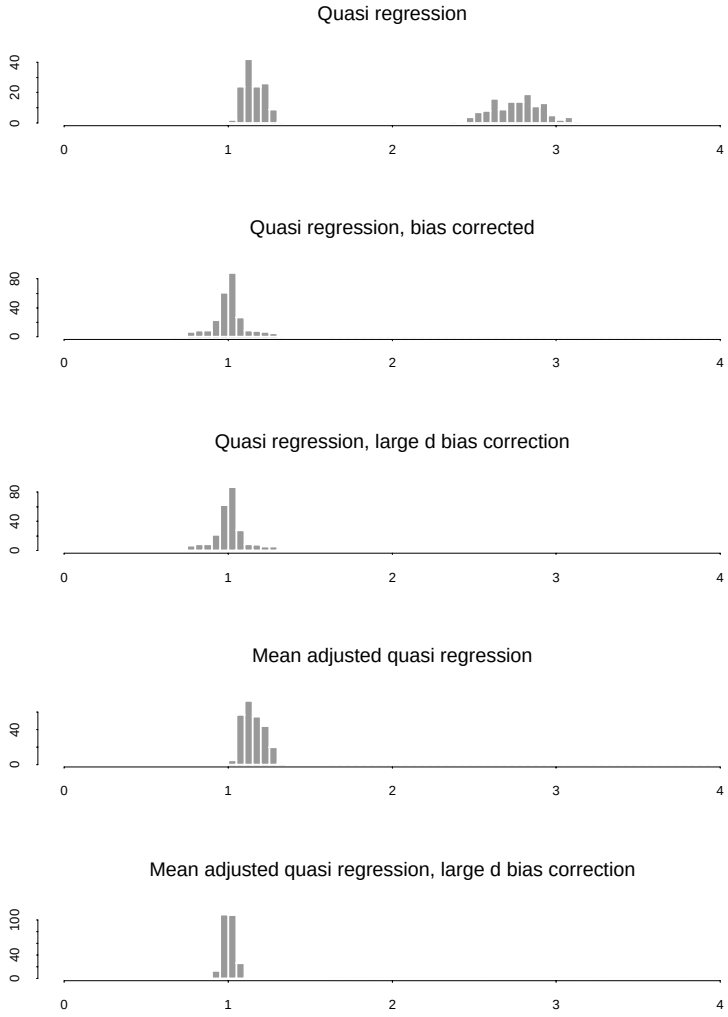

Figure 1: Experimental output. Shown are histograms of the estimated σ^2_{L} for the five methods discussed in the text, with $n = 10000$ and $d = 1000$. The points in the histogram come from 64 different functions, each of which was tested with 4 data sets. The true σ^2_{L} is always 1.0 for each function.

Figure 2 shows boxplots of the bias corrected and centered estimates $\widetilde{\sigma}^2_{\mathrm{L},CCM}$, broken down by the values of the 6 experimental factors. The only factor that affects the performance of this estimate is R, which affects the variance. A similar plot shows that R is the only important factor for $\widetilde{\sigma}^2_{\mathrm{L},C}$

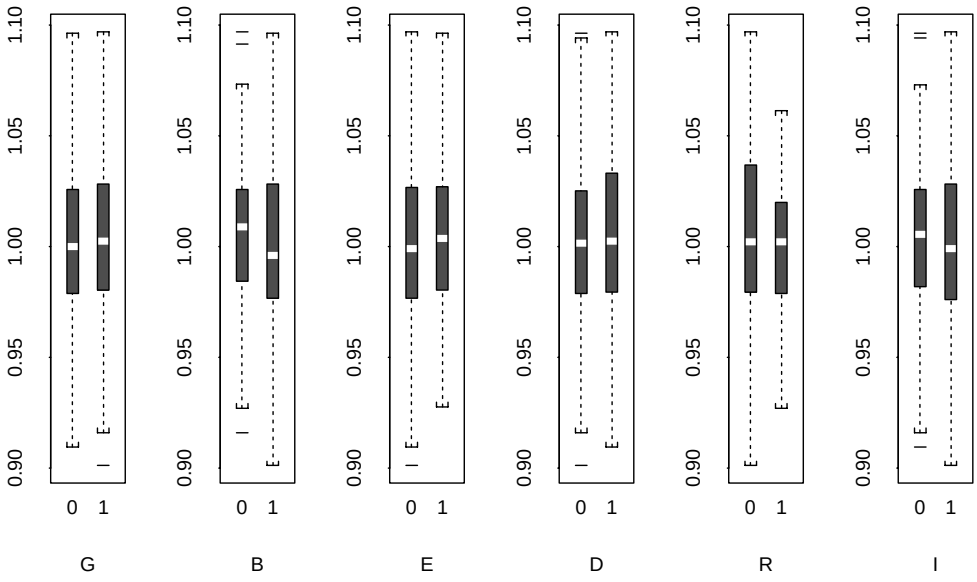

Figure 2: Shown are boxplots of the centered estimate $\widetilde{\sigma}^2_{\mathrm{L},CCM}$ of σ^2_{L}. For each of the 6 experimental factors in Table 1, there is a boxplot for each of the two levels of that factor. This experiment had $n = 10000$ and $d = 1000$. The true σ^2_{L} is always 1.0 for each function. The estimate is largely unaffected by the factors varied in this experiment, except that the factor R influences the variance.

where a large value $R^2 = .95$ reduces the bias. The only important factor for $\widetilde{\sigma}^2_{\mathrm{L},BC}$ and $\widetilde{\sigma}^2_{\mathrm{L},BCCM}$ is B, for which large intercept values yield greater variance.

Similar results are seen when the experiment is run at other values of n and d. The performance of the bias corrected estimates improves as $n/d^{2/3}$ increases, while that of the uncorrected estimates depends more on n/d.

This experiment indicates that the factors studied do not strongly affect the performance of quasi-regression estimators, and so the previous experimental situation with $d = 1,000,000$ was not misleadingly simple, except that having intercept 0 was advantageous because no centering was done.

8 Conclusions

Quasi-regression methods are able to estimate the amount of linear variation in high dimensional functions with much less computation than linear regression methods require. With bias corrected versions, it is possible to use $n \ll d$ observations in d dimensions when d is large. The natural sample size n should grow as a multiple of $d^{2/3}$ for these bias corrected estimates. This is less computation than would be required to take d divided difference estimates of the gradient of f at a single point.

The example in Section 7.1 shows that even in a $1,000,000$ dimensional space, a good estimate of σ^2_{L} can be obtained in a reasonable amount of computation. The experiment in Section 7.2 shows that performance can be improved by making an adjustment for the mean of f. The mean of f is simply the coefficient of $\phi_0(X)$ in f_{L}, and it is reasonable to expect that if any $\phi_r(X)\beta_r$ dominates f_{L} that adjusting for it, as discussed in Section 6 could further improve performance.

The experiment in Section 7.2 shows that a bias corrected version of the mean adjusted estimate works well for a set of 64 functions that vary according to the 6 binary factors in Table 1. These factors vary aspects of f that may have been especially favorable to the example in Section 7.1. None of them has a very large effect on the performance of the methods, except that a nonzero intercept can strongly bias the original quasi-regression estimate $\widetilde{\sigma}^2_{\mathrm{L}}$.

An & Owen (1999) use quasi-regression to approximate functions as expansions over large tensor product bases. The bias corrections discussed here may apply to estimates of the importance of various subsets of coefficients there.

Acknowledgements

I thank Persi Diaconis, Zhidong Bai and Jack Silverstein for helpful comments on the eigenvalues of large random covariance matrices. Thanks also to the reviewers for helpful comments on the presentation.

References

An, J. & Owen, A. B. (1999), Quasi-regression, Technical report, Stanford University, Statistics Department.

Anderson, T. W. (1984), *An Introduction to Multivariate Statistical Analysis (Second Edition)*, Wiley.

Bai, Z. (1999), 'Methodologies in spectral analysis of large dimensional random matrices, a review (with discussion)', *Statistica Sinica* **9**(3), 611–678.

Caflisch, R. E., Morokoff, W. & Owen, A. B. (1997), 'Valuation of mortgage backed securities using brownian bridges to reduce effective dimension', *Journal of Computational Finance* **1**, 27–46.

Chan, T. F., Golub, G. H. & LeVeque, R. J. (1983), 'Algorithms for computing the sample variance: Analysis and recommendations', *The American Statistician* **37**, 242–247.

Chui, C. K. & Diamond, H. (1987), 'A natural formulation of quasi-interpolation by multivariate splines', *Proceedings of the American Mathematical Society* **99**(4), 643–646.

Efromovich, S. (1992), On orthogonal series estimators for random design nonparametric regression, *in* 'Computing Science and Statistics. Proceedings of the 24rd Symposium on the Interface', pp. 375– 379.

Efron, B. M. & Tibshirani, R. J. (1993), *An Introduction to the Bootstrap*, Chapman and Hall.

Golub, G. H. & Van Loan, C. F. (1983), *Matrix Computations*, Johns Hopkins University Press.

Hickernell, F. J. & Wozniakowski, H. (2000), 'Integration and approximation in arbitrary dimensions', *Advances in Computational Mathematics* . To appear.

Koehler, J. & Owen, A. (1996), Computer experiments, *in* S. Ghosh & C. Rao, eds, 'Handbook of Statistics, 13: Design and Analysis of Experiments', North-Holland, pp. 261–308.

McKay, M. D., Beckman, R. J. & Conover, W. J. (1979), 'A comparison of three methods for selecting values of input variables in the analysis of output from a computer code', *Technometrics* **21**(2), 239–45.

Owen, A. B. (1992), 'A central limit theorem for Latin hypercube sampling', *Journal of the Royal Statistical Society, Series B* **54**, 541–551.

Paskov, S. & Traub, J. (1995), 'Faster valuation of financial derivatives', *The Journal of Portfolio Management* **22**, 113–120.

Sloan, I. H. & Wozniakowski, H. (1998), 'When are quasi-Monte Carlo algorithms efficient for high dimensional integration?', *Journal of Complexity* **14**, 1–33.

Stein, M. (1987), 'Large sample properties of simulations using Latin hypercube sampling', *Technometrics* **29**(2), 143–51.

Art B. Owen
Department of Statistics
Sequoia Hall
Stanford University
Stanford CA 94305
USA

owen@stat.stanford.edu

Abbreviated title: Assessing Linearity