COMPRESSED SENSING FOR VOLTERRA AND POLYNOMIAL REGRESSION MODELS

Vassilis Kekatos and Georgios B. Giannakis

University of Minnesota, ECE Dept.
Minneapolis, MN 55455, USA
Emails:{kekatos,georgios}@umn.edu

ABSTRACT

Volterra filtering and polynomial regression are two widely utilized tools for nonlinear system modeling and inference. They are both critically challenged by the curse of dimensionality, which is typically alleviated via kernel regression. However, exciting diverse applications ranging from neuroscience to genome-wide association (GWA) analysis call for parsimonious polynomial expansions of critical interpretative value. Unfortunately, kernel regression cannot yield sparsity in the primal domain, where compressed sampling approaches can offer a viable alternative. Following the compressed sampling principle, a sparse polynomial expansion can be recovered by far fewer measurements compared to the least squares (LS)-based approaches. But how many measurements are sufficient for a given level of sparsity? This paper is the first attempt to answer this question by analyzing the restricted isometry properties for commonly met polynomial regression settings. Additionally, the merits of compressed sampling approaches to polynomial modeling are corroborated on synthetic and real data for quantitative genotype-phenotype analysis.

1. INTRODUCTION

The Volterra filter is a well-appreciated tool for modeling nonlinear systems. It basically approximates the system output as a polynomial expansion of the input using Taylor's theorem. Widespread applications span the gamut of physiological and biological processes, power amplifiers, loudspeakers, speech, and image models, to name a few; see e.g., [1], [8]. But the notion of polynomially expanding a nonlinearity goes beyond filters. Polynomial regression aims at approximating a multivariate nonlinear function via a polynomial expansion, and has been extensively used for prediction and classification tasks [10].

Volterra system identification and polynomial regression are both studied here. Even though they model nonlinear functions, their input-output (I/O) relationship is linear with respect to the unknown parameters. Hence, the model can be estimated via least-squares (LS) [8]. The major bottleneck is the "curse of dimensionality," since for a P-th order expansion over an L-variate input, the number of regression coefficients M grows as $\mathcal{O}(L^P)$. Beyond computational and identifiability challenges involved, high values of M dictate impractically long data records N [8], [4]. Exploiting advances in machine learning and viewing nonlinear modeling as a kernel regression task, tractable solutions can be devised [4], [10].

However, various applications admit a *sparse* polynomial expansion, that is, only a few, say s out of M expansion coefficients, are nonzero. For example, Volterra filters are used to model nonlinear devices such as loudspeakers or high-power amplifiers. These devices are in cascade with long yet sparse multipath channels, yielding eventually a sparse Volterra filter. In neuroscience, parsimonious Volterra filters are used to model the causal relationships in neuronal ensembles using spike-train data recorded from individual neurons [1]. In genome-wide association (GWA) studies, sparse polynomial (logistic) regression assists geneticists in identifying which genes determine certain human genetic diseases and traits in other species.

LS-based polynomial regression methods fail to handle the dimensionality involved, whereas (even kernelized) ridge regression cannot provide sparse models. Compressed sampling methods including the basis pursuit [3] and the (weighted) Lasso [11], [15], offer valuable tools for estimating parsimonious polynomial expansions. The results obtained so far are successful [13], [14], [7], but there is no theoretical justification, nor bounds on the (s,N,M) triplets that can be supported.

Towards this challenging direction, this paper provides such an analysis through the so called restricted isometry properties (RIP) of the involved polynomial regression matrix [3]. The Volterra and the polynomial models are treated separately and different probabilistic bounds are obtained. Our RIP analysis on meaningful random input models shows that an s-sparse linear-quadratic Volterra model can be recovered if the number of measurements N scales at least as $s^2 \log M$, whereas the bound becomes $s \log^4 M$ for polynomial regression models. Interestingly, the results generalize the RIP bounds derived for the corresponding linear system identification [6], and linear regression setups [3]. A comparative study of ridge and (w)Lasso regression on synthetic and real GWA data demonstrates the potential of sparsity-aware estimators.

Notation: Lowercase (upper-case) boldface letters are reserved for column vectors (matrices); $\mathbf{1}_N$ denotes the all-ones vector of dimension N. The notation $\|\mathbf{x}\|_p := \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}$ for $p \geq 1$ stands for the ℓ_p-norm in $\mathbb{R}^n$, whereas $\|\mathbf{x}\|_0$ the ℓ_0-(pseudo)norm which equals the number of nonzero entries of $\mathbf{x}$.

2. VOLTERRA FILTERS AND POLYNOMIAL REGRESSION

Consider a nonlinear, discrete-time, and time-invariant system described by the I/O relationship $y(n) = f(\mathbf{x}_1(n))$, where the input $\mathbf{x}_1(n) := [x(n) \ \ldots \ x(n-L+1)]^T$ has memory L. Under smoothness conditions [8], this I/O relationship can be approximated by a Volterra expansion of order P as

$$y(n) = \sum_{p=0}^{P} H_p[\mathbf{x}_1(n)] + v(n) \tag{1}$$

where $v(n)$ captures unmodeled dynamics and observation noise; and $H_p[\mathbf{x}_1(n)]$ is a p-th order polynomial provided as the output of the Volterra module $h_p(k_1,\ldots,k_p)$

$$H_p[\mathbf{x}_1(n)] := \sum_{k_1=0}^{L-1} \cdots \sum_{k_p=0}^{L-1} h_p(k_1,\ldots,k_p) \prod_{i=1}^{p} x(n-k_i). \tag{2}$$

The usefulness of polynomial expansion goes beyond system modeling. Polynomial regression aims at approximating a nonlinear function $y(n) = f(\mathbf{x}_1(n))$, where now $\mathbf{x}_1(n) := [x_0(n) \ \ldots \ x_{L-1}(n)]^T$. Setting $x_l(n) = x(n-l)$ for $l = 0,\ldots,L-1$,

Dr Kekatos' work was funded by the European Community's Seventh Framework Programme (grant FP7/2008-234914). Work was also supported by NSF grants CCF-0830480, 1016605, and ECCS-0824007, 1002180.

one can readily obtain the Volterra filter as a special case of polynomial regression. For the former, being a filter, each input vector is a shifted version of the previous one. This differentiating property will be critical in our subsequent analysis. However, we choose to use common notation for the two setups; any ambiguity will be easily resolved by the context.

Given the I/O samples $\{\mathbf{x}_1(n), y(n)\}_{n=1}^N$, the goal is to estimate $h_p(k_1,\ldots,k_p)$. This problem has been extensively studied [8], but sparsity present in many polynomial representations has not been exploited.

3. ESTIMATING SPARSE POLYNOMIAL EXPANSIONS

Define the vectors $\mathbf{x}_p(n) := \mathbf{x}_{p-1}(n) \otimes \mathbf{x}_1(n)$ for $p \geq 2$ where $\otimes$ is the Kronecker product. The n-th output can then be expressed as $y(n) = \mathbf{x}^T(n)\mathbf{h}$, where $\mathbf{x}^T(n) := \left[1\ \mathbf{x}_1^T(n)\ \ldots\ \mathbf{x}_P^T(n)\right]$, and $\mathbf{h}$ contains all the polynomial expansion coefficients appropriately stacked. Upon collecting all the I/O samples in $\mathbf{y} := [y(1)\ \cdots\ y(N)]^T$ and $\mathbf{X} := [\mathbf{x}(1)\ \ldots\ \mathbf{x}(N)]^T$, the following standard linear regression model is obtained (cf. (2))

$$\mathbf{y} = \mathbf{X}\mathbf{h} + \mathbf{v}. \tag{3}$$

Estimating $\mathbf{h}$ is critically challenged by its dimensionality which is originally $\left(L^{P+1}-1\right)/(L-1)$ [8]. By discarding some redundant expansion coefficients, vectors $\mathbf{h}$ and $\mathbf{x}(n)$ can be equivalently shortened to dimension [8]

$$M := \binom{L+P}{L} \tag{4}$$

which is still large. For notational simplicity, the symbols $\mathbf{h}$ and $\mathbf{X}$ will henceforth denote the shortened versions of the variables in (3), i.e., $\mathbf{X}$ will be $N \times M$.

Based on the linear wrt to $\mathbf{h}$ model (3), one can develop standard linear regression estimators for $\mathbf{h}$ [8]. One can form LS or ridge regression estimators, which can be expressed as the minimizers of

$$\min_{\mathbf{h}}\ \|\mathbf{y}-\mathbf{X}\mathbf{h}\|_2^2 + \delta\|\mathbf{h}\|_2^2 \tag{5}$$

where $\delta > 0$ ($\delta = 0$) for the ridge (LS) case. Sparing the widely studied computational, numerical, and identifiability issues raised mainly by the high value of M [8], [4], one can immediately recognize that the these two estimates $\hat{\mathbf{h}}^{LS}$ and $\hat{\mathbf{h}}^{Ridge}$ will not be sparse.

To account for the prior information on the sparsity of the unknown $\mathbf{h}$, which is critical in many interesting applications as advocated in Section 1, one can exploit recent advances in compressed sampling. Ignoring the noise in (3), $\mathbf{h}$ can be ideally recovered by solving

$$\min_{\mathbf{h}}\ \{\|\mathbf{h}\|_0 :\ \mathbf{y} = \mathbf{X}\mathbf{h}\}. \tag{6}$$

After recognizing the NP-hardness of solving (6), compressed sampling suggests solving instead the linear program [3]

$$\min_{\mathbf{h}}\ \{\|\mathbf{h}\|_1 :\ \mathbf{y} = \mathbf{X}\mathbf{h}\} \tag{7}$$

which is also known as basis pursuit. When $\mathbf{v}$ is considered, sparsity can be effected by the ℓ_1-norm regularized regression [11]

$$\min_{\mathbf{h}}\ \frac{1}{2}\|\mathbf{y}-\mathbf{X}\mathbf{h}\|_2^2 + \lambda_N \sum_{i=1}^{M} w_i|h_i| \tag{8}$$

where h_i is the i-th entry of $\mathbf{h}$, and $w_i > 0$ for $i = 1,\ldots,M$. Two choices for the weights w_i commonly adopted are: (w1) $w_i = 1$ for all i, which corresponds to the conventional Lasso estimator [11]; or, (w2) $w_i = |\hat{h}_i^{Ridge}|^{-1}$ that leads to the weighted Lasso (wLasso) estimator [15]. Among other choices, both problems can be efficiently solved by the coordinate descent method of [5] or the adaptive RLS-type algorithm proposed in [7].

The main objective here is to investigate these recoverability guarantees in the challenging Volterra system identification and polynomial regression setup as pursued next.

4. RESTRICTED ISOMETRY PROPERTIES

One of the main tools for specifying whether the optimization problems (7) and (8) can in general recover a sparse vector is the so called *restricted isometry properties* (RIP) of the involved regression matrix $\mathbf{X}$, defined as [3]:

Definition 1 (RIP). *Matrix $\mathbf{X} \in \mathbb{R}^{N\times M}$ possesses the RIP of order s, namely $\delta_{\mathbf{X}}(s)$, if $\delta_{\mathbf{X}}(s)$ is the minimum $\delta \in (0,1)$ satisfying*

$$(1-\delta)\|\mathbf{h}\|_2^2 \leq \|\mathbf{X}\mathbf{h}\|_2^2 \leq (1+\delta)\|\mathbf{h}\|_2^2 \tag{9}$$

for all s-sparse $\mathbf{h} \in \mathbb{R}^M$.

RIP were initially derived to characterize the recoverability of an s-sparse $\mathbf{h}_o$ given noiseless linear measurements $\mathbf{y} = \mathbf{X}\mathbf{h}_o$. It has been shown that the minimization in (6) can uniquely recover $\mathbf{h}_o$ if and only if $\delta_{\mathbf{X}}(2s) < 1$. If additionally $\delta_{\mathbf{X}}(2s) < \sqrt{2}-1$, then $\mathbf{h}_o$ is the unique minimizer of basis pursuit in (7) [9]. For noise-corrupted measurements, RIP-based recoverability guarantees can be derived for the Lasso estimator as well [2].

But finding the RIP of $\mathbf{X}$ is a hard combinatorial problem. Thus, to derive sparse recoverability guarantees one usually resorts to random matrix ensembles and provides probabilistic bounds on their RIP [3], [9]. In the generic linear regression setup, it has been shown that when the entries of $\mathbf{X}$ are independently Gaussian or Bernoulli, the matrix possesses sufficiently small RIP with high probability when N scales at least as $s\log(M/s)$ [3]. In a sparse system identification setup where the regression matrix has a Toeplitz structure, the condition on the number of measurements obtained so far loosens as $s^2 \log M$ for a Gaussian, Bernoulli, or uniform input [6]. The quadratic scaling of N wrt s in the latter bound versus the linear scaling in the former can be attributed to the statistical dependencies among the entries of $\mathbf{X}$ [9]. We next characterize the RIP of $\mathbf{X}$ for both the Volterra system identification and the polynomial regression setups.

5. RIP FOR VOLTERRA SYSTEM IDENTIFICATION

In our RIP analysis for Volterra filtering, two assumptions will be considered:

(as1) $\{x_n\}$ is independent uniformly distributed in $[-1,1]$; and

(as2) $P = 2$ (linear-quadratic Volterra model).

Regarding assumption (as1), we are usually interested in a bounded-input behavior of a nonlinear system. For (as2), a second-order model is a commonly used Volterra model mainly due to computational reasons. Note that $\mathbb{E}[x_n^2] = 1/3$, $\mathbb{E}[x_n^4] = 1/5$, and $M = \binom{L+2}{2}$.

To proceed with the RIP analysis, let us first define the Grammian matrix $\mathbf{R} := \mathbf{X}^T\mathbf{X}$ and let $R_{i,j}$ denote its (i,j)-th entry. As proved in [6], the matrix $\mathbf{X}$ possesses the RIP $\delta_{\mathbf{X}}(s) \leq \delta$ if there exist positive δ_d and δ_o with $\delta_d + \delta_o = \delta$ such that $|R_{ii}-1| < \delta_d$ and $|R_{ij}| < \delta_o/s$ for every i, j with $j \neq i$. When these conditions hold, then the Geršgorin disc theorem guarantees that the eigenvalues of the Grammian matrices formed by any combination of s columns of $\mathbf{X}$ lie in the interval $[1-\delta, 1+\delta]$ and then $\delta_{\mathbf{X}}(s) < \delta$ follows by definition. In a nutshell, for a regression matrix $\mathbf{X}$ to have small $\delta_{\mathbf{X}}(s)$ and, hence, favorable compressed sampling properties, it suffices that its Grammian $\mathbf{R}$ has diagonal entries close to unity and off-diagonal entries close to zero.

In Volterra filtering, the diagonal entries $\{R_{ii}\}$ are not equal to one; but an appropriate normalization of the columns of $\mathbf{X}$ can provide at least $\mathbb{E}[R_{ii}] = 1$. The law of large numbers dictates that given sufficiently enough measurements N, the R_{ii}'s will approach their mean value. Likewise, it is desirable for the off-diagonal entries of $\mathbf{R}$ to be zero mean, so that for large N they vanish. Such a requirement is not inherently satisfied by all R_{ij}'s; e.g., the inner product between the $\mathbf{X}$ columns of the form $\left[x_n^2\ x_{n+1}^2\ \ldots\ x_{n+N-1}^2\right]^T$ and $\left[x_{n-k}^2\ x_{n-k+1}^2\ \cdots\ x_{n-k+N-1}^2\right]^T$ for some n and $k > 0$ has expected value $N\left(\mathbb{E}[x_n^2]\right)^2$ that is strictly positive.

To achieve the desired properties (p1) $\mathbb{E}[R_{ii}] = 1$, and (p2) $\mathbb{E}[R_{ij}] = 0$ for all i, j with $j \neq i$, instead of studying the RIP of $\mathbf{X}$, we focus on the modified Volterra regression matrix

$$\tilde{\mathbf{X}} := \left[\begin{array}{cccc} \tilde{\mathbf{x}}^c & \tilde{\mathbf{X}}^l & \tilde{\mathbf{X}}^q & \tilde{\mathbf{X}}^b \end{array} \right] \tag{10}$$

where $\tilde{\mathbf{x}}^c := \sqrt{\frac{1}{N}}\mathbf{1}_N$ corresponds to the intercept, $\tilde{\mathbf{X}}^l$ and $\tilde{\mathbf{X}}^q$ are two $N \times L$ Toeplitz matrices corresponding to the linear and quadratic parts defined as

$$\tilde{\mathbf{X}}^l := \sqrt{\frac{3}{N}} \left[\begin{array}{cccc} x_0 & x_{-1} & \cdots & x_{-L+1} \\ x_1 & x_0 & \cdots & x_{-L+2} \\ \vdots & \vdots & & \vdots \\ x_{N-1} & x_{N-2} & \cdots & x_{N-L+1} \end{array} \right]$$

$$\tilde{\mathbf{X}}^q := \frac{3}{2}\sqrt{\frac{5}{N}} \left[\begin{array}{cccc} x_0^2 - \frac{1}{3} & x_{-1}^2 - \frac{1}{3} & \cdots & x_{-L+1}^2 - \frac{1}{3} \\ x_1^2 - \frac{1}{3} & x_0^2 - \frac{1}{3} & \cdots & x_{-L+2}^2 - \frac{1}{3} \\ \vdots & \vdots & & \vdots \\ x_{N-1}^2 - \frac{1}{3} & x_{N-2}^2 - \frac{1}{3} & \cdots & x_{N-L+1}^2 - \frac{1}{3} \end{array} \right]$$

and $\tilde{\mathbf{X}}^b$ is a $N \times \frac{L(L-1)}{2}$ (non-Toeplitz) matrix related to the bilinear part given by

$$\tilde{\mathbf{X}}^b := \frac{3}{\sqrt{N}} \left[\begin{array}{cccc} x_0 x_{-1} & x_0 x_{-2} & \cdots & x_{-L+2} x_{-L+1} \\ x_1 x_0 & x_1 x_{-1} & \cdots & x_{-L+3} x_{-L+2} \\ \vdots & \vdots & & \vdots \\ x_{N-1} x_{N-2} & x_{N-1} x_{N-3} & \cdots & x_{N-L+2} x_{N-L+1} \end{array} \right].$$

Consider now the Grammian matrix of $\tilde{\mathbf{X}}$ defined as $\tilde{\mathbf{R}} := \tilde{\mathbf{X}}^T \tilde{\mathbf{X}}$. Comparing $\mathbf{X}$ with $\tilde{\mathbf{X}}$, the columns of $\tilde{\mathbf{X}}$ have their ℓ_2-norm normalized in expectation and, thus, $\tilde{\mathbf{R}}$ satisfies (p1). Moreover, those columns of $\tilde{\mathbf{X}}$ corresponding to the quadratic part (cf. submatrix $\tilde{\mathbf{X}}^q$) have been shifted by the variance of the input. One can readily verify that (p2) is then satisfied too.

The transition from $\mathbf{X}$ to $\tilde{\mathbf{X}}$ serves analytical only purposes, but raises a legitimate question: Do the RIP of $\tilde{\mathbf{X}}$ provide any insight on the compressed sampling guarantees for the original Volterra problem? In the noiseless scenario, we have actually substituted the optimization problem in (7) by

$$\min_{\tilde{\mathbf{h}}} \left\{ \|\tilde{\mathbf{h}}\|_1 : \ \mathbf{y} = \tilde{\mathbf{X}}\tilde{\mathbf{h}} \right\}. \tag{11}$$

Upon matching the feasible sets of the two problems, i.e., $\mathbf{X}\mathbf{h} = \tilde{\mathbf{X}}\tilde{\mathbf{h}}$, we get the one-to-one mapping

$$h_0 = \frac{1}{\sqrt{N}}\tilde{h}_0 - \frac{1}{2}\sqrt{\frac{5}{N}}\sum_{k=1}^{L}\tilde{h}_2(k,k) \tag{12a}$$

$$h_1(k) = \sqrt{\frac{3}{N}}\tilde{h}_1(k), \ \forall k \tag{12b}$$

$$h_2(k,k) = \frac{3}{2}\sqrt{\frac{5}{N}}\tilde{h}_2(k,k), \ \forall k \tag{12c}$$

$$h_2(k_1,k_2) = \frac{3}{\sqrt{N}}\tilde{h}_2(k_1,k_2), \ \forall k_1 \neq k_2. \tag{12d}$$

It is now apparent that a sparse solution for (11) translates to a sparse solution for (7) with the minor exception of the constant term in (12a). By deterministically adjusting the weights $\{w_i\}_{i=1}^M$ and the parameter λ_N in (8), this argument carries over to the Lasso and answers affirmatively the previously posed question.

To probabilistically characterize the RIP of $\tilde{\mathbf{X}}$, we use graph theoretic arguments to model the dependencies across the entries of $\mathbf{R}$ and extend the concentration results of [6] to the Volterra system identification case. One of the main results of this work is summarized in the following theorem whose proof is omitted due to space limitations.

Theorem 1 (Volterra filtering)**.** *Let $\{x_i\}_{i=-L+1}^N$ be a sequence of independent random variables drawn from $\mathscr{U}[-1,1]$. Assume the matrix $\tilde{\mathbf{X}}$ defined in* (10) *is generated by this sequence for $L \geq 7$ and $N \geq 160$. Then, for any $\delta \in (0,1)$ and for any $\gamma \in (0,1)$, whenever $N \geq \frac{5C}{(1-\gamma)\delta^2} \cdot s^2 \log L$, the matrix $\tilde{\mathbf{X}}$ possesses $\delta_{\tilde{\mathbf{X}}}(s) < \delta$ for $s \geq 2$ with probability exceeding $1 - \exp\left(-\frac{\gamma\delta^2}{C} \cdot \frac{N}{s^2}\right)$, where $C = 2,835$.*

The theorem asserts that at least $s^2 \log L$ observations are sufficient to recover an s-sparse non-homogeneous second-order Volterra filter of memory L probed by a uniformly distributed input. Since the number of Volterra coefficients M is $\mathscr{O}(L^2)$, the number of observations scales also as $s^2 \log M$. The bound agrees with the bounds obtained for the linear filtering setup [6], but here the constants are larger due to the significantly more involved dependencies among the entries of the associated regression matrix.

6. RIP FOR POLYNOMIAL REGRESSION

Consider first the linear-quadratic model

$$f(\mathbf{x}) = h_0 + \sum_{k=1}^{L} h_1(k)x_k + \sum_{k_1=1}^{L}\sum_{k_2=k_1}^{L} h_2(k_1,k_2)x_{k_1}x_{k_2}. \tag{13}$$

Given N output samples $\{y_n\}_{n=1}^N$ corresponding to input data $\{\mathbf{x}_1(n))\}_{n=1}^N$ drawn independently from $\mathscr{U}[-1,1]^L$, the goal is to recover the $M \times 1$ sparse vector $\mathbf{h}$ comprising the $h_1(k)$'s and $h_2(k_1,k_2)$'s. Note that $M = (L+1)(L+2)/2$ here. Contrary to the Volterra filtering setup, the rows of $\mathbf{X}$ are now statistically independent; a fact that differentiates the RIP analysis for polynomial regression and leads to tighter bounds.

The analysis builds on [9], which deals with finding a sparse expansion of a function $f(\mathbf{x}) = \sum_{t=1}^T c_t \psi_t(\mathbf{x})$ over a bounded orthonormal set $\{\psi_t(\mathbf{x})\}$. Considering $\mathscr{D}$ a measurable space, e.g., a measurable subset of $\mathbb{R}^L$, endowed with a probability measure ν, the set of functions $\{\psi_t(\mathbf{x}) : \mathscr{D} \to \mathbb{R}\}_{t=1}^T$ is a bounded orthonormal set if for every $t_1, t_2 = 1, \ldots, T$

$$\int_{\mathscr{D}} \psi_{t_1}(\mathbf{x})\psi_{t_2}(\mathbf{x}) d\nu(\mathbf{x}) = \delta_{t_1,t_2} \tag{14}$$

where δ_{t_1,t_2} is the Kronecker delta function, and for some constant $K \geq 1$ it holds that

$$\sup_t \sup_{\mathbf{x}\in\mathscr{D}} |\psi_t(\mathbf{x})| \leq K. \tag{15}$$

After sampling $f(\mathbf{x})$ at $\{\mathbf{x}(n) \in \mathscr{D}\}_{n=1}^N$, the involved $N \times T$ regression matrix $\mathbf{\Psi}$ with entries $\Psi_{n,t} := \psi_t(\mathbf{x}(n))$ admits the following RIP characterization [9, Th. 4.4]:

Theorem 2. *Let $\mathbf{\Psi}$ be the $N \times T$ matrix associated with a bounded orthonormal system with constant $K \geq 1$ in* (15)*. Then, for any $\delta \in (0, 0.5]$, there exist universal positive constants C and γ, such that whenever $N \geq \frac{CK^2}{\delta^2} \cdot s \log^4 T$, the matrix $\frac{1}{\sqrt{N}}\mathbf{\Psi}$ possesses $\delta_{\mathbf{\Psi}/\sqrt{N}}(s) < \delta$ with probability exceeding $1 - \exp\left(-\frac{\gamma\delta^2}{CK^2} \cdot \frac{N}{s}\right)$.*

In the regression model of (13), even though the basis functions $\{1, \{x_i\}, \{x_{i_1}x_{i_2}\}\}$ are bounded in $[-1,1]^L$, they are not orthonormal in the uniform probability measure. Fortunately, our input transformation trick devised for the Volterra model applies here too. The expansion is now over the $M = \binom{L+2}{2}$ basis functions $\{\psi_m(\mathbf{x})\}_{m=1}^M$

$$\left\{1, \{\sqrt{3}x_i\}, \left\{\frac{3\sqrt{5}}{2}\left(x_i^2 - \frac{1}{3}\right)\right\}, \{3x_{i_1}x_{i_2}\}\right\} \tag{16}$$

where the last subset contains all the unique, two-variable monomials lexicographically ordered. Upon stacking the function values $\{y_n\}_{n=1}^N$ in $\mathbf{y}$ and properly defining $\tilde{\mathbf{h}}$, the expansion $\mathbf{y}=\mathbf{X}\mathbf{h}$ can be replaced by $\mathbf{y}=\tilde{\mathbf{X}}\tilde{\mathbf{h}}$, where the entries of $\tilde{\mathbf{X}}$ are

$$\tilde{X}_{n,m} := \frac{\psi_m(\mathbf{x}(n))}{\sqrt{N}}. \tag{17}$$

Vectors $\mathbf{h}$ and $\tilde{\mathbf{h}}$ are related through the one-to-one mapping of (12). Obviously, the identifiability of a sparse $\mathbf{h}$ can be guaranteed by the RIP analysis of $\tilde{\mathbf{X}}$ presented in the next lemma.

Lemma 1 (Linear-quadratic regression)**.** *Let $x_i(n)$ for $i=1,\ldots,L$ and $n=1,\ldots,N$ independent random variables uniformly distributed in $[-1,1]$. Assume that the $N\times M$ matrix $\tilde{\mathbf{X}}$ in* (17) *is generated by this sequence for $L\geq 4$. Then, for any $\delta\in(0,0.5]$, there exist universal positive constants C and γ, such that whenever $N\geq \frac{144C}{\delta^2}\cdot s\log^4 L$, the matrix $\tilde{\mathbf{X}}$ possesses $\delta_{\tilde{\mathbf{X}}}(s)<\delta$ with probability exceeding $1-\exp\left(-\frac{\gamma\delta^2}{9C}\cdot\frac{N}{s}\right)$.*

Proof. The inputs $\mathbf{x}(n)$ are uniformly drawn over $\mathscr{D}=[-1,1]^L$, and it is easy to verify that the basis functions $\{\psi_m(\mathbf{x})\}_{m=1}^M$ in (16) form a bounded orthonormal system with $K=3$. Hence, Theorem 2 can be straightforwardly applied. Since $M\leq L^2$ for $L\geq 4$, it follows that $\log^4 M<16\log^4 L$. □

Lemma 1 assures that an s-sparse linear-quadratic L-variate expansion with independent uniformly distributed inputs can be identified with high probability from a number of observations at least in the order of $s\log^4 L$ or $s\log^4 M$. Comparing this to Theorem 1, the number of required observations here scales linearly with s. Moreover, except for the increase in the power of the logarithmic factor, the bound is close to the one obtained for random Gaussian and Bernoulli matrices in the sparse linear regression setup [3], [9]. The improvement over the Volterra RIP bound is explained by the simpler structural dependence in $\mathbf{X}$.

Another interesting regression paradigm is when $f(\mathbf{x})$ is amenable to the sparse multilinear expansion

$$f(\mathbf{x})=h_0+\sum_{k=1}^{L}h_1(k)x_k+\sum_{k_1=1}^{L}\sum_{k_2=k_1+1}^{L}h_2(k_1,k_2)x_{k_1}x_{k_2}+\ldots$$
$$+\sum_{k_1=1}^{L}\sum_{k_2=k_1+1}^{L}\ldots\sum_{k_P=k_{P-1}+1}^{L}h_P(k_1,k_2,\cdots,k_P)x_{k_1}x_{k_2}\ldots x_{k_P} \tag{18}$$

where, in contrast to the polynomial model, each variable is raised only in the first power. This is the regression model typically engaged in GWA studies [13], [14]. Because there are $\binom{L}{p}$ monomials of order p and, vector $\mathbf{h}$ here has dimension $M=\sum_{p=0}^{P}\binom{L}{p}\leq (L+1)^P$.

The goal is again to recover an s-sparse $\mathbf{h}$ given the phenotypes $\{y_n\}_{n=1}^N$ over the genotype values $\{\mathbf{x}_1(n)\}_{n=1}^N$. Vectors $\mathbf{x}_1(n)$ are drawn either from $\{-1,0,1\}^L$ or $\{-1,1\}^L$ depending on the assumed genotype model (additive for the first alphabet; dominant or recessive for the latter) [13]. Without loss of generality, consider the ternary alphabet with equal probabilities. Further, suppose for analytical convenience that the entries of $\mathbf{x}_1(n)$ are independent. Note that the input has zero mean and variance $2/3$.

The RIP analysis for the model in (18) exploits Theorem 2. The basis functions involved now, i.e., $\{1,\{x_i\},\{x_{i_1}x_{i_2}\},\ldots\}$, are orthogonal wrt the assumed point mass function. An orthonormal set $\{\psi_m(\mathbf{x})\}_{m=1}^M$ can be constructed after a simple scaling as

$$\left\{1,\left\{K^{\frac{1}{P}}x_{i_1}\right\},\left\{K^{\frac{2}{P}}x_{i_1}x_{i_2}\right\},\ldots,\left\{K^{\frac{P}{P}}x_{i_1}x_{i_2}\cdots x_{i_P}\right\}\right\} \tag{19}$$

Table 1: Experimental results for QTL data

Method	PE	MSE	NNZ	δ/λ
Synthetic data				
Ridge regression	68.10	82.29	7382	0.61 N
Lasso	12.84	15.85	200	0.19 N
wLasso	13.09	5.11	85	3.77 N
Real barley data				
Ridge regression	8.26	-	8129	$4.28\cdot 10^4$ N
Lasso	5.96	-	48	0.33 N
wLasso	5.69	-	34	6.88 N

where $K=(3/2)^{P/2}$ is the function set bound. Similar to the linear-quadratic case in (13), the original multilinear expansion $\mathbf{X}\mathbf{h}$ is transformed to $\tilde{\mathbf{X}}\tilde{\mathbf{h}}$, where $\tilde{\mathbf{X}}$ is defined as in (17) with the new basis of (19), and $\tilde{\mathbf{h}}$ is an entry-wise rescaled version of $\mathbf{h}$. Based on these facts, the RIP analysis of $\tilde{\mathbf{X}}$ follows readily from the ensuing lemma.

Lemma 2 (Multilinear expansion)**.** *Let $x_i(n)$ for $i=1,\ldots,L$ and $n=1,\ldots,N$ independent random variables equiprobably drawn from $\{-1,0,1\}$. The $N\times M$ modified multilinear regression matrix $\tilde{\mathbf{X}}$ defined in* (17) *and* (19) *is generated by this sequence. Then, for any $\delta\in(0,0.5]$, there exist universal positive constants C and γ, such that whenever $N\geq\frac{C}{\delta^2}\left(\frac{3}{2}\right)^P P^4 s\log^4(L+1)$, the matrix $\tilde{\mathbf{X}}$ possesses $\delta_{\tilde{\mathbf{X}}}(s)<\delta$ with probability exceeding $1-\exp\left(-\frac{\gamma\delta^2}{C(3/2)^P}\cdot\frac{N}{s}\right)$.*

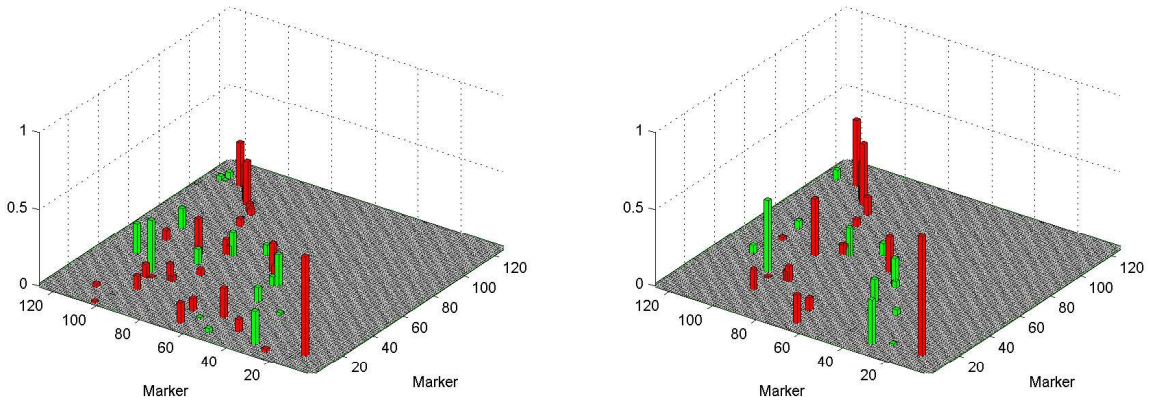
Lemma 2 guarantees the RIP to hold with high probability for N in the order of $(3/2)^P P^4 s\log^4 L$; since P is typically 2, the number of phenotype samples needed is $s\log^4 L$.

7. SIMULATED TESTS

Our tests investigate the potential of sparse polynomial regression for quantitative trait analysis. In quantitative genetics, the phenotype is a quantitative trait of an organism, e.g., the height of barley [12]. The phenotype is assumed to follow a regression model over single-gene (main) and gene-gene (epistatic) effects [14]. Determining the so called quantitative trait loci (QTL) corresponds to revealing the (pair of) genes associated with a particular trait. Since the studied population N is much smaller than the number of regressors M and given that only a few genes determine the trait considered, the sparse multilinear (for $P=2$) model of (18) is considered.

Synthetic data: The first paradigm is a synthetic QTL study detailed in [14]. A population of $N=600$ individuals is simulated over $L=121$ genes. The true population mean and variance are 5.0 and 10.0, respectively. The phenotype is assumed to be linearly expressed over the intercept, the L main effects, and the $\binom{L}{2}=7260$ epistatic effects, i.e., a total of $M=7382$ regressors. The QTLs simulated are 9 single markers and 13 marker pairs. Note that the simulation accommodates markers with main only; epistatic only; and both main and epistatic effects.

Parameters δ and λ for ridge and (w)Lasso estimators, respectively, were tuned through 10-fold cross validation based on the prediction error (PE) [10]; see Table 1. Having tuned the regularization parameters, the mean-squared error (MSE) attained by each method was averaged over 100 Monte Carlo runs. The (w)Lasso estimators were run using the glmnet software [5]. As can be seen from Table 1, Lasso attains the smaller PE, but, wLasso provides significantly higher estimation accuracy. The number of non-zero coefficients indicated in the fourth column shows that ridge regression yields an over-saturated model.



(a) Lasso

(b) wLasso

Figure 1: Regression vector estimates for the real QTL barley data. The main (epistatic) effects are shown on the diagonal (left diagonal part), while red (green) bars correspond to positive (negative) entries.

Real barley data: The second QTL experiment entails a real dataset collected as described in [12], [14]. The height of N=145 doubled-haploid lines of a cross between two barley lines, Harrington and TR306, was averaged under different environments. There were $L = 127$ markers, binary coded as $+1$ (-1) for the TR306 (Harrington) allele, yielding $M = 1 + 127 + \binom{127}{2} = 8129$ regressors. There was a 5% of missing values modeled as zeros as a simple way to minimize their effect [14].

Tuning the parameters was performed via leave-one-out cross validation [10]; see Table 1. The ridge estimator failed to handle over-fitting and δ is set to a large value yielding regression coefficients of insignificant amplitude. Using the ridge estimates to weight the regression coefficients, wLasso yielded a PE slighty smaller than the one attained by Lasso, but, more importantly, the number of spurious coefficients was reduced as shown in Fig. 1. Interestingly, some of the epistatic effects revealed involved genes not having a main effect.

8. CONCLUSIONS

Volterra filtering and polynomial regression are critically challenged by the curse of dimensionality. Exciting and diverse applications call for parsimonious polynomial expansions of critical interpretative value. Under such scenarios, the number of measurements needed to recover the sparse underlying model can be significantly decreased following the principle of compressed sampling. Existing results, including the ones conducted here on real and synthetic QTL data, indicate the potential of the venture. To theoretically quantify these limits, the RIP for Volterra filtering and polynomial regression were analyzed here. The bounds obtained is the first attempt to characterize the trade-offs between sparsity and number of measurements needed, and they generalize the bounds proved for the related linear cases.

REFERENCES

[1] T. W. Berger, D. Song, R. H. M. Chan, and V. Z. Marmarelis, "The neurobiological basis of cognition: Identification by multi-input, multi-output nonlinear dynamic modeling," *Proc. IEEE*, vol. 98, no. 3, pp. 356–374, Mar. 2010.

[2] P. J. Bickel, Y. Ritov, and A. B. Tsybakov, "Simultaneous analysis of Lasso and Dantzig selector," *Ann. Statist.*, vol. 37, no. 4, pp. 1705–1732, 2009.

[3] E. J. Candès and T. Tao, "Decoding by linear programming," *IEEE Trans. Inf. Theory*, vol. 51, no. 12, pp. 4203–4215, Dec. 2005.

[4] M. Franz and B. Scholkopf, "A unifying view of Wiener and Volterra theory and polynomial kernel regression," *Neural Computation*, vol. 18, pp. 3097–3118, 2006.

[5] J. H. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *Journal of Statistical Software*, vol. 33, no. 1, pp. 1–22, 2010.

[6] J. Haupt, W. U. Bajwa, G. Raz, and R. Nowak, "Toeplitz compressed sensing matrices with applications to channel sensing," *IEEE Trans. Inf. Theory*, 2011, *(to appear)*.

[7] V. Kekatos, D. Angelosante, and G. B. Giannakis, "Sparsity-aware estimation of nonlinear Volterra kernels," in *Proc. CAMSAP*, Aruba, Dutch Antilles, Dec. 2009.

[8] V. Mathews and G. Sicuranza, *Polynomial Signal Processing*. John Wiley & Sons Inc., 2000.

[9] H. Rauhut, "Compressive sensing and structured random matrices," *Theoretical Foundations and Numerical Methods for Sparse Recovery*, vol. 9 of Radon Series Comp. Appl. Math., pp. 1–92, 2010.

[10] B. Scholkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA: MIT Press, 2002.

[11] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *J. R. Stat. Soc. Ser.*, vol. 58, no. 1, pp. 267–288, 1996.

[12] N. A. Tinker, D. E. Mather, B. G. Rossnagel, K. J. Kasha, A. Kleinhofs, P. M. Hayes, and D. E. Falk, "Regions of the genome that affect agronomic performance in two-row barley," *Crop Science*, vol. 36, pp. 1053–1062, 1996.

[13] T. T. Wu, Y. F. Chen, T. Hastie, E. Sobel, and K. Lange, "Genome-wide association analysis by Lasso penalized logistic regression," *Bioinformatics*, vol. 25, no. 6, pp. 714–721, 2009.

[14] S. Xu, "An empirical Bayes method for estimating epistatic effects of quantitative trait loci," *Biometrics*, vol. 63, no. 2, pp. 513–521, 2007.

[15] H. Zou, "The adaptive Lasso and its oracle properties," *J. of the American Stat. Assoc.*, vol. 101, no. 476, pp. 1418–1429, Dec. 2006.