Delft University of Technology

Blind Polynomial Regression

Natali, Alberto; Leus, Geert

Citation (APA)
Natali, A., & Leus, G. (2023). Blind Polynomial Regression. In *Proceedings of the ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* IEEE. https://doi.org/10.1109/ICASSP49357.2023.10095361

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

BLIND POLYNOMIAL REGRESSION

Alberto Natali and Geert Leus

Faculty of Electrical Engineering, Mathematics and Computer Science
Delft University of Technology, Delft, The Netherlands

ABSTRACT

Fitting a polynomial to observed data is an ubiquitous task in many signal processing and machine learning tasks, such as interpolation and prediction. In that context, input and output pairs are available and the goal is to find the coefficients of the polynomial. However, in many applications, the input may be partially known or not known at all, rendering conventional regression approaches not applicable. In this paper, we formally state the (potentially partial) blind regression problem, illustrate some of its theoretical properties, and propose an algorithmic approach to solve it. As a case-study, we apply our methods to a jitter-correction problem and corroborate its performance.

Index Terms— polynomial regression, interpolation, Vandermonde, matrix factorization, MUSIC

1. INTRODUCTION

Parameter estimation is a classical task in signal processing [1], where the objective is to estimate the parameters of a hypothesized model that best explains the relationship between observed input-output data pairs. When the input is not directly observable, standard estimation approaches such as the ordinary least squares, cannot be directly applied, and new estimation techniques are required.

Motivated by a recent problem arising in graph signal processing [2] in the context of graph learning [3], in this work we consider a polynomial regression model, i.e., where the output variable is expressed as the linear combination of powers of the input variable, with possible additive noise. As we will see, even though the problem can be approached through the lens of matrix factorization [4] with a Vandermonde factor, a rather natural interpretation is given by that of *sampling multiple unknown polynomials at the same unknown locations*. Specifically, we assume to observe a certain number of sampled polynomials, whose coefficients and sample locations are unknown or noisy, and the goal is to jointly recover them from the available observations.

With reference to prior work, one of the earliest works in the context of sampling at unknown locations is that of [5], which considers discrete time bandlimited signals where only a subset of its values are known in correspondence to an ordered unknown subset of the index set. The resulting combinatorial optimization problem is then approached with an exhaustive search and two heuristics methods. The work in [6] considers the continuous case, and proposes an alternating least squares method that converges to a local minimum. More similar to our work is that of [7], which considers sampled polynomials (and bandlimited functions), and gives the conditions for the unique identifiability of the sampled polynomial. Specifically, it is shown that unique identifiability is achieved by constraining the sampling locations to be described by rational functions.

Our work can be considered complementary to that of [7] with some distinctions; first, we do not constrain the sample locations, and assume to observe different sampled polynomials all sampled at the same *unknown* locations. This implicitly restricts the feasible sample locations. Then, we study which are the model ambiguities, representing an equivalence class and show that for some problems, namely the one in [3], every element of this class (which is a solution for the problem) is a good solution. Finally, we propose a subspace-based method to solve the problem. Numerical simulations in the context of sampling-jitter corroborate our theory.

2. PROBLEM DEFINITION

Consider a simple polynomial regression model [8] of the form:

$$y = w_0 + w_1 x + \ldots + w_{K-1} x^{K-1} + e \qquad (1)$$

$$= \mathbf{w}^\top \mathbf{v}(x) + e, \qquad (2)$$

where y is the independent variable (*observation*), x is the dependent variable (*regressor*), $\mathbf{w} := [w_0, \ldots, w_{K-1}]^\top$ are the coefficients associated to the different polynomial degrees and $\mathbf{v}(x) := [1, x, \ldots, x^{K-1}]^\top$ is the Vandermonde vector with parameter $x \in \mathbb{R}$; $e \sim \mathcal{N}(0, \sigma^2)$ is a noise term.

Although polynomial regression fits a non linear model to the (available) data, from an estimation point of view it is linear, meaning that for a fixed input variable x, the dependence of the output variable y on the weights in $\mathbf{w}$ is linear. This data-availability scenario is often encountered in signal processing in the context (for instance) of signal interpolation and prediction, where input-output pairs $\{(x_i, y_i)\}_{i=0}^{N-1}$ are available by sampling an unknown function in N distinct points, and the goal is to estimate the coefficient vector $\mathbf{w}$ which, for a specified K, best represents the unknown polynomial relationship between x and y.

Expressing (1) for all the input-output pairs leads to the matrix-vector form[1]:

$$\mathbf{y} = \mathbf{V}(\mathbf{x})\mathbf{w} \qquad (3)$$

where $\mathbf{x} := [x_0, \ldots, x_{N-1}]^\top$ is the vector of inputs, $\mathbf{y} = [y_0, \ldots, y_{N-1}]^\top$ is the vector of outputs, and where $\mathbf{V}(\mathbf{x})$ is the $N \times K$ Vandermonde matrix, defined as:

$$\mathbf{V}(\mathbf{x}) := \begin{bmatrix} 1 & x_0 & \cdots & x_0^{K-1} \\ 1 & x_1 & \cdots & x_1^{K-1} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_{N-1} & \cdots & x_{N-1}^{K-1} \end{bmatrix} \qquad (4)$$

The unique solution of (3), guaranteed when all x_i's are different and $N > K$, can be found as $\hat{\mathbf{w}} = (\mathbf{V}^\top \mathbf{V})^{-1} \mathbf{V}^\top \mathbf{y}$. As a shorthand notation, we will use $\mathbf{V}$ to refer to the Vandermonde matrix

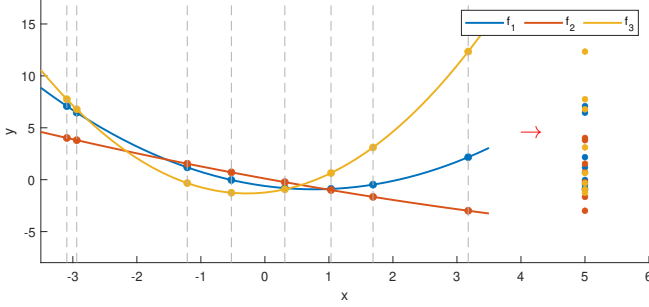[1]We will neglect the error term in the rest of the exposition.

Fig. 1: Illustration of the problem: a number of functions ($L = 3$) are sampled (filled points) in $N = 8$ locations (gray dotted line). The problem is to recover the sampling locations and the polynomial coefficients with the only knowledge of the function values (shown in the right in the figure).

with parameter $\mathbf{x}$ in (4); when convenient for clarity of exposition, we will explicitly write $\mathbf{V}(\mathbf{x})$.

However, it is often the case that variable x is not perfectly known or not known at all. For instance, in analog-to-discrete conversion of signals, undesired jitter on the clock signal leads to a clock drift with respect to the reference clock and a re-synchronization scheme is necessary. For this reason, in this work we consider the scenario in which not only the linear weights $\mathbf{w}$ are unknown, but also the vector $\mathbf{x}$. In order to overcome unique identifiability and solvability issues, we assume to obtain $L \geq K$ observation vectors $\{\mathbf{y}_i\}_{i=0}^{L-1}$ of (3) associated to different *unknown* coefficient vectors $\{\mathbf{w}_i\}_{i=0}^{L-1}$ yet having the same *unknown* vector $\mathbf{x}$, that is,:

$$\mathbf{Y} = \mathbf{V}(\mathbf{x})\mathbf{W} \qquad (5)$$

where $\mathbf{Y} := [\mathbf{y}_0, \ldots, \mathbf{y}_{L-1}]$ and $\mathbf{W} := [\mathbf{w}_0, \ldots, \mathbf{w}_{L-1}]$. Such a model can be motivated by the fact that the L observation channels are controlled by the same clock and hence have the same potential clock jitter. Based on (5), the problem statement can be formalized as follows:

Problem Statement. *Given the matrix $\mathbf{Y} \in \mathbb{R}^{N \times L}$, recover the input vector $\mathbf{x} \in \mathbb{R}^N$ and the coefficient matrix $\mathbf{W} \in \mathbb{R}^{K \times L}$ such that* (5) *holds.*

Although the problem can be approached from a pure algebraic point of view as a structured matrix factorization, a pleasing geometrical interpretation of (5) is given in Fig. 1. Each vector $\mathbf{y}_0, \ldots, \mathbf{y}_{L-1}$ can be interpreted as function values obtained by sampling L distinct polynomials $y_0(x), \ldots, y_{L-1}(x)$, all with degree $K - 1$, in the same N unknown locations $x_0, \ldots, x_{N-1}$. The goal is to recover the original locations (and polynomial coefficients) from the available sampled function values (filled points on the right of the figure).

Remark 1. At first glance, model (5) shares similarities with fundamental array signal processing models encountered for instance in direction of arrival (DoA) estimation [9]. However, the geometry of our problem is different (transposed) and complicates its analysis in that ready-to-use algorithms, such as MUSIC [10], are not directly applicable; see also Remark 2.

3. PROPOSED PROBLEM SOLUTIONS

Before delving into the algorithmic framework, we ask ourselves which are the solution pairs $(\mathbf{x}, \mathbf{W})$ that can be considered a "valid solution" for problem (5). Indeed notice how (5) is not free of model

ambiguities, and different pairs $(\mathbf{x}, \mathbf{W})$ may lead to the same observations matrix $\mathbf{Y}$. Due to its theoretical relevance (with practical consequences), this will be the subject of study of this section. Subsequently, we will propose two different approaches to solve the problem: the first based on a selection sampling scheme, the second based on subspace fitting.

Ambiguities. If $\mathbf{V}$ and $\mathbf{W}$ are the true matrix factors satisfying (5), then for any $K \times K$ invertible matrix $\mathbf{T}$, it holds:

$$\mathbf{Y} = \mathbf{V}\mathbf{T}\mathbf{T}^{-1}\mathbf{W} = \mathbf{V}'\mathbf{W}'. \qquad (6)$$

However, $\mathbf{V}'$ needs to be Vandermonde in order for (6) to be consistent with the model structure in (5). For this reason, we now investigate which class of matrices maps a Vandermonde matrix to another Vandermonde matrix, and is thus responsible for the ambiguities in (5).

To study the possible ambiguities of the system $\mathbf{V}_1\mathbf{T} = \mathbf{V}_2$, with $\mathbf{V}_1$ and $\mathbf{V}_2$ Vandermonde matrices, means to study which linear transformation $\mathbf{T}$ transforms a Vandermonde vector with parameter $x \in \mathbb{R}$ to another Vandermonde vector with parameter y (independent of the specific parameter instantiation), i.e.,:

$$\underbrace{[1, y, y^2, \ldots, y^{K-1}]}_{\mathbf{v}(y)^\top} = \underbrace{[1, x, x^2, \ldots, x^{K-1}]}_{\mathbf{v}(x)^\top} \mathbf{T} \qquad (7)$$

for some $y \in \mathbb{R}$. We show how matrix $\mathbf{T}$ needs to be an upper (generalized) Pascal matrix [11] with parameters (t_0, t_1), denoted as $\mathbf{T}_{(t_0, t_1)}$; that is, a matrix of the form[2]:

$$\mathbf{T}_{(t_0,t_1)} := \begin{bmatrix} 1 & t_0 & t_0^2 & t_0^3 \\ 0 & t_1 & 2t_0t_1 & 3t_0^2t_1 \\ 0 & 0 & t_1^2 & 3t_0t_1^2 \\ 0 & 0 & 0 & t_1^3 \end{bmatrix} \qquad (8)$$

Theorem 1. *The equality $\mathbf{v}(y)^\top = \mathbf{v}(x)^\top \mathbf{T}$ [cf. (7)] holds for generic x, y if and only if $\mathbf{T} \in \mathbb{R}^{K \times K}$ is a Pascal matrix with parameters (t_0, t_1).*

Proof. $\Longleftarrow$) If $\mathbf{T}_{(t_0, t_1)}$ is Pascal with parameters t_0 and t_1, then it directly follows that $\mathbf{v}(x)^\top \mathbf{T}_{(t_0,t_1)} = \mathbf{v}(t_0 + t_1 x) = \mathbf{v}(y)^\top$.

$\Longrightarrow$) Consider $\mathbf{T}(:, 1) = [t_0, t_1, \ldots, t_{K-1}]^\top$ so that $y = t_0 + t_1 x + t_2 x^2 + \ldots + t_{K-1} x^{K-1}$, which models the second column of $\mathbf{v}(y)^\top$ as a polynomial of order $K - 1$ with coefficients given by the second column of $\mathbf{T}$. The last element of $\mathbf{v}(y)^\top$ is then given by:

$$y^{K-1} = (t_0 + t_1 x + t_2 x^2 + \ldots + t_{K-1} x^{K-1})^{K-1} \qquad (9)$$

whose highest order term of is $t_{K-1}^{K-1} x^{(K-1)^2}$. Since $x^{(K-1)^2}$ is not present in $\mathbf{v}(x)^\top$, t_{K-1} necessarily needs to be equal to zero. Next, all the coefficients associated to the monomial x^j, with $j > (K-2)(K-1)$, include t_{K-1}, which is thus also equal to zero. The next monomial which includes terms with a coefficient different from zero is $x^{(K-2)(K-1)}$, for which the only non-zero term is t_{K-2}. However, since such monomial is not present in $\mathbf{v}(x)^\top$, we conclude $t_{K-2} = 0$. With the same logic one can show that only t_0 and t_1, associated to the zeroth- and first-order monomial, can be different from zero. In other words, the entry (i, j) of $\mathbf{T}$ represents the coefficient associated to x^i when the binomial $y(x) = t_0 + t_1 x$ is raised to the power j, with i, j starting from zero. $\square$

Identifiability and solvability. One of the consequences of Theorem 1 is that, without additional constraints, every shifted and scaled

---

[2]We show $K = 4$ to ease the visualization.

version of the groundtruth parameter x (for an appropriate $\mathbf{W}$), perfectly fits the observation model [cf. (6)]. Thus the model is identifiable up to this ambiguity, which is sufficient in many contexts. For instance, in the context of (eigenvalues) graph learning [3], a shift and scale of the graph eigenvalues maintains the same topological structure of the original graph (removing the self loops caused by the shift). As we will see, identifiability alone (which is mainly a theoretical property) does not provide us a guarantee of solvability (meaning, global convergence).

3.1. Subspace Fitting
In a subspace fitting approach [12], we start by considering the following optimization problem to estimate $\mathbf{x}$ and $\mathbf{W}$:

$$\min_{\mathbf{x}, \mathbf{W}} \frac{1}{2} \|\mathbf{Y} - \mathbf{V}(\mathbf{x})\mathbf{W}\|_F^2, \tag{10}$$

which can be solved, for instance, with an alternating minimization approach, without guarantee of convergence to a global optimum.

To improve the conditioning of the problem and to reduce the estimation effort when $L \gg K$, we consider the economy-size SVD of matrix $\mathbf{Y}$, i.e., $\mathbf{Y} = \mathbf{U}\mathbf{\Sigma}\mathbf{Z}^\top$, where $\mathbf{U} \in \mathbb{R}^{N \times K}$ and $\mathbf{Z} \in \mathbb{R}^{L \times K}$ are the left and right singular vectors, respectively, and $\mathbf{\Sigma} \in \mathbb{R}^{K \times K}$ is the diagonal matrix of singular values. Since both $\mathbf{V}(\mathbf{x})$ and $\mathbf{U}$ represent a basis for the column space of $\mathbf{Y}$, there exists a non-singular matrix $\mathbf{S} \in \mathbb{R}^{K \times K}$ such that $\mathbf{V} = \mathbf{U}\mathbf{S}$. The subspace fitting problem reads then as:

$$\min_{\mathbf{x}, \mathbf{S}} \frac{1}{2} \|\mathbf{V}(\mathbf{x}) - \mathbf{U}\mathbf{S}\|_F^2, \tag{11}$$

which, upon substituting the pseudoinverse solution $\hat{\mathbf{S}} = \mathbf{U}^\dagger \mathbf{V}$ into (11), can be casted as the following equivalent problem:

$$\min_{\mathbf{x}} \left\{ f(\mathbf{x}) := \frac{1}{2} \|\mathbf{P}\mathbf{V}(\mathbf{x})\|_F^2 \right\}, \tag{12}$$

with $\mathbf{P} := \mathbf{I}_N - \mathbf{U}\mathbf{U}^\dagger$ the orthogonal projection matrix onto the orthogonal complement of $\mathbf{U}$. In other words, problem (12) aims to find a vector $\mathbf{x}$ such that the Vandermonde matrix $\mathbf{V}(\mathbf{x})$ is orthogonal to the subspace spanned by the orthogonal complement of $\mathbf{U}$.

Problem (11) is not convex in $\mathbf{x}$ due to the polynomial degree K (unless $K - 1 = 1$, i.e., the model in (3) is linear). To tackle the non-convexity of the problem, we resort to sequential convex programming (SCP) [13], a local optimization method that leverages convex optimization, where the non-convex portion of the problem is modeled by convex functions that are (at least locally) accurate. As in any non-convex problem, the initial starting point plays a big role; thus, if no prior information on the variable is given, a multi-starting point approach is advisable.

Remark 2. As mentioned in Remark 1, the nature of the problem and the formulation (12) share similarities with the MUSIC algorithm [10]. However, while in MUSIC every column of matrix $\mathbf{V}(\mathbf{x})$ (containing the steering vectors of the array manifold) depends solely on one scalar parameter and N independent 1-dimensional searches can be carried out, here each column contains all the N variables. An N-dimensional search is thus needed, which makes a "scanning" of the vector variable $\mathbf{x}$ infeasible, unless N is very small.

SCP. The general idea of SCP is to maintain, at each iteration r, an estimate $\mathbf{x}^r$ and a respective convex trust region $\mathcal{T}^r \subseteq \mathbb{R}^N$ over which we trust our solution to reside. The next solution $\mathbf{x}^{r+1}$ is then obtained by minimizing, over the defined trust region $\mathcal{T}^r$, a convex approximation $\hat{f}(\cdot)$ of $f(\cdot)$ around the previous estimate $\mathbf{x}^r$. In our

case, we define as trust region the set:

$$\mathcal{T}^r := \{\mathbf{x} \mid \|\mathbf{x} - \mathbf{x}^r\|_p^2 \leq \rho(r)\} \tag{13}$$

where $\rho : \mathbb{Z}_+ \to \mathbb{R}_{++}$ is an iteration-dependent mapping indicating the maximum admissible length of the convex p-norm ball in (13).

Next, a feasible intermediate iterate is found by:

$$\hat{\mathbf{x}} = \operatorname*{argmin}_{\mathbf{x} \in \mathcal{T}^{[r]}} f(\mathbf{x}^r) + \nabla_{\mathbf{x}} f(\mathbf{x}^r)^\top (\mathbf{x} - \mathbf{x}^r) \tag{14}$$

i.e., by minimizing the (convex) first-order Taylor approximation of $f(\cdot)$ around $\mathbf{x}^r$, with $\nabla_{\mathbf{x}} f(\cdot) \in \mathbb{R}^N$ the gradient of function $f(\cdot)$; see Appendix A for the explicit gradient computation. Notice that due to the non-convexity of $f(\cdot)$ [cf. (12)], its value $f(\hat{\mathbf{x}})$ at the new feasible estimate $\hat{\mathbf{x}}$ is not guaranteed to be lower than the value at $\mathbf{x}^r$. Thus, to find the optimal solution at iteration $(r + 1)$, we pick the best convex combination of $\mathbf{x}^r$ and $\hat{\mathbf{x}}$ achieving the lowest function value, i.e.,:

$$\alpha_r^\star = \operatorname*{argmin}_{\alpha_r} f(\alpha_r \mathbf{x}^r + (1 - \alpha_r)\hat{\mathbf{x}}), \quad \alpha \in (0, 1) \tag{15}$$

$$\mathbf{x}^{r+1} = \alpha_r^\star \mathbf{x}^r + (1 - \alpha_r^\star)\hat{\mathbf{x}}. \tag{16}$$

We stress that a zero fitting error of the cost function can be achieved only by the true unknown parameter and its model ambiguities, which as such represent global minima of the function and solutions of the problem.

4. NUMERICAL RESULTS
We now perform numerical simulations to assess the validity of the subspace approach, with focus on a practical case-study: clock-synchronization after sampling jitter.

Scenario. In the context of signal sampling, jitter is an undesirable effect caused by, for instance, interference of the clock signal, and consisting of a deviation from the true sampling pattern which we consider here to be periodic. When the time between samples varies an instantaneous signal error arises and a jitter-correction (clock-recovery) is desirable. To this extent, consider the time-interval $[u_0, u_1, \ldots, u_{N-1}]$ of uniformly sampled locations (timings) with period T, i.e., such that $u_n - u_{n-1} = T$ for $n = 1, \ldots, N-1$. Because of jitter, the actual sampling locations are given by $x_n = u_n + j_n$, with $j_n \sim \mathcal{N}(0, (\delta T/2)^2)$ a Gaussian distribution truncated at one standard deviation. The parameter δ specifies whether an overlap among adjacent samples x_n is possible ($\delta \geq 1$) or not ($\delta < 1$). Associated to each x_n is a signal value y_n corresponding to the sampling of an unknown function $f(x_n)$.

Metrics. To assess the validity of the proposed approach, recall that we can recover the solution of problem (12) up to a shift and scaling of the true positions [cf. Section 3]. Thus, we use the normalized error modulo Pascal (PNE) as performance metric, defined as:

$$\text{PNE}(\hat{\mathbf{x}}, \mathbf{x}) = \min_{t_0, t_1} \frac{\|\mathbf{x} - (t_0 \mathbf{1} + t_1 \hat{\mathbf{x}})\|_2}{NT} \tag{17}$$

which measures how far the true locations are from a linear transformation of the recovered estimates. Clearly (17) is zero whenever $\hat{\mathbf{x}}$ is a solution for (12).

Results. For a fixed number $N = 30$ of locations, we run the algorithm 1000 times to randomize the coefficient matrix $\mathbf{W}$ used to generate the polynomials, for different values of the polynomial order $K = \{3, 4\}$, number of polynomials $L = \{3, 4\}$, $L \geq K$, and parameter $\delta \in (1, 140)$. We focus our attention on quadratic ($K = 3$) and cubic ($K = 4$) polynomials because for higher degrees the polynomials are almost flat in the domain of interest and they tend to be highly sensitive to round-off due to the very ill-conditioning of the
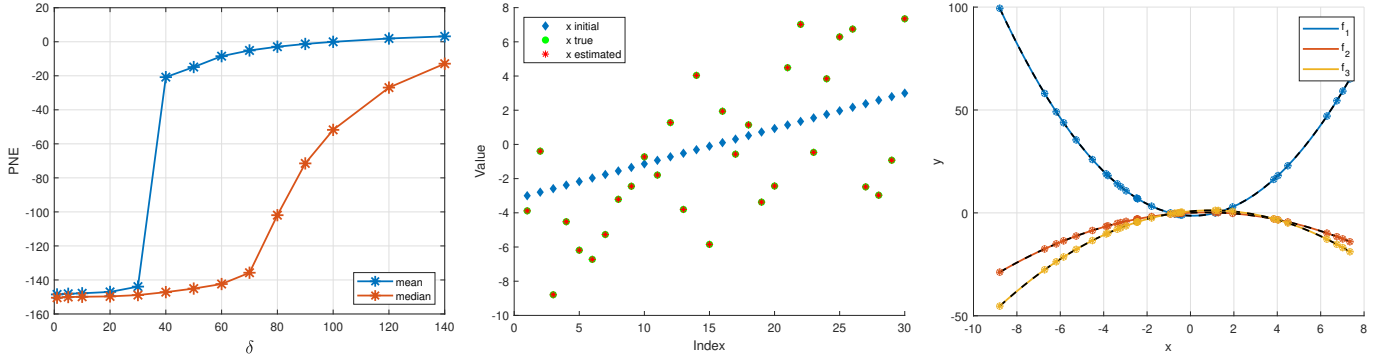
Fig. 2: For $K = L = 3$: (left) Average and median PNE in dB over 1000 runs of the algorithm; (middle) True locations (green circles), inferred locations (red asterisks) and uniform locations (blue diamonds); the latter is used as the starting point for the algorithm; (right) Original polynomials (solid lines) with associated function samples ($\circ$) in correspondence of the original uniform locations, and *inferred* polynomials (black dotted lines) with associated function samples ($*$) in correspondence of the *inferred* locations.

Vandermonde matrix. As function domain we consider $[-3, 3]$, and discretize it with a sampling period of $T = 6/(N-1)$. We initialize the starting point of the algorithm with the uniform sampled domain, i.e., $\mathbf{x}^0 = [u_0, \dots, u_{N-1}]^\top$.

In Fig. 2 (left) is shown (in dB) the average and median PNE over 1000 simulations for $K = L = 3$ and δ ranging from 1 to 140. The first remarkable result is that the algorithm is able to perfectly recover the solution (up to a shift and scaling) without ending in a local minimum for very high values of δ, resulting in big overlaps among the samples. In particular, both the mean and the median are zero up to $\delta = 30$, which is the setting for which one sample can overlap with 15 of its nearby samples. The median, which is more robust than the mean in terms of sensitivity to outliers, is basically zero up to $\delta = 60$, i.e., the setting for which a sample can overlap with all the others, rendering it almost a random configuration of points. Nonetheless, also for the extreme case of $\delta = 140$, the median still achieves an error of $1/100$th of a sampling period, which is an acceptable error. To enjoy the algorithm's performance, in Fig. 2 (middle) we show for $K = L = 3$ and $\delta = 60$, the estimated locations (red, after ambiguity correction), together with the true locations (green) and the initial starting point of the algorithm (blue). Correspondingly, in Fig. 2 (right) we show the original and inferred polynomials, constructed as $\hat{\mathbf{Y}} = \mathbf{V}(\hat{\mathbf{x}})\mathbf{V}(\hat{\mathbf{x}})^\dagger \mathbf{Y}$.

5. CONCLUSION

In this work, we studied the problem of polynomial regression when the coefficients and the regressor are both unknown. We have shown how the model is identifiable up to a generalized Pascal transformation, responsible to scale and shift the input parameter, which is sufficient in many recovery problems. We then proposed an algorithmic routine to tackle the problem based on a subspace fitting approach. Finally, we performed numerical simulations in the context of sampling jitter and show the effectiveness of the proposed approach to recover the original sample locations. Interesting future research directions include: *i)* studying the influence of input perturbation (δ) and polynomial order (K) on the reconstruction error for jitter scenarios; *ii)* understanding the recoverability property in case of noise in the output; *iii)* studying the problem under a purely (structured) matrix factorization approach with a Vandermonde factor.

A. GRADIENT COMPUTATION

Consider the function $f(\mathbf{x}) = \|\mathbf{P}\mathbf{V}(\mathbf{x})\|_F^2$ and define the composite matrix function $\mathbf{M}(\mathbf{V}(\mathbf{x})) = \mathbf{P}\mathbf{V}(\mathbf{x})$. We use the chain rule and to compute the gradient of $f(\cdot)$ w.r.t. $\mathbf{x}$, i.e., $\partial f(\cdot)/\partial \mathbf{x}$. We have $f(\mathbf{M}) = \frac{1}{2}\text{tr}(\mathbf{M}^\top \mathbf{M}) = \frac{1}{2}\mathbf{M} : \mathbf{M}$. The differential of $f(\cdot)$ is

$df(\mathbf{M}, d\mathbf{M}) = \mathbf{M} : d\mathbf{M} = \mathbf{M} : \mathbf{P}d\mathbf{V}(\mathbf{x})$. For the differential $d\mathbf{V}(\mathbf{x}, d\mathbf{x})$ notice that, by selecting one column, e.g., $\mathbf{x}^3$, we have $(\mathbf{x} + d\mathbf{x})^3 = \mathbf{x}^3 + 3\mathbf{x}^2 d\mathbf{x} + \mathcal{O}(d\mathbf{x}^2)$, and more generally:

$$d\mathbf{V}(\mathbf{x}) = [\mathbf{0} \; \mathbf{1} \dots (K-1)\mathbf{x}^{K-2}] \odot [d\mathbf{x} \; d\mathbf{x} \dots d\mathbf{x}] \quad (18)$$

$$= (\mathbf{V}(\mathbf{x})\mathbf{D}_k) \odot (d\mathbf{x}\mathbf{1}^\top) = \text{Diag}(d\mathbf{x})(\mathbf{V}(\mathbf{x})\mathbf{D}_k) \quad (19)$$

where $\mathbf{D}_k = \text{SupDiag}(1, \dots, K-1) \in \mathbb{R}^{K \times K}$ is the differentiation matrix for polynomials and $\odot$ is the element-wise product. By plugging this differential into the previous expression for df we obtain:

$$df(\mathbf{M}, d\mathbf{M}) = \mathbf{M} : \mathbf{P}d\mathbf{V}(\mathbf{x}) = \mathbf{M} : \mathbf{P}\text{Diag}(d\mathbf{x})(\mathbf{V}(\mathbf{x})\mathbf{D}_k)$$

$$= \mathbf{P}^\top \mathbf{M}(\mathbf{V}(\mathbf{x})\mathbf{D}_k)^\top : \text{Diag}(d\mathbf{x}) = \text{diag}(\mathbf{P}^\top \mathbf{M}(\mathbf{V}(\mathbf{x})\mathbf{D}_k)^\top) : d\mathbf{x}$$

We conclude $\partial f(\mathbf{x})/\partial \mathbf{x} = \text{diag}(\mathbf{P}^\top \mathbf{P}\mathbf{V}(\mathbf{x})\mathbf{D}_k^\top \mathbf{V}(\mathbf{x})^\top)$.

B. REFERENCES

[1] Steven M Kay, *Fundamentals of statistical signal processing: estimation theory*, Prentice-Hall, Inc., 1993.

[2] David I Shuman, Sunil K Narang, Pascal Frossard, Antonio Ortega, and Pierre Vandergheynst, "The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains," *IEEE signal processing magazine*, vol. 30, no. 3, pp. 83–98, 2013.

[3] Alberto Natali and Geert Leus, "A general convolution theorem for graph data," in *2022 55th Asilomar Conference on Signals, Systems, and Computers*. IEEE, p. To appear.

[4] Xiao Fu, Kejun Huang, Nicholas D Sidiropoulos, and Wing-Kin Ma, "Nonnegative matrix factorization for signal and data analytics: Identifiability, algorithms, and applications.," *IEEE Signal Process. Mag.*, vol. 36, no. 2, pp. 59–80, 2019.

[5] Pina Marziliano and Martin Vetterli, "Reconstruction of irregularly sampled discrete-time bandlimited signals with unknown sampling locations," *IEEE Transactions on Signal Processing*, vol. 48, no. 12, pp. 3462–3471, 2000.

[6] John Browning, "Approximating signals from nonuniform continuous time samples at unknown locations," *IEEE transactions on signal processing*, vol. 55, no. 4, pp. 1549–1554, 2007.

[7] Golnoosh Elhami, Michalina Pacholska, Benjamín Béjar Haro, Martin Vetterli, and Adam Scholefield, "Sampling at unknown locations: Uniqueness and reconstruction under constraints," *IEEE Transactions on Signal Processing*, vol. 66, no. 22, pp. 5862–5874, 2018.

[8] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman, *The elements of statistical learning: data mining, inference, and prediction*, vol. 2, Springer, 2009.

[9] Pei-Jung Chung, Mats Viberg, and Jia Yu, "Doa estimation methods and algorithms," in *Academic Press Library in Signal Processing*, vol. 3, pp. 599–650. Elsevier, 2014.

[10] Ralph Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE transactions on antennas and propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[11] Gregory S. Call and Daniel J. Velleman, "Pascal's matrices," *The American Mathematical Monthly*, vol. 100, no. 4, pp. 372–376, 1993.

[12] Mats Viberg and Bjorn Ottersten, "Sensor array processing based on subspace fitting," *IEEE Transactions on signal processing*, vol. 39, no. 5, pp. 1110–1121, 1991.

[13] Stephen Boyd, "Sequential convex programming," *Lecture Notes, Stanford University*, 2008.