A Proof of the VARiable PROjection (VARPRO) Method in Hilber Space

Qiang Ning

Department of Electrical and Computer Engineering

Beckman Institute for Advanced Science and Techonology

University of Illinois at Urbana-Champaign

Abstract

The variable projection (VARPRO) method deals with those nonlinear least squares problems whose variables separate into two categories: linear variables and nonlinear variables. This report focuses on the original paper proposing the VARPRO method, corrects its typos and generalizes the proof to infinite dimensional spaces and complex numbers. One key proposition that the author used without proof in the paper is also proved here. We also implemented the VARPRO method based on Gauss-Newton algorithm and the corresponding experimental results are shown and discussed in this report.

I. Introduction

In model-based estimation problems, we need to estimate the parameters, $\boldsymbol{\alpha}$, in the proposed model, $\eta(\boldsymbol{\alpha};t)$, given the observation data $y(t)$. One of the commonly used penalty functions that describe the loyalty to the observation data is the so-called least squares function,

$$r(\boldsymbol{\alpha};t) = y(t) - \eta(\boldsymbol{\alpha};t), \tag{1}$$

where

$$\boldsymbol{\alpha} = [\alpha_1, \alpha_2, ..., \alpha_k]^T \in \mathbb{R}^k,$$

and we usually obtain an estimation for $\boldsymbol{\alpha}$ by maximizing the 2-norm of the least squares function,

$$\hat{\boldsymbol{\alpha}}(\boldsymbol{y}) = \arg\min_{\boldsymbol{\alpha}} \|\boldsymbol{r}(\boldsymbol{\alpha})\|_2, \tag{2}$$

where

$$\boldsymbol{y} = [y(t_1), \ldots, y(t_m)]^T \in \mathbb{R}^m,$$

$$\boldsymbol{r}(\boldsymbol{\alpha}) = [r(\boldsymbol{\alpha}; t_1), \ldots, r(\boldsymbol{\alpha}; t_m)]^T \in \mathbb{R}^m.$$

Besides describing the discrepancy between observation and the proposed model, least squares function also has another property. When the observation data are corrupted by Gaussian noise, i.e.,

$$y(t) = \eta(\boldsymbol{\alpha}; t) + \xi(t), \tag{3}$$

where $\xi(t) \sim \mathcal{N}(0, \sigma^2)$ stands for a Gaussian noise with variance σ^2, then the maximum likelihood estimator is

$$\begin{aligned} \hat{\boldsymbol{\alpha}}_{ML}(\boldsymbol{y}) &= \arg\max_{\boldsymbol{\alpha}} \ln f_{Y|\alpha}(\boldsymbol{y}|\boldsymbol{\alpha}) \\ &= \arg\max_{\boldsymbol{\alpha}} \ln \prod_i \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y(t_i)-\eta(\boldsymbol{\alpha};t_i))}{2\sigma^2}} \\ &= \arg\min_{\boldsymbol{\alpha}} \|\boldsymbol{r}(\boldsymbol{\alpha})\|_2. \end{aligned} \tag{4}$$

Therefore, the least squares estimator gives exactly the same result as the maximum likelihood estimator in this case. As the maximum likelihood estimator has some attractive properties such as asymptotically unbiased and asymptotically efficient (i.e., its variance of error asymptotically reaches its Cramer-Rao Lower Bound) [1], the least squares problem is very useful and ubiquitous in real applications if there is no prior information available and there are no other assumptions made.

If the model we proposed above, $\boldsymbol{\eta}(\cdot)$, is linear with respect to $\boldsymbol{\alpha}$, then the least squares problem turns into a linear least squares problem, which has already been well-defined and solved in Hilbert spaces. If $\boldsymbol{\eta}(\boldsymbol{\alpha}) = A\boldsymbol{\alpha}$, where A represents the linear mapping $\boldsymbol{\eta}$, then the least squares solution must satisfy

$$A^* A \hat{\boldsymbol{\alpha}}_{LS} = A^* \boldsymbol{y}, \tag{5}$$

where A^* is the adjoint of A which exists and is unique by Lemma 7.3 (Textbook [2]). As long as the range space of $\boldsymbol{\eta}$ or A is complete, the least squares solution exists (Theorem 7.13 Textbook [2]). Or, if we are more interested in the least squares solution that has the minimum norm, then we can further projecting any least squares solution to the orthogonal complement of the null space of $\boldsymbol{\eta}$, which has also been solved in our class.

However, in many cases such as harmonic retrieval and direction of arrival problems, the model $\boldsymbol{\eta}(\boldsymbol{\alpha})$ is unfortunately not linear with respect to its parameter $\boldsymbol{\alpha}$. Although people have developed many state space methods, for example, Prony's method, Kumaresan and Tufts method, and the Multiple Signal Classification (MUSIC) method, all of these methods are specifically designed for one type of applications and lacks flexibility to other problems. Therefore, solving the nonlinear least squares problem is also meaningful under many occasions.

In the remainder of this report, we are going to focus on one kind of nonlinear least squares problems whose variables separate into two categories: linear variables and nonlinear variables. The method is called the Variable Projection (VARPRO) method and was proposed by Golub and Pereyra in 1973 [3]. Section II will be the formulation of the problems that VARPRO handles and will be specifying the infinite dimensional Hilbert space we are dealing with. The following section III will be a detailed proof of the theorems and lemmas from the paper [3], with extensions to the infinite dimensional case. We have also implemented the VARPRO method using the Gauss-Newton algorithm as described by section IV, with an extension from real number case to complex number case. And the experimental results are shown in section V. Finally an analysis of the results and a more general discuss over the VARPRO method will be provided in section VI.

II. Formulation

Those nonlinear least squares problems whose variables separate are such a kind of problems:

$$(\hat{\boldsymbol{c}}(\boldsymbol{y}), \hat{\boldsymbol{\alpha}}(\boldsymbol{y})) = \arg\min_{\boldsymbol{c},\boldsymbol{\alpha}} \|\boldsymbol{r}(\boldsymbol{c},\boldsymbol{\alpha})\|_2, \tag{6}$$

where

$$\boldsymbol{\alpha} = [\alpha_1, \alpha_2, ..., \alpha_k]^T \in \mathbb{R}^k,$$

$$\boldsymbol{c} = [c_1, c_2, ..., c_n]^T \in \mathbb{R}^n,$$

$$\boldsymbol{y} = [y(t_1), \dots, y(t_m)]^T \in \mathbb{R}^m,$$

$$\boldsymbol{r}(\boldsymbol{c},\boldsymbol{\alpha}) = [r(\boldsymbol{c},\boldsymbol{\alpha}; t_1), \dots, r(\boldsymbol{c},\boldsymbol{\alpha}; t_m)]^T \in \mathbb{R}^m,$$

and

$$r(\boldsymbol{c},\boldsymbol{\alpha}; t) = y(t) - \eta(\boldsymbol{c},\boldsymbol{\alpha}; t). \tag{7}$$

Note here we have $0 < n < \infty$ and $0 < k, m \leq \infty$, which means that we have generalized VARPRO to cope with infinite nonlinear variables $\boldsymbol{\alpha}$ and infinite observations $\boldsymbol{y}$.

Here we specifically replace the original parameter vector $\boldsymbol{\alpha}$ in (1) and (2) by $\boldsymbol{c}$ and $\boldsymbol{\alpha}$ is to explicitly show that the variables separate into two categories, where $\boldsymbol{c}$ represents the linear variables and $\boldsymbol{\alpha}$ represents those nonlinear variables. To be more precise, the model $\boldsymbol{\eta}(\cdot)$ has the following form:

$$\eta(\boldsymbol{c}, \boldsymbol{\alpha}; t) = \sum_{j=1}^{n} c_j \varphi_j(\boldsymbol{\alpha}; t), \tag{8}$$

where $\varphi_j(\boldsymbol{\alpha}; t), j = 1, 2, \ldots, n$ and $n < \infty$ is a set of basis functions.

If we rewrite Eqn. (7) in vector form, we have

$$\boldsymbol{r}(\boldsymbol{c}, \boldsymbol{\alpha}) = \boldsymbol{y} - \boldsymbol{\Phi}(\boldsymbol{\alpha})\boldsymbol{c}, \tag{9}$$

where $\boldsymbol{\Phi}(\boldsymbol{\alpha}) \in \mathbb{R}^{m \times n}$ and $\boldsymbol{\Phi}_{ij} = \varphi_j(\boldsymbol{\alpha}; t_i)$, for $i = 1, 2, \ldots, m,\ j = 1, 2, \ldots, n$.

Given Eqn. (6), we want to set this problem into the scope of vector spaces, such that we can use what we have developed for vector spaces to solve this least squares problem. It is feasible if we already know the value of $\boldsymbol{\alpha}$ such that the basis functions $\varphi(\boldsymbol{\alpha}; t)$ are fully determined. We can project the observation $\boldsymbol{y}$ onto the span of $\varphi(\cdot)$'s and obtain $\hat{\boldsymbol{c}}$.

To be more rigorous, we need to firstly define a Hilbert space where these observations lie in:

$$\mathcal{X} = \{l_2(\mathbb{Z}^+), < \cdot, \cdot >\}, \tag{10}$$

where the inner product $< \cdot, \cdot >$ is defined as following:

$$< \boldsymbol{u}, \boldsymbol{v} >= \sum_{j=0}^{\infty} u[j]v[j], \forall \boldsymbol{u}, \boldsymbol{v} \in \mathcal{X}. \tag{11}$$

Because the signals in real world must have finite energy, the so-defined $\mathcal{X}$ indeed contains the observation data $\boldsymbol{y}$. Note here the dimension of $\boldsymbol{y}$, which is m as we mentioned before, is not necessarily finite. But it is obvious that $\mathcal{X}$ is a Hilbert space (Theorem 6.2 Textbook [2]), and the inner product we defined is valid for real valued data.

If the parameter vector $\boldsymbol{\alpha}$ is unknown and the order of the model n is fixed, then the set of all the linear combinations of such φ's does not form a subspace at all. If the order n is not fixed but finite, then the set is a subspace but not complete, for the same reason as the set of all finite length discrete time signals. Neither of these two cases is desirable because we cannot construct a linear projector and solve the linear variables $\boldsymbol{c}$ on these two sets (Theorem 7.6 Textbook [2]).

However, if the parameter vector $\boldsymbol{\alpha}$ is available somehow and the model order n is fixed, then the span of the basis functions $\varphi_j(\boldsymbol{\alpha}; t), j = 1, 2, \ldots, n$ form an n dimensional subspace $\mathcal{S} \subset \mathcal{X}$. Because $\mathcal{S}$ is finite dimensional, it is obviously complete. By Theorem 7.6 (Textbook [2]), we know that there exists a projector, $P_{\mathcal{S}}$, from $\mathcal{X}$ onto the subspace $\mathcal{S}$. By applying this projector to our observation $\boldsymbol{y}$, we would obtain a unique $P_{\mathcal{S}}\boldsymbol{y} \in \mathcal{S}$ and its coordinate under basis $\varphi_j, j = 1, \ldots, n$ is exactly the desired $\hat{\boldsymbol{c}}$. This provides an intuition for the VARPRO method.

The VARPRO method is to replace (9) by

$$\boldsymbol{r}_2(\boldsymbol{\alpha}) = \boldsymbol{y} - \boldsymbol{\Phi}(\boldsymbol{\alpha})\boldsymbol{\Phi}^{\dagger}(\boldsymbol{\alpha})\boldsymbol{y}. \tag{12}$$

And the original problem (6) is turned into the following two-step problem (which we call VARPRO in the following):

$$\hat{\boldsymbol{\alpha}}(\boldsymbol{y}) = \arg\min_{\boldsymbol{\alpha}} \|\boldsymbol{r}_2(\boldsymbol{\alpha})\|_2, \tag{13}$$

$$\hat{\boldsymbol{c}}(\boldsymbol{y}) = \boldsymbol{\Phi}^{\dagger}(\hat{\boldsymbol{\alpha}})\boldsymbol{y}, \tag{14}$$

where the last equation comes from Definition 7.3 (Definition of pseudo-inverse in Hilbert spaces) from our textbook [2].

An immediate concern would be whether VARPRO is equivalent to its original problem, which the paper [3] was focused on.

Note from the analysis above, we can see that the VARPRO does not cope with the situation where the model order n is variant (otherwise the projection cannot be made). This is to say that VARPRO cannot be generalized to deal with the infinite dimensional model $\eta(\cdot; t)$ where n is not fixed but finite. However, we can see later that VARPRO can indeed be generalized such that the nonlinear variables and the observations can be infinite dimensional.

III. Theorems and Lemmas

A. Frechet Derivative

The proof of the theorems from the paper [3] are based on a generalization of derivative to Banach spaces, the Frechet derivative.

Let V and W be two Banach spaces. Note here V and W are not necessarily finite dimensional. The only requirement is that they are complete normed spaces. Take a function $f : V \mapsto W$.

It is Frechet differentiable at $x \in V$ if and only if there exists a bounded, linear operator $A_x : V \mapsto W$, such that

$$\lim_{h \to 0} \frac{\|f(x+h) - f(x) - A_x(h)\|_W}{\|h\|_V} = 0,$$

and we denote A_x by $Df(x)$

This definition is a generalization to Banach space derivatives. To see this, if we take $V = \mathbb{R}^p$ and $W = \mathbb{R}$, and take the norm in both V and W to be the 2-norm, then by the Taylor expansion of this multiple variables function f, we have

$$f(\boldsymbol{x} + \boldsymbol{h}) = f(\boldsymbol{x}) + \nabla f_{\boldsymbol{x}} \boldsymbol{h} + \frac{1}{2} \boldsymbol{h}^T H_{\boldsymbol{x}} \boldsymbol{h} + o(\|\boldsymbol{h}\|_V^2),$$

where ∇f denotes the gradient vector of f and H denotes the Hessian matrix of f. Then apparently the Frechet derivative at point x, A_x, is the same with ∇f_x.

If we change W to $\mathbb{R}^q$ and remains $V = \mathbb{R}^p$ and the norm definition to be the same, we can still obtain the Taylor expansion for the vector function $\boldsymbol{f}$:

$$\boldsymbol{f}(\boldsymbol{x} + \boldsymbol{h}) = \boldsymbol{f}(\boldsymbol{x}) + J_{\boldsymbol{x}} \boldsymbol{h} + o(\|\boldsymbol{h}\|_V),$$

and $D\boldsymbol{f}(\boldsymbol{x}) = J_{\boldsymbol{x}}$, where J is the Jacobian matrix of $\boldsymbol{f}$.

One important proposition that the author repeatedly used without proof in [3] is:

$$D(AB) = (DA)B + A(DB), \tag{15}$$

where A and B are matrices.

However, this is counter-intuitive, in fact. Because the product of two matrices represents a composition of two linear mappings. For example, if A and B are respectively the matrix form of mapping $\mathcal{A}$ and mapping $\mathcal{B}$, then AB is the matrix representation of $\mathcal{A} \circ \mathcal{B}$. It can be easily found in any materials of Frechet derivative that the chain rule of Frechet derivative is

$$D(\mathcal{A} \circ \mathcal{B})(x) = D\mathcal{A}(\mathcal{B}(x)) D\mathcal{B}(x), \tag{16}$$

which is contradictory to (15). This difference comes from the fact that we did not clarify the two Banach spaces V and W where the Frechet derivatives were defined with. The Banach spaces associated with (15) and (16) were actually different, which is why we got different results above.

In the derivation of [3], the author frequently used Eqn. (15). In that case, the two Banach spaces are defined as following:

$$V = \mathbb{R}^k, W = \mathbb{R}^{m \times m},$$

Note V is the nonlinear variable vector space. Because we can have infinitely many nonlinear variables to estimate ($k \leq \infty$), V is an infinite dimensional space and needs to check against completeness. W, which is a matrix vector space, is $m \times m$ and can also be infinite dimensional ($m \leq \infty$). Instead of immediately telling whether V and W are complete, we need to define norms associated with them:

$$\|\boldsymbol{v}\| = \left(\sum_{j=1}^{k} v^2[j] \right)^{\frac{1}{2}}, \forall \boldsymbol{v} \in V,$$

$$\|\boldsymbol{w}\| = \left(\sum_{i=1}^{m} \sum_{j=1}^{m} w^2[i,j] \right)^{\frac{1}{2}}, \forall \boldsymbol{w} \in W.$$

If we further assume that the norms defined above are bounded, then V and W are just $l_2(\mathbb{Z}^+)$ and $l_2(\mathbb{Z}^+ \times \mathbb{Z}^+)$ and apparently Banach spaces (Theorem 6.2 Textbook [2]).

Next we try to prove (15) as the following proposition by ourselves. This is not trivial because it involves infinite dimensional vector spaces and completely different from the property of conventional differentiation (however, the proof of $dxy = ydx + xdy$ indeed provides intuition).

Proposition 1 (Property of Frechet Derivatives): Let $\mathcal{A}$ and $\mathcal{B}$ be two mappings from V to W, where V and W are defined above as Banach spaces. Then we have

$$D(\mathcal{A}(\boldsymbol{v})\mathcal{B}(\boldsymbol{v})) = (D\mathcal{A}(\boldsymbol{v}))\mathcal{B}(\boldsymbol{v}) + \mathcal{A}(\boldsymbol{v})(D\mathcal{B}(\boldsymbol{v})) \tag{17}$$

for all $\boldsymbol{v} \in V$.

Proof: For all $\boldsymbol{v} \in V$, let $A(\boldsymbol{v}) = \mathcal{A}(\boldsymbol{v}) \in W$, $B(\boldsymbol{v}) = \mathcal{B}(\boldsymbol{v}) \in W$ and $C(\boldsymbol{v}) = A(\boldsymbol{v})B(\boldsymbol{v})$. Then,

$$C(\boldsymbol{v})[i,j] = \sum_{l=1}^{m} A(\boldsymbol{v})[i,l]B(\boldsymbol{v})[l,j].$$

By the definition of norm on W, we have

$$\begin{aligned} & \|C(\boldsymbol{v}+\boldsymbol{h}) - C(\boldsymbol{v}) - DC(\boldsymbol{v})(\boldsymbol{h})\|_W \\ = & \ [\sum_{i=1}^{m} \sum_{j=1}^{m} (C_{ij}(\boldsymbol{v}+\boldsymbol{h}) - C_{ij}(\boldsymbol{v}) \\ & -DC_{ij}(\boldsymbol{v})(\boldsymbol{h}))^2]^{\frac{1}{2}} \end{aligned} \tag{18}$$

Assume $DC(\boldsymbol{v})$ satisfies (17), then,

$$\begin{aligned} & C_{ij}(\boldsymbol{v}+\boldsymbol{h}) - C_{ij}(\boldsymbol{v}) - DC_{ij}(\boldsymbol{v})(\boldsymbol{h}) \\ = & \sum_{l=1}^{m} A_{il}(\boldsymbol{v}+\boldsymbol{h})B_{lj}(\boldsymbol{v}+\boldsymbol{h}) \\ & -A_{il}(\boldsymbol{v})B_{lj}(\boldsymbol{v}) \\ & -(DA_{il})(\boldsymbol{v})B_{lj}(\boldsymbol{v})\boldsymbol{h} \\ & -A_{il}(\boldsymbol{v})(DB_{lj})(\boldsymbol{v})\boldsymbol{h} \\ = & \sum_{l=1}^{m} A_{il}(\boldsymbol{v}+\boldsymbol{h})(B_{lj}(\boldsymbol{v}+\boldsymbol{h}) - B_{lj}(\boldsymbol{v}) + B_{lj}(\boldsymbol{v})) \\ & -A_{il}(\boldsymbol{v})B_{lj}(\boldsymbol{v}) \\ & -(\nabla_{il}A(\boldsymbol{v}))B_{lj}(\boldsymbol{v})\boldsymbol{h} \\ & -(A_{il}(\boldsymbol{v}) - A_{il}(\boldsymbol{v}+\boldsymbol{h}) + A_{il}(\boldsymbol{v}+\boldsymbol{h}))(\nabla_{lj}B(\boldsymbol{v}))\boldsymbol{h} \\ = & \sum_{l=1}^{m} A_{il}(\boldsymbol{v}+\boldsymbol{h})(B_{lj}(\boldsymbol{v}+\boldsymbol{h}) - B_{lj}(\boldsymbol{v}) - \nabla_{lj}B(\boldsymbol{v})\boldsymbol{h}) \\ & +(A_{il}(\boldsymbol{v}+\boldsymbol{h}) - A_{il}(\boldsymbol{v}) - \nabla_{il}A(\boldsymbol{v})\boldsymbol{h})B_{lj}(\boldsymbol{v}) \\ & -(A_{il}(\boldsymbol{v}) - A_{il}(\boldsymbol{v}+\boldsymbol{h}))\nabla_{lj}B(\boldsymbol{v})\boldsymbol{h} \\ = & \sum_{l=1}^{m} o(\|\boldsymbol{h}\|_V), \end{aligned} \tag{19}$$

This is a sum of infinitely many $o(\cdot)$'s, but since elements in W all has finite norms, it is easy to see that

$$\frac{\|C(\boldsymbol{v}+\boldsymbol{h}) - C(\boldsymbol{v}) - DC(\boldsymbol{v})(\boldsymbol{h})\|_W}{\|\boldsymbol{h}\|_V} \to 0,$$

still holds, where $DC(\boldsymbol{v})$ satisfies (17).

■

B. Equivalence

One of the major results of [3] is the following theorem. Here we will go through it and fill up its missing steps in the original paper.

Theorem 1 (Equivalence): Let $\boldsymbol{r}(\boldsymbol{c}, \boldsymbol{\alpha})$ and $\boldsymbol{r}_2(\boldsymbol{\alpha})$ be defined as above. Assume that in an open set $\Omega \subset \mathbb{R}^k$, $\boldsymbol{\Phi}(\boldsymbol{\alpha})$ has constant rank $r \leq \min(m, n)$.

a) If $\boldsymbol{\alpha}$ is a critical point (or a global minimizer in Ω) of $\boldsymbol{r}_2(\boldsymbol{\alpha})$, and

$$\hat{\boldsymbol{c}} = \boldsymbol{\Phi}^{\dagger}(\hat{\boldsymbol{\alpha}})\boldsymbol{y}, \tag{20}$$

then $(\hat{\boldsymbol{c}}, \hat{\boldsymbol{\alpha}})$ is a critical point of $\boldsymbol{r}(\boldsymbol{c}, \boldsymbol{\alpha})$ (or a global miminizer for $\boldsymbol{\alpha} \in \Omega$) and $\boldsymbol{r}(\hat{\boldsymbol{c}}, \hat{\boldsymbol{\alpha}}) = \boldsymbol{r}_2(\hat{\boldsymbol{\alpha}})$.

b) If $(\hat{\boldsymbol{c}}, \hat{\boldsymbol{\alpha}})$ is a global minimizer of $\boldsymbol{r}(\boldsymbol{c}, \boldsymbol{\alpha})$ for $\boldsymbol{\alpha} \in \Omega$, then $\hat{\boldsymbol{\alpha}}$ is a global minimizer of $\boldsymbol{r}_2$ in Ω and $\boldsymbol{r}_2(\hat{\boldsymbol{\alpha}}) = \boldsymbol{r}(\hat{\boldsymbol{c}}, \hat{\boldsymbol{\alpha}})$. Furthermore, if there is a unique $\hat{\boldsymbol{c}}$ among the minimizing pairs of $\boldsymbol{r}(\boldsymbol{c}, \boldsymbol{\alpha})$, then $\hat{\boldsymbol{c}}$ must satisfy (20).

Before proving Theorem 1, we need to firstly prove a lemma, which was also provided by [3].

Lemma 1: Let $A^{-}(\boldsymbol{\alpha})$ be an $n \times m$ matrix function such that $AA^{-}A = A$ and $(AA^{-})^T = AA^{-}$. Then

$$DP_A = P_{A^{\perp}}(DA)A^{-} + (P_{A^{\perp}}(DA)A^{-})^T. \tag{21}$$

In order to generalize this lemma to infinite dimensional spaces, we have to require in addition that the range space of A is complete, and the whole space containing A should be Hilbert.

Although we have an additional requirement for A, this is easily satisfied in our scope because the range space of $\boldsymbol{\Phi}$ is exactly $\mathcal{X}$ which is indeed complete as we defined in section II.

Proof of Lemma 1: Since P_A is a projector onto the subspace of A, we have

$$P_A A = A,$$

by Theorem 7.4 i) (Textbook [2]). Therefore,

$$\begin{aligned} DA &= D(P_A A) = (DP_A)A + P_A(DA), \\ (DP_A)A &= DA - P_A(DA) = P_{A^{\perp}}(DA). \end{aligned} \tag{22}$$

We can see that the A^{-} here is a generalization or relaxation of the moore-penrose pseudo-inverse of A. It can be verified that AA^{-} is also a valid projector onto the subspace of A, because it satisfies the orthogonality principle. Moreover, because of the uniqueness of a projector, we have $P_A = AA^{-}$.

Because one projector is idempotent, we have $P_A^2 = P_A$ (Theorem 7.10 Textbook [2]), and then

$$\begin{aligned} DP_A &= D(P_A^2) \\ &= (DP_A)P_A + ((DP_A)P_A)^T \\ &= (DP_A)AA^- + ((DP_A)AA^-)^T \end{aligned} \tag{23}$$

By using (22), we have the result of (21). ■

As $P_{A^\perp} = I - P_A$ (Lemma 7.1 Textbook [2]), we easily have $DP_{A^\perp} = -DP$, which is useful in the proof following. And we should also notice that if A^- is replaced by $A^\dagger$ in (21), the result remains unchanged.

Proof of Theorem 1:

a) As we only consider an open set Ω, the critical points are only those with zero gradient. Hence, to show that these two problems' critical points coincide, we need to find their corresponding gradients to begin with.

Since $\|\boldsymbol{r}_2(\boldsymbol{\alpha})\|_2^2 = \|P_{\boldsymbol{\Phi}^\perp(\boldsymbol{\alpha})}\boldsymbol{y}\|_2^2$, by the chain rule of gradient, we have

$$\begin{aligned} \frac{1}{2}\nabla\|\boldsymbol{r}_2\|_2^2 &= \boldsymbol{y}^T P_{\boldsymbol{\Phi}^\perp}(DP_{\boldsymbol{\Phi}^\perp})\boldsymbol{y} \\ &= -\boldsymbol{y}^T P_{\boldsymbol{\Phi}^\perp}[P_{\boldsymbol{\Phi}^\perp}(D\boldsymbol{\Phi})\boldsymbol{\Phi}^\dagger \\ &\quad +(P_{\boldsymbol{\Phi}^\perp}(D\boldsymbol{\Phi})\boldsymbol{\Phi}^\dagger)^T]\boldsymbol{y}. \end{aligned} \tag{24}$$

Because $P_{\boldsymbol{\Phi}^\perp(\boldsymbol{\alpha})}(\boldsymbol{\Phi}^\dagger)^T = (\boldsymbol{\Phi}^\dagger)^T - \boldsymbol{\Phi}\boldsymbol{\Phi}^\dagger(\boldsymbol{\Phi}^\dagger)^T = 0$, the expression above can be simplified to

$$\frac{1}{2}\nabla\|\boldsymbol{r}_2(\boldsymbol{\alpha})\|_2^2 = -\boldsymbol{y}^T P_{\boldsymbol{\Phi}^\perp}(D\boldsymbol{\Phi})\boldsymbol{\Phi}^\dagger\boldsymbol{y}. \tag{25}$$

Similarly, for $\|\boldsymbol{r}(\boldsymbol{c},\boldsymbol{\alpha})\|_2^2$, we also have

$$\frac{1}{2}\nabla\boldsymbol{r}(\boldsymbol{c},\boldsymbol{\alpha}) = -(\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{c})^T((D\boldsymbol{\Phi})c \oplus \boldsymbol{\Phi}),$$

where the $\oplus$ sign denote direct sum, which comes from taking gradient with respect to $\boldsymbol{\alpha}$ and $\boldsymbol{c}$, respectively.

Because $P_{\boldsymbol{\Phi}^\perp}\boldsymbol{\Phi} = 0$, the expression above can be simplified to

$$\frac{1}{2}\nabla\boldsymbol{r}(\boldsymbol{\Phi}^\dagger\boldsymbol{y},\boldsymbol{\alpha}) = \frac{1}{2}\nabla\boldsymbol{r}_2(\boldsymbol{\alpha}) \oplus \boldsymbol{0}. \tag{26}$$

Thus, if $\hat{\boldsymbol{\alpha}}$ is a critical point of $\boldsymbol{r}_2$, and $\hat{\boldsymbol{c}} = \boldsymbol{\Phi}^{\dagger}(\boldsymbol{\alpha})\boldsymbol{y}$, then $(\hat{\boldsymbol{c}}, \hat{\boldsymbol{\alpha}})$ is a critical point of $\boldsymbol{r}$ as well.

Next we need to show that the global miminizer of these two problems coincide. Assume $\hat{\boldsymbol{\alpha}}$ is a global minimizer of $\boldsymbol{r}_2$, and $\hat{\boldsymbol{c}}$ is calculated as (20). Then obviously, $\boldsymbol{r}_2(\hat{\boldsymbol{\alpha}}) = \boldsymbol{r}(\hat{\boldsymbol{c}}, \hat{\boldsymbol{\alpha}})$. Suppose that there exists $(\boldsymbol{c}^*, \boldsymbol{\alpha}^*)$, $\boldsymbol{\alpha}^* \in \Omega$, s.t. $\boldsymbol{r}(\boldsymbol{c}^*, \boldsymbol{\alpha}^*) < \boldsymbol{r}(\hat{\boldsymbol{c}}, \hat{\boldsymbol{\alpha}})$. Because VARPRO takes an extra step to minimize over $\boldsymbol{c}$, we have $\boldsymbol{r}_2(\boldsymbol{\alpha}) \leq \boldsymbol{r}(\boldsymbol{c}, \boldsymbol{\alpha})$, $\forall \boldsymbol{\alpha}, \boldsymbol{c}$. Then it follows that $\boldsymbol{r}(\boldsymbol{\alpha}^*) \leq \boldsymbol{r}(\boldsymbol{c}^*, \boldsymbol{\alpha}^*) < \boldsymbol{r}(\hat{\boldsymbol{c}}, \hat{\boldsymbol{\alpha}}) = \boldsymbol{r}_2(\hat{\boldsymbol{\alpha}})$. But this is contradictory to the fact that $\hat{\boldsymbol{\alpha}}$ was a global minimizer of $\boldsymbol{r}_2$ in Ω. So the hypothesis is not valid, and $(\hat{\boldsymbol{c}}, \hat{\boldsymbol{\alpha}})$ is also a global minimizer of $\boldsymbol{r}$ in Ω.

b) Conversely, if $(\hat{\boldsymbol{c}}, \hat{\boldsymbol{\alpha}})$ is a global minimizer of $\boldsymbol{r}$ in Ω and let $\boldsymbol{c}^* = \boldsymbol{\Phi}^{\dagger}(\hat{\boldsymbol{\alpha}})\boldsymbol{y}$, then $\boldsymbol{r}_2(\hat{\boldsymbol{\alpha}}) = \boldsymbol{r}(\boldsymbol{c}^*, \hat{\boldsymbol{\alpha}}) \leq \boldsymbol{r}(\hat{\boldsymbol{c}}, \hat{\boldsymbol{\alpha}})$. Since $\boldsymbol{r}(\hat{\boldsymbol{c}}, \hat{\boldsymbol{\alpha}})$ is a global minimizer, it has to be equal sign in the formal inequality. Because for any $\boldsymbol{\alpha}$, $\boldsymbol{r}_2(\boldsymbol{\alpha}) \geq \boldsymbol{r}(\hat{\boldsymbol{c}}, \hat{\boldsymbol{\alpha}}) = \boldsymbol{r}_2(\hat{\boldsymbol{\alpha}})$, we proved $\hat{\boldsymbol{\alpha}}$ is a global minimizer of $\boldsymbol{r}_2$ as well.

Furthermore, if $\hat{\boldsymbol{c}}$ is unique, then it must be the same as $\boldsymbol{c}^* = \boldsymbol{\Phi}^{\dagger}(\hat{\boldsymbol{\alpha}})\boldsymbol{y}$, because $\boldsymbol{r}(\boldsymbol{c}^*, \hat{\boldsymbol{\alpha}}) = \boldsymbol{r}(\hat{\boldsymbol{c}}, \hat{\boldsymbol{\alpha}})$. Therefore, part b) of Theorem 1 is also proved.

■

C. Differentiation of pseudo-inverses

Another result from the paper [3] is the differentiation of pseudo-inverses, which is described as following.

Theorem 2 (Differentiation of Pseudo-inverses): Let $\Omega \subset \mathbb{R}^k$ be an open set and for $\boldsymbol{\alpha} \in \Omega$, let $A(\boldsymbol{\alpha})$ be an $m \times n$ Frechet differentiable matrix function with local constant rank $r \leq \min(m, n)$ in Ω. Then for any $\boldsymbol{\alpha} \in \Omega$, we have

$$\begin{aligned} DA^{\dagger}(\boldsymbol{\alpha}) &= -A^{\dagger}(DA)A^{\dagger} + A^{\dagger}(A^{\dagger})^T(DA)^T P_{A^{\perp}} \\ &\quad + P_{(A^{\dagger})^{\perp}}(DA)^T(A^{\dagger})^T A^{\dagger}. \end{aligned} \tag{27}$$

Similar to Theorem 1, we need to require in addition that the range space of A is complete to generalize this theorem to infinite dimensional spaces.

Proof of Theorem 2: Since the proof provided by its original paper [3] was very clear and detailed, we will not be going through it here. It can be found as Theorem 4.3 on pp. 420-421 [3]. ■

IV. Algorithm

The paper [3] provided us theoretical guarantee that obtaining the critical point or global minimum of the VARPRO functional (13) would be the same as solving the original nonlinear least squares problem (6). We are going to solve (13) using the Gauss-Newton algorithm, which is a special kind of the descent algorithms for optimization problems.

The general descent algorithm for optimization of $f(x)$ is as following [4].

1) Given a starting point $x^{(0)}$.
2) Repeat until stopping criterion is satisfied.
   a) Determine a descent search direction $\Delta x^{(k)}$.
   b) Choose a step length $t > 0$.
   c) Update. $x^{(k+1)} := x^{(k)} + t^{(k)} \Delta x^{(k)}$.

Specifically, the Gaussian-Newton algorithm [4] calculates the search direction by:

$$\boldsymbol{r}_2(\boldsymbol{\alpha}) \approx \boldsymbol{r}_2(\boldsymbol{\alpha}^{(k)}) + D\boldsymbol{r}_2(\boldsymbol{\alpha}^{(k)})(\boldsymbol{\alpha} - \boldsymbol{\alpha}^{(k)}),$$

$$\Delta\boldsymbol{\alpha} \approx \arg\min_{\Delta\boldsymbol{\alpha}} \|\boldsymbol{r}_2(\boldsymbol{\alpha}^{(k)}) + D\boldsymbol{r}_2(\boldsymbol{\alpha}^{(k)})\Delta\boldsymbol{\alpha}\|_2.$$

This is a least squares problem and can be solved by pseudo-inverse of $D\boldsymbol{r}_2$:

$$\begin{aligned} \Delta\boldsymbol{\alpha}^{(k)} &= -(D\boldsymbol{r}_2)^\dagger \boldsymbol{r}_2, \\ D\boldsymbol{r}_2 &= D(\boldsymbol{\Phi}\boldsymbol{\Phi}^\dagger)\boldsymbol{y} \\ &= (D\boldsymbol{\Phi})\boldsymbol{\Phi}^\dagger\boldsymbol{y} + \boldsymbol{\Phi}(D\boldsymbol{\Phi}^\dagger)\boldsymbol{y}, \end{aligned} \tag{28}$$

where the last equal sign comes from our Proposition 1. As we already have the derivative of pseudo-inverses as shown in Theorem 2, Eqn. (28) can be easily handled. Since $D\boldsymbol{r}_r$ is a Jacobian matrix, its j-th column is as following:

$$D\boldsymbol{r}_2(\boldsymbol{\alpha})[:, j] = (\frac{\partial\boldsymbol{\Phi}}{\partial\alpha_j})\boldsymbol{\Phi}^\dagger\boldsymbol{y} + \boldsymbol{\Phi}(\frac{\partial\boldsymbol{\Phi}^\dagger}{\partial\alpha_j})\boldsymbol{y},$$

where

$$\begin{aligned} \frac{\partial\boldsymbol{\Phi}^\dagger}{\partial\alpha_j} &= -\boldsymbol{\Phi}^\dagger(\frac{\partial\boldsymbol{\Phi}}{\partial\alpha_j})\boldsymbol{\Phi}^\dagger + \boldsymbol{\Phi}^\dagger(\boldsymbol{\Phi}^\dagger)^T(\frac{\partial\boldsymbol{\Phi}}{\partial\alpha_j})^T P_{\boldsymbol{\Phi}^\perp} \\ &\quad + P_{(\boldsymbol{\Phi}^\dagger)^\perp}(\frac{\partial\boldsymbol{\Phi}}{\partial\alpha_j})^T(\boldsymbol{\Phi}^\dagger)^T\boldsymbol{\Phi}^\dagger. \end{aligned} \tag{29}$$

A. Extension to the Complex Number Case

In the analysis of [3], the author assumed that all the variables and observations were real numbers. However, in many real applications, the observations are usually complex numbers. Therefore, we extend the algorithm such that it can handle complex observations, i.e., $\mathcal{X}$ is a Hilbert space where its elements are complex vectors. We can further assume that the variables remain to be real without loss of generality, since if they are not we can directly separate them into real parts and imaginary parts.

In the complex number case, we can define a new functional $\tilde{\boldsymbol{r}}_2(\boldsymbol{\alpha})$:

$$\tilde{\boldsymbol{r}}_2(\boldsymbol{\alpha}) = \begin{bmatrix} \mathrm{Re}(\boldsymbol{r}_2) \\ \mathrm{Im}(\boldsymbol{r}_2) \end{bmatrix}.$$

It is easy to verify that $\|\boldsymbol{r}_2(\boldsymbol{\alpha})\|_2 = \|\tilde{\boldsymbol{r}}_2(\boldsymbol{\alpha})\|_2$. Thus, solving (13) is equivalent to solving

$$\begin{aligned} \hat{\boldsymbol{\alpha}}(\boldsymbol{y}) &= \arg\min_{\boldsymbol{\alpha}} \|\tilde{\boldsymbol{r}}_2(\boldsymbol{\alpha})\|_2, & (30) \\ \hat{\boldsymbol{c}}(\boldsymbol{y}) &= (\tilde{\boldsymbol{\Phi}})^{\dagger}\tilde{\boldsymbol{y}}, & (31) \end{aligned}$$

where

$$\tilde{\boldsymbol{\Phi}} = \begin{bmatrix} \mathrm{Re}(\boldsymbol{\Phi}) \\ \mathrm{Im}(\boldsymbol{\Phi}) \end{bmatrix},$$

$$\tilde{\boldsymbol{y}} = \begin{bmatrix} \mathrm{Re}(\boldsymbol{y}) \\ \mathrm{Im}(\boldsymbol{y}) \end{bmatrix}.$$

Since we keep the variables $\boldsymbol{\alpha} \in \mathbb{R}^k$, then the Frechet derivative of $\boldsymbol{r}_2$ can still be calculated by (28). And it can be justified that the derivative of $\tilde{\boldsymbol{r}}_2$ and $\boldsymbol{r}_2$ satisfies the following simple relationship:

$$D\tilde{\boldsymbol{r}}_2 = \begin{bmatrix} \mathrm{Re}(D\boldsymbol{r}_2) \\ \mathrm{Im}(D\boldsymbol{r}_2) \end{bmatrix}.$$

Therefore, if $\boldsymbol{y} \in \mathbb{C}^m$, the VARPRO method can be easily modified to copy with that situation.

V. Experiment

We take the MRI signals (damped sinusoid signals) as an example to show our VARPRO algorithm. The formulation of MRI signals is

$$d(t) = \sum_{n=1}^{N} c_n e^{-\frac{t}{T_{2n}^*}} e^{-i2\pi f_n t}, t \geq 0,$$

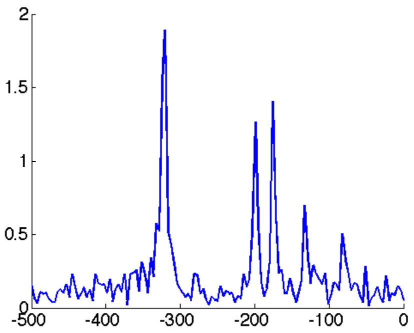

Fig. 1. Spectrum of Observation with Noise

where N is the number of metabolites in the experimented subject, and c_n, T_{2n}^* and f_n are the concentration, T2 relaxation time and resonance frequency of the n-th metabolites, respectively. We can see here c_n's are the linear variables, while T_{2n}^*'s and f_n's are nonlinear variables. Thus, this is a model whose variables separate. If the observation data are corrupted by Gaussian noise and we are going to estimate those variables by solving a least squares problem, then the VARPRO algorithm we implemented above can be utilized.

A. Simulation Setup

The spectrum of our simulation observation data is shown in Fig. 1, the x-axis of which is in Hertz. We set $N = 6$, which means we had six metabolites in total. And the ground truth is shown in Table I.

B. VARPRO

Firstly we used HSVD [5] to obtain an initial guess of all the variables. Then, we used the VARPRO algorithm we implemented, and the effect of our fitting is shown in Fig. 2, where the red line represents the estimated signal and the black line represents the residual. We also list all the estimated variables in Table II for comparison.

VI. Discussion

The original paper of VARPRO [3] provided us confidence that if the nonlinear least squares problem has separable variables, then estimating the nonlinear variables and linear variables separately by VARPRO can always give us a solution that coincides the solution of the original

TABLE I
SIMULATION SETUP

Relative c_n	T_{2n}^* (ms)	f_n (Hz)
1.98	201.7	321.9
0.56	54.6	280.5
1.11	182.1	198.1
1.19	206.9	174.8
1	68.7	132.6
0.56	140.2	80.7

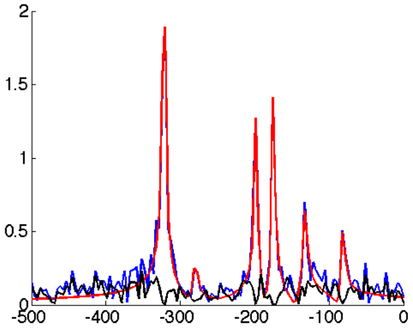

Fig. 2. Spectrum of Estimated Signal

problem. And the result of the derivative of pseudo-inverses also provided us a powerful tool for many problems, otherwise we could only handle the pseudo-inverses of full row ranked or full column ranked matrices.

We can see from Fig. 2 that the VARPRO algorithm gave us a very good fitting of the spectrum, even for the small peak around 280Hz. And by comparison between Table I and II, we can see the estimation is accurate to some extent. However, due to our time limit, we were not able to explore more extensive experiments showing the features of VARPRO and its comparison to its original problem. We are not clear about the performance of VARPRO if the signal-to-noise ratio becomes much worse, or if the number of variables grows to be much larger, either. But

TABLE II

ESTIMATION RESULTS

Relative c_n	T_{2n}^* (ms)	f_n (Hz)
2.16	166.9	322.0
0.37	87.1	279.9
1.12	189.8	198.0
1.17	234.5	174.8
1.09	68.5	132.8
0.56	135.4	81.2

we are clear that VARPRO is supposed to be more robust than its original problem because it has fewer variables to estimate (i.e., lower order) at each step.

In terms of the generalization to infinite dimensional spaces, we proved that all the proposition, lemma and theorems we proved above is valid if the dimension of observation (m) and the dimension of nonlinear variables (k) are infinite, while the model order n should be a fixed and finite number. We have also generalized VARPRO to the complex number case, which is also very common in real applications.

REFERENCES

[1] B. C. Levy, *Principles of signal detection and parameter estimation.* Springer, 2008, ch. 4.4, pp. 131–150.

[2] Y. Bresler, S. Basu, and C. Couvreur, "Vector spaces and least squares methods for signal processing," February 2009, uiuc-ece513-textbook.

[3] G. H. Golub and V. Pereyra, "The differentiation of pseudo-inverses and nonlinear least squares problems whose variables separate," *SIAM Journal of Numerical Analysis*, vol. 10, no. 2, April 1973.

[4] S. P. Boyd and L. Vandenberghe, *Convex optimization.* Cambridge University Press, 2004.

[5] H. Barkhuijsen, R. D. Beer, and D. V. Ormondt, "Improved algorithm for noniterative time-domain model fitting to exponentially damped magnetic resonance signals," *Journal of Magneic Resonance*, pp. 553–557, 1987.