# Vector Valued Regression for Iron Overload Estimation

Luca Baldassarre
*DIFI - DISI*
*via Dodecaneso 33 - 35*
*I-16146 Genova*
*baldassarre@disi.unige.it*

Annalisa Barla
*DISI*
*via Dodecaneso 35*
*I-16146 Genova*
*barla@disi.unige.it*

Barbara Gianesin
Mauro Marinelli
*DIFI*
*via Dodecaneso 33*
*I-16146 Genova*
*{gianesin, marinelli}@ge.infn.it*

## Abstract

*In this work we present and discuss in detail a novel vector-valued regression technique: our approach allows for an all-at-once estimation, as opposed to solve a number of scalar-valued regression tasks. Despite its general purpose nature, the method has been designed to solve a delicate medical issue: a reliable and non-invasive assessment of body-iron overload.*

*The Magnetic Iron Detector (MID) measures the magnetic track of a person, which depends on the anthropometric characteristics and the body-iron burden. We aim to provide an estimate of this signal in absence of iron overload. We show how this question can be formulated as the estimation of a vector-valued function which encompasses the prior knowledge on the shape of the magnetic track. This is accomplished by designing an appropriate vector-valued feature map. We successfully applied the method on a dataset of 84 volunteers.*

## 1 Introduction

Certain blood diseases, such as thalassemia and hemochromatosis, are characterized by the accumulation of iron in the body organs, mainly in the liver. The therapies for these disorders require to accurately evaluate the iron overload in the patients. Marinelli and colleagues [7, 6] have developed a room-temperature biosusceptometer, the Magnetic Iron Detector (MID), that allows the non-invasive assessment of the iron overload in the whole liver, as opposed to the invasive liver biopsy, which still is the most used diagnostic tool. The biosusceptometer is composed of an AC magnetic source and a pickup coil which measures the electromotive force *emf* produced by the oscillation of the magnetic field flux. When a body is inserted under the magnetic field it produces a variation of the *emf* which depends on the magnetic properties of the body and on its position relative to the magnetic field axis.

Due to the oscillating field, the magnetic signal generated by the human body has two independent sources: the magnetization signal, from the magnetic properties of the tissues, and the eddy current signal, from their electrical conductivity. For each patient a double track is recorded: only the magnetization signal depends on the iron overload. Given an estimate of the background signal of a patient, that is the signal that would be generated in absence of iron overload, it is possible to recover the liver iron overload by subtracting the two signals.

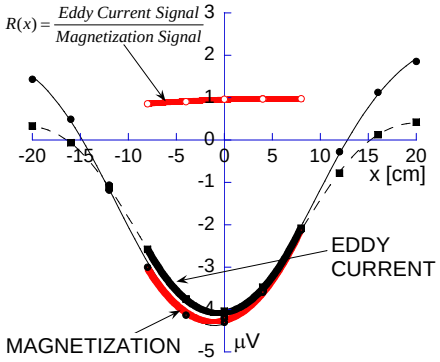


**Figure 1.** Eddy current and magnetization signals of a volunteer.

The parametric model developed by Marinelli and coworkers [6] is currently used at the "E.O. Ospedali Galliera" Hospital in Genoa, Italy, for assessing the iron overload. The model has been trained on a dataset of 84 healthy volunteers and it estimates the ratio $R(x)$ between the two signals shown in Fig.1. The core idea behind their approach is that the magnetization signal of a well-treated patient is indistinguishable from the one

of a healthy volunteer with the same biometric features, see Fig.2. Furthermore, they assume that the ratio $R(x)$ of the two signals, evaluated only in the range between -8cm and 8cm, resembles a parabola.

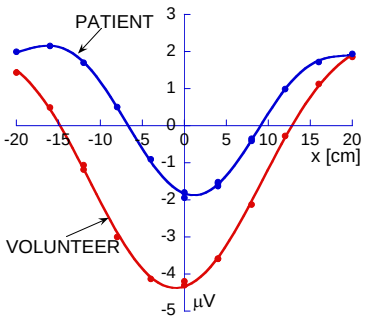


**Figure 2.** Magnetic tracks of a healthy volunteer and of a patient with similar biometric features: the liver iron overload produces an evident variation of the signal in the left part of the track. Position x = 0cm corresponds to the center of the body; negative positions to the liver side, positive positions to the spleen side.

We reformulate this problem in the context of Statistical Learning presenting a method to transform a curve fitting task into a vector-valued regression model. Since the measures are always taken at fixed positions along the measurement axis, they can be thought of as components of a vector and a high correlation among them can be assumed, because they approximately lie on a parabola. In this way we eschew from directly estimating the magnetization curve. The method described in Sec. 2.1 can be implemented by means of iterative algorithms, see [5], such as Landweber, $\nu$-method or the sparsity enforcing *l1l2* regularization [3, 4].

## 2 Vector-valued regression

Given a training set of example points $D_n \equiv \{(\mathbf{x}_i, \mathbf{y}_i) \in \mathcal{X} \times \mathcal{Y}\}_{i=1}^n$, obtained by sampling $\mathcal{X} \times \mathcal{Y}$ according to the unknown probability $P(\mathbf{x}, \mathbf{y})$, the problem of learning consists in providing a deterministic estimator $f : \mathcal{X} \to \mathcal{Y}$ with good generalization properties on unseen examples.

The best possible estimator would be the one minimizing the so-called *expected risk*, that is:

$$I[f] \equiv \int_{\mathcal{X} \times \mathcal{Y}} V(\mathbf{y}, f(\mathbf{x})) P(\mathbf{x}, \mathbf{y})\, d\mathbf{x} d\mathbf{y},$$

where $V(\mathbf{y}, f(\mathbf{x}))$ is the loss function measuring the error of predicting $\mathbf{y}$ by $f(\mathbf{x})$. Since $P$ is unknown, we could try to minimize the *empirical risk*: $\mathcal{E}_n(f) = \frac{1}{n} \sum_{i=1}^n V(\mathbf{y}_i, f(\mathbf{x}_i))$, but we would end up in solving an ill-posed problem, since the solution is not unique, not stable and with poor generalization properties. Therefore we select the minimizer of

$$\mathcal{E}_n(f) + \lambda ||f||_K^2 \tag{1}$$

in a Reproducing Kernel Hilbert Space (RKHS), provided with a kernel function $K$. The second term in (1) represents the *complexity* of the function $f$. In our setting $\mathcal{Y} \subseteq \mathbb{R}^d$, therefore the kernel function $K$ is $K : \mathcal{X} \times \mathcal{X} \to \mathcal{L}_+(\mathbb{R}^d)$.

The representer theorem [8] guarantees that the solution can always be written as:

$$f(\mathbf{x}) = \sum_{i=1}^n K(\mathbf{x}, \mathbf{x}_i) \mathbf{c}_i,$$

where the coefficients $\mathbf{c}_i$ depend on the data, on the loss function, on the kernel choice and on the regularization parameter $\lambda$. For vector-valued functions the $\mathbf{c}_i$ are $d$-dimensional vectors, while $K(\mathbf{x}, \mathbf{x}_i)$ is a $d \times d$ matrix.

The direct approach, as in [8], is computationally expensive since it requires to invert a $nd \times nd$ matrix.

To overcome this issue, we propose an extension to the vector-valued case of iterative algorithms methods, originally developed for scalar regression [9, 5]. The main idea of these techniques is to start with an approximate solution and iteratively add a correction in the direction opposite to the gradient of the empirical risk. By early stopping the procedure, a regularized solution is achieved. The number of iterations $m$ plays the role of the regularization parameter $\lambda$. In practice we obtain a sequence of solutions, one for each iteration, avoiding to directly compute the solution for each regularization parameter, therefore allowing a faster selection of the optimal stopping point. Another advantage of such approaches consists in solving the minimization problem without inverting any matrix.

We also extend to the vector-valued case the *l1l2* regularization and assess its performance on this problem. The *l1l2* regularization is a sparsification method initially proposed by [10] and studied in [3]. This method iteratively minimizes the following functional, derived from (1) with a square loss and the addition of a $l1$ penalty term:

$$\frac{1}{n} \sum_{i=1}^n ||(\mathbf{y}_i - f(\mathbf{x}_i))||_d^2 + \lambda(1-\alpha)||f||_K^2 + \lambda\alpha||f||_{l1}.$$

### 2.1 Designing the feature map

For each person $i = 1, \ldots, 84$, we consider only the 5 measures $y_{ik}$ at positions $t_k = \{-8, -4, 0, 4, 8\} cm$. The measures can be thought as the components of a five-dimensional vector and lie approximately on a

parabola, hence we can model them as $y_{ik} = f(\mathbf{x_i})^k + \epsilon_{ik}$, where $\mathbf{x_i}$ stands for the biometric data of the volunteer $i$, $\epsilon_{ik}$ representing the noise and:

$$f(\mathbf{x_i})^k = c_0(\mathbf{x_i}) + c_1(\mathbf{x_i})t_k + c_2(\mathbf{x_i})t_k^2. \qquad (2)$$

In our model, we assume that the coefficients $c_0, c_1, c_2$ depend linearly on $\mathbf{x}$: $c_j(\mathbf{x}) = \boldsymbol{\beta}_j \cdot \mathbf{x}$, for $j = 0, 1, 2$ and introduce the vector-valued feature map, $\varphi : \mathcal{X} \to \mathbb{R}^{5\times 3p}$, ($\mathcal{X} \subseteq \mathbb{R}^p$, $p = 22$):

$$\varphi(\mathbf{x}) = \left(\begin{array}{c|c|c} \mathbf{x} & \mathbf{x}t_1 & \mathbf{x}t_1^2 \\ \mathbf{x} & \mathbf{x}t_2 & \mathbf{x}t_2^2 \\ \mathbf{x} & \mathbf{x}t_3 & \mathbf{x}t_3^2 \\ \mathbf{x} & \mathbf{x}t_4 & \mathbf{x}t_4^2 \\ \mathbf{x} & \mathbf{x}t_5 & \mathbf{x}t_5^2 \end{array}\right). \qquad (3)$$

Hence, the vector-valued estimator can be written as a linear combination of the features $\varphi(\mathbf{x})$:

$$\mathbf{f}(\mathbf{x}) = \varphi(\mathbf{x})\boldsymbol{\beta}, \qquad \boldsymbol{\beta} \in \mathbb{R}^{3p}, \qquad (4)$$

where $\boldsymbol{\beta}$ is obtained by concatenating the vectors $\boldsymbol{\beta}_j$.

Choosing the square loss, the empirical risk is:

$$\mathcal{E}_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{y}_i - \varphi(\mathbf{x}_i)\boldsymbol{\beta}\|^2_{\mathbb{R}^5}.$$

Our aim is to compare the estimates of $\boldsymbol{\beta}$ obtained with three algorithms: Landweber, $\nu$-method and *l1l2*.

These methods require the computation of the empirical risk gradient, which, for this specific case, is:

$$\begin{aligned} \nabla\mathcal{E}_n(\boldsymbol{\beta}) &= -\frac{2}{n}(\varphi Y - \varphi^T\varphi\boldsymbol{\beta}) \qquad (5)\\ (\varphi Y)_\gamma &= \sum_{i=1}^{n} < \varphi_\gamma(\mathbf{x}_i), \mathbf{y}_i >_{\mathbb{R}^5} \\ (\varphi^T\varphi)_{\gamma,\gamma`} &= \sum_{i=1}^{n} < \varphi_\gamma(\mathbf{x}_i), \varphi_{\gamma`}(\mathbf{x}_i) >_{\mathbb{R}^5} \end{aligned}$$

where $\varphi_\gamma(\mathbf{x})$ corresponds to the $\gamma$-th row of $\varphi(\mathbf{x})$, $\varphi Y \in \mathbb{R}^{3p}$ and $\varphi^T\varphi \in \mathbb{R}^{3p\times 3p}$.

The simple Landweber approach [1] starts with the null solution which is updated by adding the negative of the gradient multiplied by a constant step size, $\eta$:

$$\boldsymbol{\beta}_{m+1} = \boldsymbol{\beta}_m - \eta\nabla\mathcal{E}_n(\boldsymbol{\beta}_m), \qquad \boldsymbol{\beta}_0 = (0, \ldots, 0)$$

The number of iterations $m$ corresponds to the regularization parameter $\lambda^{-1}$.

The $\nu-$method [5] extends Landweber by using a dynamic step size and introducing an inertial term which keeps memory of the previous update:

$$\boldsymbol{\beta}_{m+1} = \boldsymbol{\beta}_m + u(\boldsymbol{\beta}_m - \boldsymbol{\beta}_{m-1}) - w\eta\nabla\mathcal{E}_n(\boldsymbol{\beta}_m),$$

where $w$ and $u$ change at each iteration. The number of iterations corresponds to $\lambda^{-2}$.

*l1l2* regularization iteratively minimizes the following functional:

$$\frac{1}{n}\sum_{i=1}^{n} \|\mathbf{y}_i - \varphi(\mathbf{x}_i)\boldsymbol{\beta}\|^2_{\mathbb{R}^5} + \lambda(1-\alpha)||\boldsymbol{\beta}||_2^2 + \lambda\alpha||\boldsymbol{\beta}||_1$$

The $l1$ penalty term forces many of the coefficients $\beta_{lj}$ to be zero: the corresponding variables are irrelevant and can be discarded. The iterations are essentially of the Landweber type, but at each step the coefficients are soft-thresholded and shrunk:

$$\boldsymbol{\beta}_{m+1} = H(\boldsymbol{\beta}_m - \eta\nabla\mathcal{E}_n(\boldsymbol{\beta}_m), \lambda\alpha)/(1 + \lambda(1-\alpha))$$

and $H(\boldsymbol{\beta}, \tau)$ is the soft-thresholding operator, which sets to zero all components of $\boldsymbol{\beta}$ within $[-\tau, \tau]$ and shifts towards zero by $\tau$ the remaining ones. The algorithm stops accordingly to a convergence criterion [3].

From the vector-valued feature map $\varphi$ we can derive the corresponding matrix-valued kernel. Following [2]:

$$(K(x,s))_{pq} = \sum_{k=1}^{5} \varphi_{kp}(x)\varphi_{kq}(t) = (x\cdot s)(1 + t_p t_q + t_p^2 t_q^2).$$

This kernel is of the form $K(x,s) = k(x,s)\mathbf{A}$, with $k(x,s)$ a scalar kernel and $\mathbf{A}$ a positive semi-definite matrix. Further prior information can be included by changing the scalar kernel $k$ or the matrix $\mathbf{A}$.

### 2.2 Model selection and assessment

We adopt an experimental protocol in order to select the model parameters and assess the generalization capabilities of our method in an unbiased way, performing two nested loops of K-fold Cross Validation. The inner loop is a 5-fold Cross Validation and is performed to select the regularizing parameter. For each value of the parameter, an estimate of the generalization error is computed. The value that minimizes the error is used for training. The outer loop is a Leave One Out (LOO) Cross Validation evaluating the performance of the chosen model. The estimate of the generalization error is the mean of the $n$ empirical errors.

## 3 Results

The data set is composed of 84 healthy volunteers represented by their biometric data and the five measures of the eddy-current signal. This features are highly inhomogeneous and can lead to numerical problems, therefore we normalized our data. We set the columns of the $n\times p$ data matrix $X = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$, to

have zero mean and fixed range and changed the variable $t$ from $\{-8, -4, 0, 4, 8\}$ to $\{-1, -0.5, 0, 0.5, 1\}$, since it only represents a label for the components of the vector $\mathbf{y}$. Thus, each element of the 3d matrix $\varphi(X) \in \mathbb{R}^{n \times 3p \times 5}$, obtained by applying the feature map to the data matrix $X$, belongs to $[-1, 1]$. In the test phase, we apply these normalizing factors to the test data.

**Table 1. Model parameters**

| Model | Number of iterations | $\lambda$ | Time |
|---|---|---|---|
| Landweber | 397 | n.a. | 1605 s |
| $\nu$-method | 68 | n.a. | 114 s |
| *l1l2* | $23 * 10^4$ | $10^{-5}$ | 5300 s |

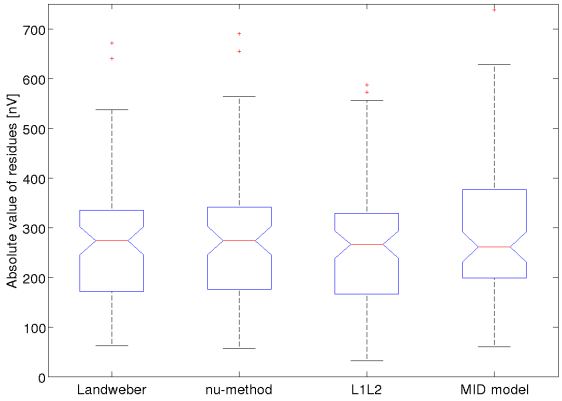


**Figure 3.** LOO errors distributions. The first three models are obtained from the vector-valued one by the indicated algorithms.

For model selection and assessment we used the experimental protocol outlined in Sec.2.2: the model parameters to be selected are the number of iterations $m$ for the Landweber and $\nu$-method algorithms and the regularizing parameter $\lambda$ for the *l1l2* method. In the latter case, $\alpha$ was set to $0.9$ to enforce maximum sparsity while retaining correlated features [3]. For all the algorithms we set the step size $\eta = (2\|\varphi^T \varphi\|)^{-1}$ which guarantees their convergence [9, 3, 5]. From Tab.1 it can be noted that the $\nu$-method is significantly faster than the other two methods.

The boxplots shown in Fig.3 represent the LOO errors distributions for the different algorithms and for the model of [6], obtained with the same protocol. As expected, we observe that the LOO errors show a high variance. The advantage of modelling the problem as vector-valued regression is supported by the fact that the three algorithms consistently lead to error distributions that are closer to zero with respect to the MID model.

The accuracies obtained correspond to a precision in the iron overload estimation of about $0.8g$. Iron overload lower than $1g$ is considered mild: currently no model is capable to detect this kind of iron burden.

## 4 Conclusions

The model proposed is a general method to approach vector-valued regression problems. Moreover, it can be used to estimate a curve explained by a variable that is always sampled at fixed values. Prior knowledge (e.g. the shape of the curve w.r.t. the parametrizing variable, or the correlation among the elements of the vector-valued function to be estimated) can be easily incorporated into the feature map or the kernel function.

Our results show that the iterative algorithms can be applied to the vector-valued case with success. They also provide an efficient alternative to the direct computation of the inverse of a $nd \times nd$ matrix. The model selection and validation protocol adopted leads to an unbiased solution, avoiding overfitting and unreliable estimates of the performance.

## References


[1] P. Bühlmann and B. Yu. Boosting with the l2- loss: Regression and classification. *J. Amer. Statistical Assoc.*, 98, 2002.

[2] A. Caponnetto, C. Micchelli, M. Pontil, and Y. Ying. Universal kernels for multi-task learning. *Journal of Machine Learning Research*, submitted.

[3] C. De Mol, E. De Vito, and L. Rosasco. Consistency of elastic-net regularization. Technical report, DISI, 2008.

[4] C. De Mol, S. Mosci, M. Traskine, and A. Verri. A regularized method for selecting nested groups of relevant genes from microarray data. Technical report, DISI, 2007.

[5] L. Lo Gerfo, L. Rosasco, F. Odone, E. De Vito, and A. Verri. Spectral algorithms for supervised learning. *Neural Computation*, to appear.

[6] M. Marinelli, S. Cuneo, B. Gianesin, A. Lavagetto, M. Lamagna, E. Oliveri, G. Sobrero, L. Terenzani, and G. Forni. Non-invasive measurement of iron overload in the human body. *IEEE Trans. on applied superconductivity*, 16(2), June 2006.

[7] M. Marinelli, B. Gianesin, M. Lamagna, A. Lavagetto, E. Oliveri, M. Saccone, G. Sobrero, L. Terenzani, and G. Forni. Whole liver iron overload measurement by a non-cryogenic magnetic susceptometer. In *Proc. of New Frontiers in Biomagnetism*, Vancouver, Canada, 2007.

[8] C. A. Micchelli and M. Pontil. On learning vector-valued functions. *Neural Computation*, 17, 2005.

[9] Y. Yao, L. Rosasco, and A. Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, August 2007.

[10] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *J. Roy. Statistical Society: Series B*, 67(2), 2005.