Multi-Output Learning with Spectral Filters

Luca Baldassarre

Università di Genova
Corso di Dottorato in Fisica
Ciclo XXII
DIFI - SlipGuru

March 26, 2010

Outline

Outline

1. Motivations
2. Supervised learning
   - Supervised learning basics
   - Problem setting
   - Spectral filters
   - Theoretical results
3. Experiments
   - Simulated vector fields
   - Magnetic Iron Detector
4. Conclusions

Motivations

- There are many processes for which an explicit modeling is unfeasible
- We can **learn** a **predictive** model from a **training set** of input-output *examples*.
- Many processes require the estimation of several **related** outputs simultaneously
- We show a *unified framework to solve them efficiently*.

Motivations

- There are many processes for which an explicit modeling is unfeasible
- We can **learn** a **predictive** model from a **training set** of input-output *examples*.
- Many processes require the estimation of several **related** outputs simultaneously
- We show a *unified framework to solve them efficiently*.

Multi-output learning problems

- **Multi-class classification**: classify a datum into one of several categories.

Motivations

- There are many processes for which an explicit modeling is unfeasible
- We can **learn** a **predictive** model from a **training set** of input-output *examples*.
- Many processes require the estimation of several **related** outputs simultaneously
- We show a *unified framework to solve them efficiently*.

Multi-output learning problems

- **Multi-class classification**: classify a datum into one of several categories. Face Recognition

Motivations

- There are many processes for which an explicit modeling is unfeasible
- We can **learn** a **predictive** model from a **training set** of input-output *examples*.
- Many processes require the estimation of several **related** outputs simultaneously
- We show a *unified framework to solve them efficiently*.

Multi-output learning problems

- **Multi-class classification**: classify a datum into one of several categories. Face Recognition
- **Multi-task learning**: many related scalar regression tasks, each provided with its own training set.

Motivations

- There are many processes for which an explicit modeling is unfeasible
- We can **learn** a **predictive** model from a **training set** of input-output *examples*.
- Many processes require the estimation of several **related** outputs simultaneously
- We show a *unified framework to solve them efficiently*.

Multi-output learning problems

- **Multi-class classification**: classify a datum into one of several categories. Face Recognition
- **Multi-task learning**: many related scalar regression tasks, each provided with its own training set. Consumer preferences

Motivations

- There are many processes for which an explicit modeling is unfeasible
- We can **learn** a **predictive** model from a **training set** of input-output *examples*.
- Many processes require the estimation of several **related** outputs simultaneously
- We show a *unified framework to solve them efficiently*.

Multi-output learning problems

- **Multi-class classification**: classify a datum into one of several categories. Face Recognition
- **Multi-task learning**: many related scalar regression tasks, each provided with its own training set. Consumer preferences
- **Vector-valued learning**: a regression task where we have multiple outputs but only one training set.

Motivations

- There are many processes for which an explicit modeling is unfeasible
- We can **learn** a **predictive** model from a **training set** of input-output *examples*.
- Many processes require the estimation of several **related** outputs simultaneously
- We show a *unified framework to solve them efficiently*.

Multi-output learning problems

- **Multi-class classification**: classify a datum into one of several categories. Face Recognition
- **Multi-task learning**: many related scalar regression tasks, each provided with its own training set. Consumer preferences
- **Vector-valued learning**: a regression task where we have multiple outputs but only one training set. Velocity field

Main ingredients for Multi-Output Learning

Key Requirements

1. **Generalization**: ability to predict outside the training set.
2. Methods that deal with *few and noisy* data.
3. *Model-free* methods that
4. allow for the incorporation of *prior* information.
5. **Consistency**: guarantee that increasing the number of examples leads to optimal estimators.

Main ingredients for Multi-Output Learning

Key Requirements

1. **Generalization**: ability to predict outside the training set.
2. Methods that deal with *few and noisy* data.
3. *Model-free* methods that
4. allow for the incorporation of *prior* information.
5. **Consistency**: guarantee that increasing the number of examples leads to optimal estimators.

Key Ingredients

1. Proper *Hypothesis Spaces* where to search for *estimators*
2. **Robust** and **efficient** estimation methods

Outline

1. Motivations
2. Supervised learning
   - Supervised learning basics
   - Problem setting
   - Spectral filters
   - Theoretical results
3. Experiments
   - Simulated vector fields
   - Magnetic Iron Detector
4. Conclusions

Supervised learning

Training set

$$\mathbf{z} = \{(x_1, y_1), ..., (x_n, y_n)\} \subset \mathcal{X} \times \mathcal{Y}$$

$\mathcal{X} = \mathbb{R}^p$ **input space**
$\mathcal{Y} = \mathbb{R}^d$ **output space**

Supervised learning

Training set

$$\mathbf{z} = \{(x_1, y_1), ..., (x_n, y_n)\} \subset \mathcal{X} \times \mathcal{Y}$$

$\mathcal{X} = \mathbb{R}^p$ **input space**
$\mathcal{Y} = \mathbb{R}^d$ **output space**

Estimator

The goal is to learn a function that *generalizes* well to unseen examples

$$f_{\mathbf{z}}^n : \mathbb{R}^p \to \mathbb{R}^d$$

Supervised learning

Training set

$$\mathbf{z} = \{(x_1, y_1), ..., (x_n, y_n)\} \subset \mathcal{X} \times \mathcal{Y}$$

$\mathcal{X} = \mathbb{R}^p$ **input space**
$\mathcal{Y} = \mathbb{R}^d$ **output space**

Estimator

The goal is to learn a function that *generalizes* well to unseen examples

$$f_{\mathbf{z}}^n : \mathbb{R}^p \to \mathbb{R}^d$$

Scalar case

The theory of supervised learning in the **scalar case** (i.e. $\mathcal{Y} = \mathbb{R}$) has been extensively treated ([Vapnik and Chervonenkis, 1974, Girosi et al., 1995, Evgeniou et al., 2000, Cucker and Smale, 2001]), but still presents some interesting challenges.

Supervised learning

Training set

$$\mathbf{z} = \{(x_1, y_1), ..., (x_n, y_n)\} \subset \mathcal{X} \times \mathcal{Y}$$

$\mathcal{X} = \mathbb{R}^p$ **input space**
$\mathcal{Y} = \mathbb{R}^d$ **output space**

Estimator

The goal is to learn a function that *generalizes* well to unseen examples

$$f_{\mathbf{z}}^n : \mathbb{R}^p \rightarrow \mathbb{R}^d$$

Multi-output case

A comprehensive theory for **multi-output learning** is still at its infancy ([Micchelli and Pontil, 2005, Carmeli et al., 2006, Caponnetto et al., 2008]), despite some extensions of scalar methods have been proposed.

Supervised learning

Training set

$$\mathbf{z} = \{(x_1, y_1), ..., (x_n, y_n)\} \subset \mathcal{X} \times \mathcal{Y}$$

$\mathcal{X} = \mathbb{R}^p$ **input space**
$\mathcal{Y} = \mathbb{R}^d$ **output space**

Estimator

The goal is to learn a function that *generalizes* well to unseen examples

$$f_{\mathbf{z}}^n : \mathbb{R}^p \to \mathbb{R}^d$$

Unknown Probability Distribution

We suppose that the given examples and the future data are *identically, independently sampled* from an **unknown** probability distribution

$$p(x, y) = p(y|x)p(x) \text{ on } \mathcal{X} \times \mathcal{Y}$$

Some definitions

Hypothesis space - where we look for candidate estimators

$$\mathcal{H} \subseteq \{f : \mathbb{R}^p \to \mathbb{R}^d\}$$

Some definitions

Hypothesis space - where we look for candidate estimators

$$\mathcal{H} \subseteq \{f : \mathbb{R}^p \to \mathbb{R}^d\}$$

Expected risk - evaluates the performance of a candidate estimator

$$\mathrm{I}[f] = \int_{\mathcal{X}\times\mathcal{Y}} ||y - f(x)||_d^2 p(x, y) dxdy$$

Some definitions

Hypothesis space - where we look for candidate estimators

$$\mathcal{H} \subseteq \{f : \mathbb{R}^p \to \mathbb{R}^d\}$$

Expected risk - evaluates the performance of a candidate estimator

$$\mathrm{I}[f] = \int_{\mathcal{X}\times\mathcal{Y}} ||y - f(x)||_d^2 p(x,y) dxdy$$

Regression function and best estimator in $\mathcal{H}$

$$f_\rho(x) = \int_{\mathcal{Y}} y p(y|x) dy, \qquad \mathrm{I}[f_\rho] = \min_f \mathrm{I}[f], \qquad f_{\mathcal{H}} = \operatorname*{argmin}_{f \in \mathcal{H}} \mathrm{I}[f]$$

Some definitions

Hypothesis space - where we look for candidate estimators

$$\mathcal{H} \subseteq \{f : \mathbb{R}^p \to \mathbb{R}^d\}$$

Expected risk - evaluates the performance of a candidate estimator

$$\mathrm{I}[f] = \int_{\mathcal{X}\times\mathcal{Y}} ||y - f(x)||_d^2 p(x,y)dxdy$$

Regression function and best estimator in $\mathcal{H}$

$$f_\rho(x) = \int_{\mathcal{Y}} yp(y|x)dy, \qquad \mathrm{I}[f_\rho] = \min_f \mathrm{I}[f], \qquad f_{\mathcal{H}} = \operatorname*{argmin}_{f\in\mathcal{H}} \mathrm{I}[f]$$

Empirical Risk - all we have access to

$$\mathrm{I_S}[f] = \frac{1}{n}\sum_{i=1}^{n} ||y_i - f(x_i)||_d^2$$

Kernels and RKHS

Kernel for vector valued functions

A **kernel** is a *symmetric matrix valued* function

$$\Gamma : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}^{d \times d}$$

that satisfies a *positivity* constraint.

Kernel for vector valued functions

A **kernel** is a *symmetric matrix valued* function

$$\Gamma : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}^{d \times d}$$

that satisfies a *positivity* constraint.

Given some points $\{x_1, \ldots, x_n\}$, we can write a function $f : \mathbb{R}^p \to \mathbb{R}^d$ as

$$f(x) = \sum_{i=1}^{n} \Gamma(x, x_i) c_i, \qquad c_i \in \mathbb{R}^d.$$

Kernels and RKHS

Kernel for vector valued functions

A **kernel** is a *symmetric matrix valued* function

$$\Gamma : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}^{d \times d}$$

that satisfies a *positivity* constraint.

Given some points $\{x_1, \ldots, x_n\}$, we can write a function $f : \mathbb{R}^p \to \mathbb{R}^d$ as

$$f(x) = \sum_{i=1}^{n} \Gamma(x, x_i) c_i, \qquad c_i \in \mathbb{R}^d.$$

A kernel uniquely defines a Hilbert space of functions $f : \mathbb{R}^p \to \mathbb{R}^d$ called *Reproducing Kernel Hilbert Space*.

Empirical Risk Minimization

Empirical risk

$$\mathrm{I_S}[f] = \frac{1}{n}\sum_{i=1}^{n} ||y_i - f(x_i)||_d^2$$

Empirical Risk Minimization

Empirical risk

$$\mathrm{I_S}[f] = \frac{1}{n}\sum_{i=1}^{n} ||y_i - f(x_i)||_d^2$$

Minimizer in RKHS with kernel Γ

$$f_{\mathbf{z}}^n(x) = \sum_{i=1}^{n} \Gamma(x, x_i) c_i$$

where the coefficients $c_i \in \mathbb{R}^d$ satisfy

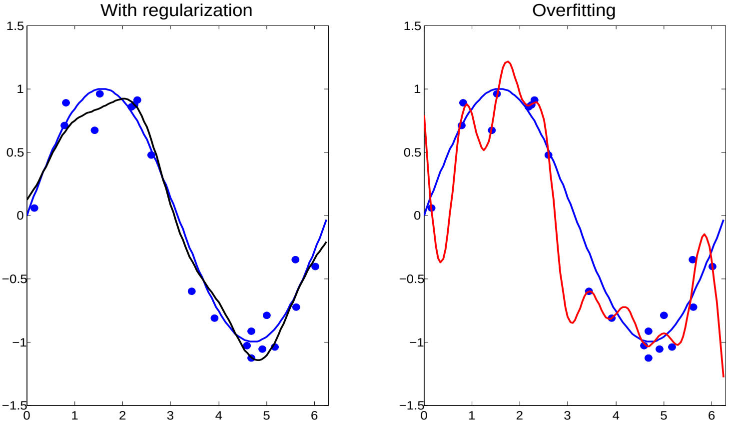
$$\mathbf{\Gamma C} = \mathbf{Y}$$

- $\mathbf{\Gamma}$ is a $n \times n$ block matrix, whose $d \times d$ (i, j) block is $\Gamma(x_i, x_j)$
- $\mathbf{C} = (c_1, \ldots, c_n)$
- $\mathbf{Y} = (y_1, \ldots, y_n)$

Empirical Risk Minimization and Overfitting

Overfitting

If $\mathcal{H}$ is too large, by minimizing the Empirical Risk, we will **fit the noise** in the data and will generalize poorly on new data.



Control Overfitting

Regularization

A technique borrowed from the Inverse Problems Theory literature [Tikhonov and Arsenin, 1977, Engl et al., 1996, De Vito et al., 2005].

$$\frac{1}{n}\sum_{i=1}^{n} ||y_i - f(x_i)||_d^2 + \lambda ||f||_{\mathcal{H}}^2$$

The norm usually controls the **smoothness** of the estimator.

Tikhonov regularization or Regularized Least Squares

Tikhonov functional - avoids overfitting - stable solution

$$\frac{1}{n}\sum_{i=1}^{n} ||y_i - f(x_i)||_d^2 + \lambda ||f||_\Gamma^2$$

Tikhonov regularization or Regularized Least Squares

Tikhonov functional - avoids overfitting - stable solution

$$\frac{1}{n}\sum_{i=1}^{n}||y_i - f(x_i)||_d^2 + \lambda||f||_\Gamma^2$$

Minimizer in RKHS with kernel Γ [Micchelli and Pontil, 2005]

$$\begin{aligned} f_{\mathbf{z}}^n(x) &= \sum_{i=1}^{n}\Gamma(x, x_i)c_i, \qquad c_i \in \mathbb{R}^d \\ \mathbf{C} &= (\boldsymbol{\Gamma} + n\lambda\mathbf{I})^{-1}\mathbf{Y}. \end{aligned}$$

The penalty term helps stabilizing the inverse of $\boldsymbol{\Gamma}$.

Spectral Filters

Idea

Instead of $(\mathbf{\Gamma} + \lambda n \mathbf{I})^{-1}$, use other *regularized* matrices $g_\lambda(\mathbf{\Gamma})$, defined by the **spectral filters** g_λ, such that

$$\lim_{\lambda \to 0} g_\lambda(\mathbf{\Gamma}) = \mathbf{\Gamma}^{-1}$$

$$\mathbf{C} = g_\lambda(\mathbf{\Gamma})\mathbf{Y}$$

Spectral Filters

Idea

Instead of $(\mathbf{\Gamma} + \lambda n \mathbf{I})^{-1}$, use other *regularized* matrices $g_\lambda(\mathbf{\Gamma})$, defined by the **spectral filters** g_λ, such that

$$\lim_{\lambda \to 0} g_\lambda(\mathbf{\Gamma}) = \mathbf{\Gamma}^{-1}$$

$$\mathbf{C} = g_\lambda(\mathbf{\Gamma})\mathbf{Y}$$

Advantages

1. **Strong statistical properties** derived from Inverse Problems
2. **Computational efficiency** of iterative algorithms
3. Regularization achieved by **early stopping** [Yao et al., 2007]
4. Not necessary to run the whole algorithm for every regularization parameter value

Iterative spectral filters

Landweber or L2 Boosting [Bühlmann and Yu, 2002, Yao et al., 2007]

Essentially it is **gradient descent** of the empirical risk with early stopping

$$\begin{aligned} \mathbf{C}^0 &= 0 \\ \mathbf{C}^t &= \mathbf{C}^{t-1} + \eta(\mathbf{Y} - \mathbf{\Gamma}\mathbf{C}^{t-1}) \end{aligned}$$

Iterative spectral filters

Landweber or L2 Boosting [Bühlmann and Yu, 2002, Yao et al., 2007]

Essentially it is **gradient descent** of the empirical risk with early stopping

$$\begin{aligned}\mathbf{C}^0 &= 0 \\ \mathbf{C}^t &= \mathbf{C}^{t-1} + \eta(\mathbf{Y} - \mathbf{\Gamma}\mathbf{C}^{t-1})\end{aligned}$$

ν-method or Accelerated L2 Boosting [Lo Gerfo et al., 2008]

Accelerated version of the previous algorithm.

$$\begin{aligned}\mathbf{C}^0 &= 0 \\ \mathbf{C}^t &= \mathbf{C}^{t-1} + u_t(\mathbf{C}^{t-1} - \mathbf{C}^{t-2}) + \frac{\omega_t}{n}(\mathbf{Y} - \mathbf{\Gamma}\mathbf{C}^{t-1})\end{aligned}$$

Error Analysis

Expected risk - evaluates the performance of a candidate estimator

$$\mathrm{I}[f] = \int_{\mathcal{X}\times\mathcal{Y}} ||y - f(x)||_d^2 p(x,y) dx dy$$

Error Analysis

Expected risk - evaluates the performance of a candidate estimator

$$\mathrm{I}[f] = \int_{\mathcal{X}\times\mathcal{Y}} ||y - f(x)||_d^2 p(x,y) dxdy$$

Regression function - f_ρ

$$\mathrm{I}[f_\rho] = \min_f \mathrm{I}[f]$$

Error Analysis

Expected risk - evaluates the performance of a candidate estimator

$$\mathrm{I}[f] = \int_{\mathcal{X}\times\mathcal{Y}} ||y - f(x)||_d^2 p(x,y)dxdy$$

Regression function - f_ρ

$$\mathrm{I}[f_\rho] = \min_f \mathrm{I}[f]$$

Excess Risk - how well we are doing compared to the best

$$\mathrm{I}[f_{\mathbf{z}}^n] - \mathrm{I}[f_\rho]$$

Consistency of spectral filters [Baldassarre et al., 2010b]

Theorem - Finite sample bound on the Excess Risk

Let $\mathbf{f}_{\mathbf{z}}^{\lambda_\mathbf{n}}$ be the **estimator** obtained with a spectral filter $\mathbf{g}_{\lambda_\mathbf{n}}$,
where $\lambda(n) = \lambda_n$. **Fix** a confidence $0 < \eta < 1$.
Given *reasonable assumptions* on f_ρ, $\mathcal{Y}$ and the kernel Γ, we have

$$\mathrm{I}(f_{\mathbf{z}}^{\lambda_n}) - \mathrm{I}(f_\rho) \leq \frac{C \log 4/\eta}{\sqrt{n}}$$

with probability $1 - \eta$.
C is a constant that depends on the assumptions and other constants characterizing the spectral filters.

Consistency of spectral filters [Baldassarre et al., 2010b]

Theorem - Finite sample bound on the Excess Risk

Let $\mathbf{f}_{\mathbf{z}}^{\lambda_n}$ be the **estimator** obtained with a spectral filter $\mathbf{g}_{\lambda_n}$, where $\lambda(n) = \lambda_n$. **Fix** a confidence $0 < \eta < 1$.
Given *reasonable assumptions* on f_ρ, $\mathcal{Y}$ and the kernel Γ, we have

$$\mathrm{I}(f_{\mathbf{z}}^{\lambda_n}) - \mathrm{I}(f_\rho) \leq \frac{C \log 4/\eta}{\sqrt{n}}$$

with probability $1 - \eta$.
C is a constant that depends on the assumptions and other constants characterizing the spectral filters.

Theorem - Consistency

$$\lim_{n \to \infty} \mathrm{P}\left[\mathrm{I}(f_{\mathbf{z}}^{\lambda_n}) - \mathrm{I}(f_\rho) > \varepsilon\right] = 0$$

for any $\varepsilon > 0$

Results for a special class of kernels

Decomposable kernels [Caponnetto et al., 2008]

$$\Gamma(x, x') = K(x, x')A$$

- $K : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ is a scalar kernel that encodes the similarity between the input points.
- A is a positive semi-definite $d \times d$ matrix that encodes the relationships between the outputs

Results for a special class of kernels

Decomposable kernels [Caponnetto et al., 2008]

$$\Gamma(x, x') = K(x, x')A$$

- $K : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ is a scalar kernel that encodes the similarity between the input points.
- A is a positive semi-definite $d \times d$ matrix that encodes the relationships between the outputs

Proposition [Baldassarre et al., 2010b]

Let $f = (f^1, \ldots, f^d)$, with $f \in \mathcal{H}_K$, then if $\Gamma = KA$

$$||f||^2_\Gamma = \sum_{\ell,q=1}^{d} A^\dagger_{\ell q} \langle f^\ell, f^q \rangle_K$$

where $A^\dagger$ is the pseudo-inverse of A.

Results for a special class of kernels

Decomposable kernels [Caponnetto et al., 2008]

$$\Gamma(x, x') = K(x, x')A$$

- $K : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ is a scalar kernel that encodes the similarity between the input points.
- A is a positive semi-definite $d \times d$ matrix that encodes the relationships between the outputs

Decomposition scheme [Baldassarre et al., 2010b]

The vector valued learning problem can be **decomposed** into **d** essentially independent scalar problems, where the output data is *projected onto the eigenvectors of the matrix* **A**, with a **reduction** in computational complexity (i.e. speed).

Outline

Helmholtz Theorem and kernels

Helmoltz Theorem

A vector field that is

1. twice continuous differentiable and
2. vanishes faster than $1/r$ at infinity

can be decomposed into a divergence-free and a curl-free part.

Helmholtz Theorem and kernels

Helmoltz Theorem

A vector field that is

1. twice continuous differentiable and
2. vanishes faster than $1/r$ at infinity

can be decomposed into a divergence-free and a curl-free part.

Divergence-free and curl-free kernels

[Macêdo and Castro, 2008] introduced two kernels, Γ_{df} and Γ_{cf}, that yield vector fields that are either divergence-free or curl-free.

Helmholtz Theorem and kernels

Helmoltz Theorem

A vector field that is

1. twice continuous differentiable and
2. vanishes faster than $1/r$ at infinity

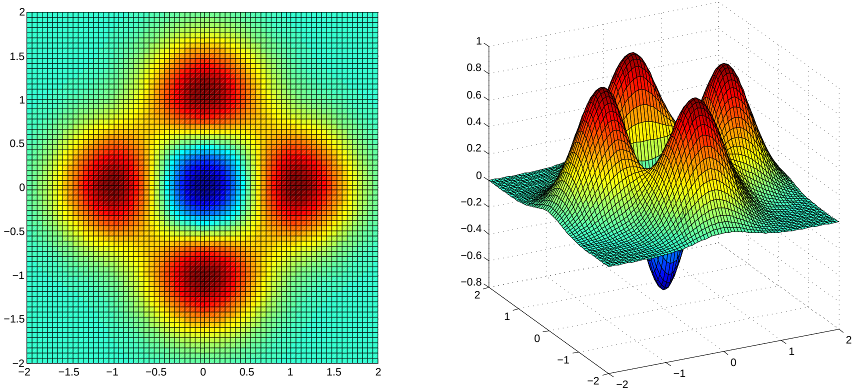
can be decomposed into a divergence-free and a curl-free part.

Divergence-free and curl-free kernels

[Macêdo and Castro, 2008] introduced two kernels, Γ_{df} and Γ_{cf}, that yield vector fields that are either divergence-free or curl-free.

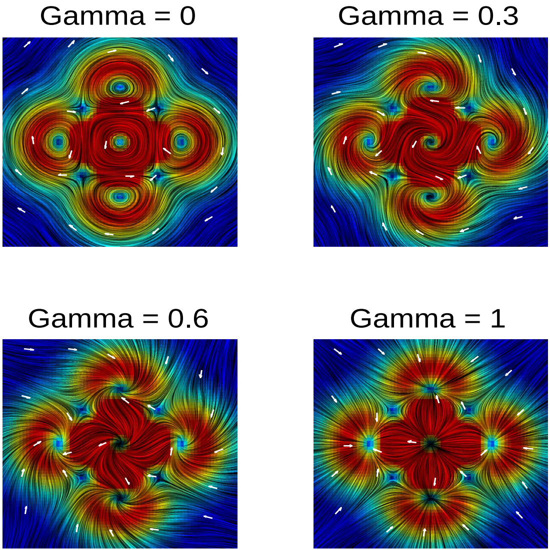
With a kernel $\Gamma = \gamma\Gamma_{df} + (1-\gamma)\Gamma_{cf}$ it is possible to learn a vector field that satisfies the hypothesis of Helmholtz Theorem and reconstruct the two parts separately.

Vector Field [Baldassarre et al., 2010b]



1. Compute the gradient and the field perpendicular to it
2. Consider a convex combination of these two vector fields, controlled by a parameter γ

Vector Field

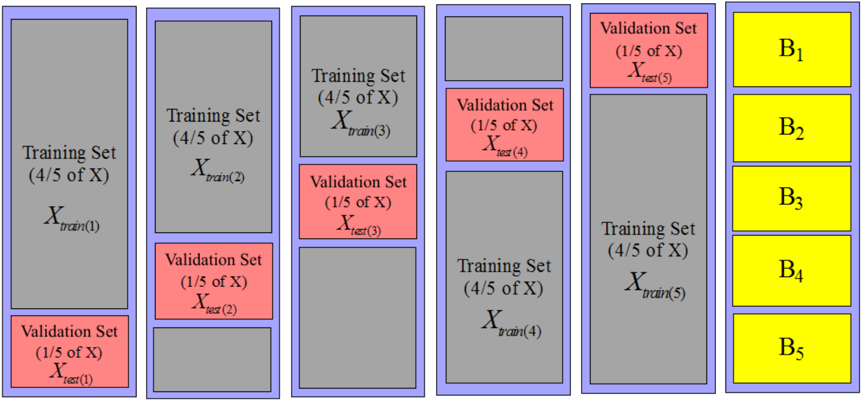


Experimental Protocol

- Field computed on a 70×70 grid on the $[-2\ 2]^2$ square
- Sampling of $10, 20, \ldots, 100, 150, 200, 400, 600$ points for training
- Remaining points used for evaluating performance
- Model parameters found via 5-fold Cross Validation

Experimental Protocol

- Field computed on a 70×70 grid on the $[-2\ 2]^2$ square
- Sampling of $10, 20, \ldots, 100, 150, 200, 400, 600$ points for training
- Remaining points used for evaluating performance
- Model parameters found via 5-fold Cross Validation



Evaluation

- We use the ν-method for learning (it is the fastest!)

Evaluation

- We use the ν-method for learning (it is the fastest!)
- We use the divergence-free and curl-free kernels

Evaluation

- We use the ν-method for learning (it is the fastest!)
- We use the divergence-free and curl-free kernels
- Model parameters:
  - Number of iterations
  - Weight balancing the two kernels

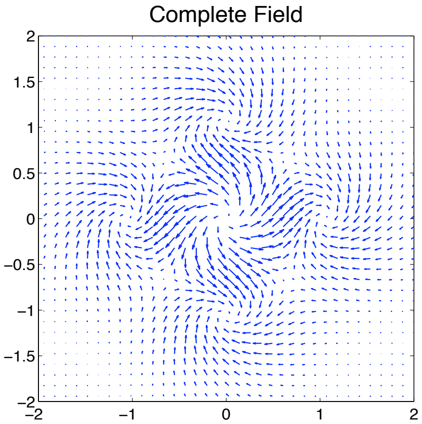
Evaluation

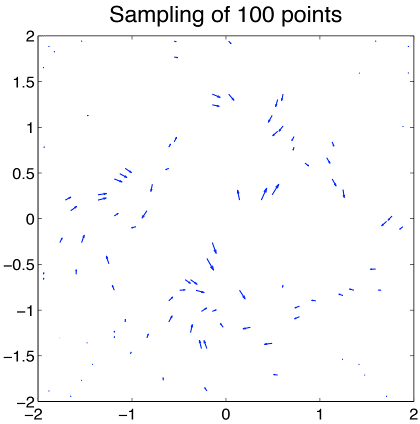
- We use the ν-method for learning (it is the fastest!)
- We use the divergence-free and curl-free kernels
- Model parameters:
  - Number of iterations
  - Weight balancing the two kernels
- We first consider the case without output noise

Evaluation

- We use the ν-method for learning (it is the fastest!)
- We use the divergence-free and curl-free kernels
- Model parameters:
  - Number of iterations
  - Weight balancing the two kernels
- We first consider the case without output noise
- Secondly we treat the case with independent gaussian noise of standard deviation 0.3

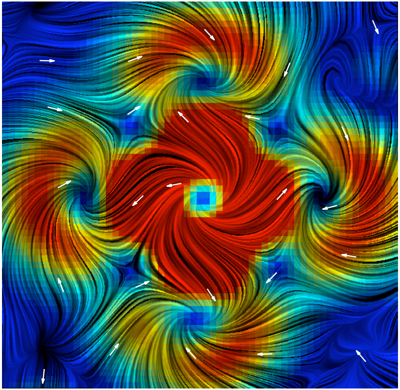
Vector field - $\gamma = 0.5$



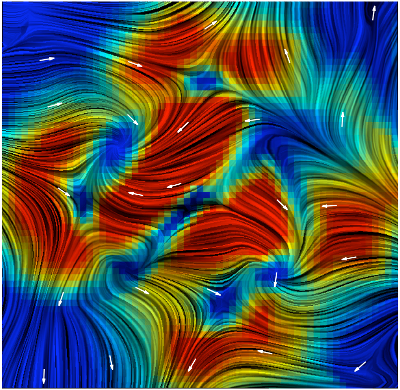


Vector field - $\gamma = 0.5$

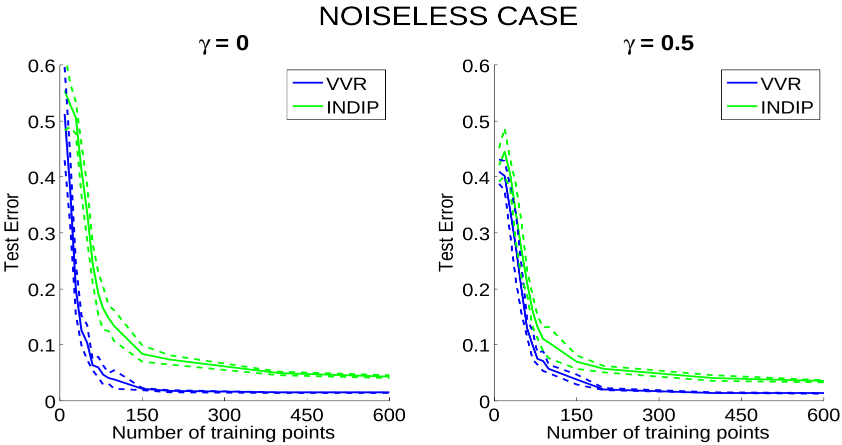
Reconstruction with my method



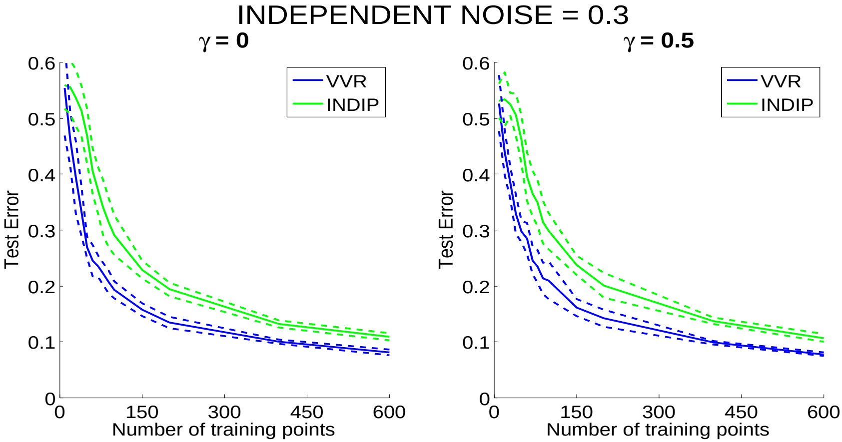
Reconstruction with interpolation



Vector field - Results I [Baldassarre et al., 2010b]

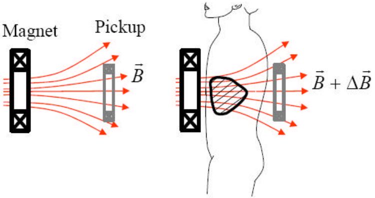


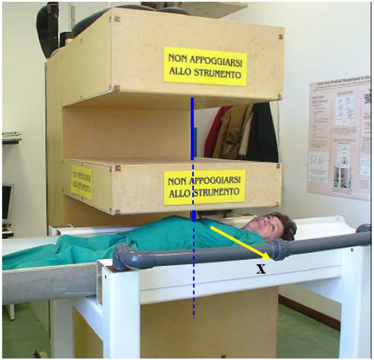
Vector field - Results II [Baldassarre et al., 2010b]



MID - The medical problem

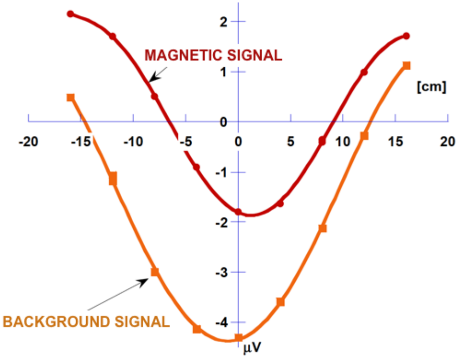
- The treatment of *Thalassemia* and *Hemochromatosis* requires the evaluation of the **iron overload** in the patient **liver**.
- The biosusceptometer MID can evaluate the iron overload in a **non-invasive** manner [Marinelli et al., 2006, Marinelli et al., 2007].
- The transductor measures the **magnetic field variation** when the patient is positioned between the magnet and the pickup.
- The magnetic field variation depends on the **geometry of the patient**, on the **magnetic properties of the tissues** and on the **patient position** (X axis).





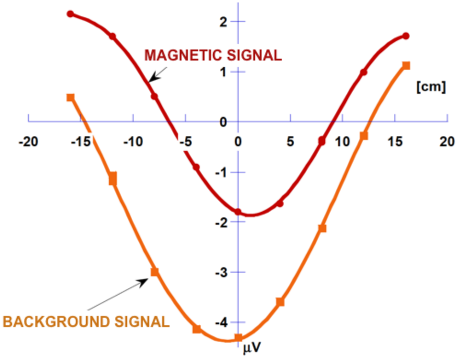
MID - The signal

- **STEP 1**: Measurement of the patient's **magnetic signal**;



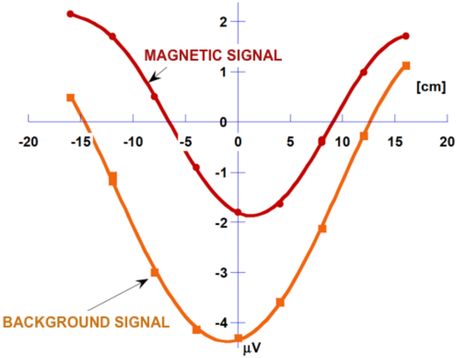
MID - The signal

- **STEP 1**: Measurement of the patient's **magnetic signal**;
- **STEP 2** : Estimation of the patient's magnetic track without iron overload (**background signal**).



MID - The signal

- **STEP 1**: Measurement of the patient's **magnetic signal**;
- **STEP 2** : Estimation of the patient's magnetic track without iron overload (**background signal**).
- **STEP 3**: The amount of iron overload is obtained using the **difference** between the two signals



MID - The model I

The idea

The **background signal** of a patient is *similar* to the **magnetic signal** of a healthy person with *similar* anthropometric features (height, weight, body shape, BMI etc).

MID - The model I

The idea

The **background signal** of a patient is *similar* to the **magnetic signal** of a healthy person with *similar* anthropometric features (height, weight, body shape, BMI etc).

The data

In order to estimate the background signal, we used the magnetic signals recorded from a pool of **84 volunteers**.

MID - The model II [Baldassarre et al., 2008]

Vector valued model

- The input examples contain the anthropometric features.
- The output examples contain the measures of the magnetic signal.

MID - The model II [Baldassarre et al., 2008]

Vector valued model

- The input examples contain the anthropometric features.
- The output examples contain the measures of the magnetic signal.
- Each measure is considered as a component of a vector.
- The measures lie on a parabola with a small approximation error.

MID - The model II [Baldassarre et al., 2008]

Vector valued model

- The input examples contain the anthropometric features.
- The output examples contain the measures of the magnetic signal.
- Each measure is considered as a component of a vector.
- The measures lie on a parabola with a small approximation error.
- We design a matrix-valued kernel that imposes a parabolic correlation among the components

$$\Gamma(x, x')_{pq} = (x \cdot x')(1 + t_p t_q + t_p^2 t_q^2)$$

  with t indicating the measurement position.
- It is of the form $\Gamma = KA$, with K a simple linear kernel.

MID - The Experimental Protocol

- No test set available to compare the algorithms.
- A first Leave-One-Out Cross Validation to evaluate performance.
- On the remaining N-1 examples perform another LOOCV to select optimal algorithm parameters.

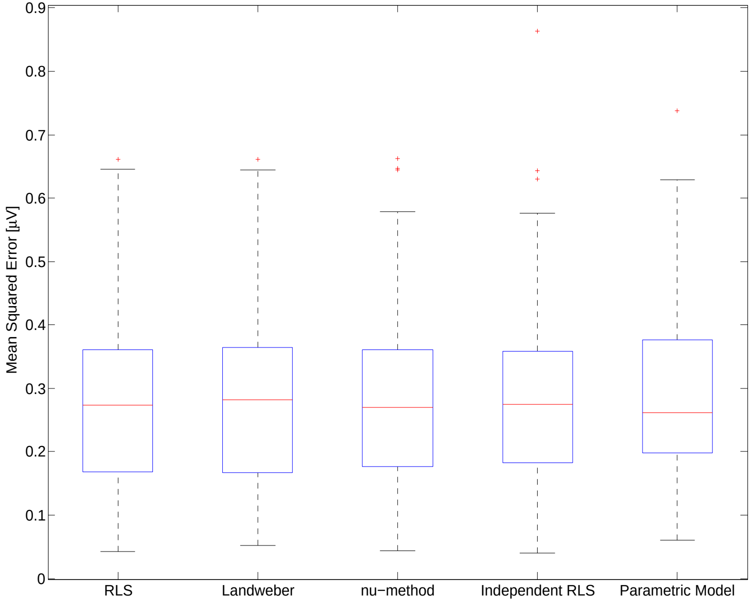
MID - The Experimental Protocol

- No test set available to compare the algorithms.
- A first Leave-One-Out Cross Validation to evaluate performance.
- On the remaining N-1 examples perform another LOOCV to select optimal algorithm parameters.
- Compare vector valued Tikhonov (RLS), Landweber, ν-method and scalar Tikhonov on each measure separately.

Algorithm	Average Time [s]
Tikhonov	4.4
Landweber	1.2
ν-method	0.31
Independent Tikhonov	0.17

Table: Average computation times for each loop of the first LOOCV.

MID - Results (84 volunteers) [Baldassarre et al., 2008]



MID - Considerations

Considerations

- Our model is now used at the Hospital since it has proven *more robust* in estimating the signal for patients that are poorly represented by the volunteer populations (i.e. small kids, very fat or very slim people).

MID - Considerations

Considerations

- Our model is now used at the Hospital since it has proven *more robust* in estimating the signal for patients that are poorly represented by the volunteer populations (i.e. small kids, very fat or very slim people).
- Our model seems to *generalize* better...

MID - Considerations

Considerations

- Our model is now used at the Hospital since it has proven *more robust* in estimating the signal for patients that are poorly represented by the volunteer populations (i.e. small kids, very fat or very slim people).
- Our model seems to *generalize* better...
- and is *faster*!

MID - Considerations

Considerations

- Our model is now used at the Hospital since it has proven *more robust* in estimating the signal for patients that are poorly represented by the volunteer populations (i.e. small kids, very fat or very slim people).
- Our model seems to *generalize* better...
- and is *faster*!
- The kernel adopted might not reflect the real dependencies among the measures.

MID - Considerations

Considerations

- Our model is now used at the Hospital since it has proven *more robust* in estimating the signal for patients that are poorly represented by the volunteer populations (i.e. small kids, very fat or very slim people).
- Our model seems to *generalize* better...
- and is *faster*!
- The kernel adopted might not reflect the real dependencies among the measures.
- The anthropometric features measure do not correlate enough with the magnetic signal.

Outline

1. Motivations
2. Supervised learning
   - Supervised learning basics
   - Problem setting
   - Spectral filters
   - Theoretical results
3. Experiments
   - Simulated vector fields
   - Magnetic Iron Detector
4. Conclusions

Conclusions

Main contributions

- Connection between the norm in a vector valued RKHS to regularization terms on the components.
- Finite sample bound on the excess risk.
- **Faster learning scheme when** $\Gamma = KA$.
- **Complexity analysis**.
- **Multi-class and Multi-task extensions**.
- Real world applications:
  1. MID [Baldassarre et al., 2008]
  2. **HAND** [Noceti et al., 2009, Baldassarre et al., 2010a]

Conclusions

Main contributions

- Connection between the norm in a vector valued RKHS to regularization terms on the components.
- Finite sample bound on the excess risk.
- **Faster learning scheme when** $\Gamma = KA$.
- **Complexity analysis**.
- **Multi-class and Multi-task extensions**.
- Real world applications:
  1. MID [Baldassarre et al., 2008]
  2. **HAND** [Noceti et al., 2009, Baldassarre et al., 2010a]

Open problems

- No kernel good for all seasons, especially for multi-class.
- Estimation of the kernel from the data.
- Incorporation of prior information **not** in the kernel.

Baldassarre, L., Barla, A., Gianesin, B., and Marinelli, M. (2008).
Vector valued regression for iron overload estimation.
In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*.

Baldassarre, L., Barla, A., Noceti, N., and Odone, F. (2010a).
Learning how to grasp objects.
In *Proceedings of ESANN*.

Baldassarre, L., Barla, A., Rosasco, L., and Verri, A. (2010b).
Multi-output learning via spectral filtering.
Machine Learning (submitted).

Bühlmann, P. and Yu, B. (2002).
Boosting with the l_2-loss: Regression and classification.
Journal of American Statistical Association, 98:324–340.

Caponnetto, A., Micchelli, C., Pontil, M., and Ying, Y. (2008).
Universal kernels for multi-task learning.
Journal of Machine Learning Research, 9:1615–1646.

Carmeli, C., De Vito, E., and Toigo, A. (2006).
Vector valued reproducing kernel Hilbert spaces of integrable functions and Mercer theorem.
Anal. Appl. (Singap.), 4(4):377–408.

Cucker, F. and Smale, S. (2001).
On the mathematical foundations of learning.
Bullettin of The American Mathematical Society, 39:1–49.

De Vito, E., Rosasco, L., Caponnetto, A., De Giovannini, U., and Odone, F. (2005).
Learning from examples as an inverse problem.
Journal of Machine Learning Research, 6:883–904.

Engl, H. W., Hanke, M., and Neubauer, A. (1996).
Regularization of inverse problems, volume 375 of *Mathematics and its Applications*.
Kluwer Academic Publishers Group, Dordrecht.

Evgeniou, T., Pontil, M., and Poggio, T. (2000).
Regularization networks and support vector machines.

Advances in Computational Mathematics, 13(1):1–50.

Girosi, F., Jones, M., and Poggio, T. (1995).
Regularization theory and neural networks architectures.
Neural computation, 7(2):219–269.

Lo Gerfo, L., Rosasco, L., Odone, F., De Vito, E., and Verri, A. (2008).
Spectral algorithms for supervised learning.
Neural Computation.

Macêdo, I. and Castro, R. (2008).
Learning divergence-free and curl-free vector fields with matrix-valued kernels.
Technical report, Instituto Nacional de Matematica Pura e Aplicada.

Marinelli, M., Cuneo, S., Gianesin, B., Lavagetto, A., Lamagna, M., Oliveri, E., Sobrero, G., Terenzani, L., and Forni, G. (2006).
Non-invasive measurement of iron overload in the human body.
IEEE Transactions on Applied Superconductivity, 16(2).

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

Micchelli, C. and Pontil, M. (2005).
On learning vector–valued functions.
Neural Computation, 17:177–204.

Noceti, N., Caputo, B., Castellini, C., Baldassarre, L., Barla, A., Rosasco, L., Odone, F., and Sandini, G. (2009).
Towards a theoretical framework for learning multi-modal patterns for embodied agents.
In *IEEE Proceedings of ICIAP*.

Tikhonov, A. N. and Arsenin, V. Y. (1977).
Solutions of Ill-posed Problems.
John Wiley.

Vapnik, V. and Chervonenkis, A. (1974).

Theory of pattern recognition.
Nauka, Moscow.

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