Supervised Nonlinear Factorizations Excel In Semi-supervised Regression

Josif Grabocka[1], Erind Bedalli[2], and Lars Schmidt-Thieme[1]

[1]ISML Lab, University of Hildesheim
Samelsonplatz 22, 31141 Hildesheim, Germany
{josif,schmidt-thieme}@ismll.uni-hildesheim.de
[2]Department of Mathematics and Informatics, University of Elbasan
Rruga Rinia, Elbasan, Albania
erind.bedalli@uniel.edu.al

Abstract. Semi-supervised learning is an eminent domain of machine learning focusing on real-life problems where the labeled data instances are scarce. This paper innovatively extends existing factorization models into a supervised nonlinear factorization. The current state of the art methods for semi-supervised regression are based on supervised manifold regularization. In contrast, the latent data constructed by the proposed method jointly reconstructs both the observed predictors and target variables via generative-style nonlinear functions. Dual-form solutions of the nonlinear functions and a stochastic gradient descent technique which learns the low dimensionality data are introduced. The validity of our method is demonstrated in a series of experiments against five state-of-art baselines, clearly improving the prediction accuracy in eleven real-life data sets.

Keywords: Supervised Matrix Factorization, Nonlinear Dimensionality Reduction, Feature Exctraction

1 Introduction

Regression is a core task of machine learning, aiming at identifying the relationship between a series of predictor variables and a special target variable (labeled instances) of interest [1]. Practitioners often face budget constraints in recording/measuring instances of the target variable, in particular due to the need for domain expertise [2]. On the other hand, the instances composed of predictor variables alone (unlabeled instances) appear in abundant amounts because they typically originate from less expensive automatic processes. Eventually the research community realized the potential of unlabeled instances as an important guidance in the learning process, establishing the rich domain of semi-supervised learning [2]. Semi-supervised learning is expressed on two flavors: regression and classification, depending on the metric used to evaluate the prediction of the target variable.

The principle of incorporating unlabeled instances relies heavily on exploring the geometric structure of unlabeled data, addressing the synchronization of the

detected structural regularities against the positioning of the labeled instances. A stream of research focuses on the notion of clusters, where the predicted target values were influenced by connections to labeled instances through dense data regions [3, 4]. The other prominent stream elaborates on the idea that data is closely encapsulated in a reduced dimensionality space, known as the manifold principle. Subsequently, the method of Manifold Regularization restricted the learning algorithm by imposing the manifold geometry via the addition of structural regularization penalty terms [5]. Discretized versions of the manifold regularization highlighted the structural understanding of data through elaborating the graph Laplacian regularization [6]. The extrapolating power of manifold regularization have been extended to involve second-order Hessian energy regularization [7], and parallel vector field regularization [8].

Throughout this study we introduce a semi-supervised regression model. The underlying foundation of our approach considers the observed data variables to be dependent on a smaller set of hidden/latent variables. The proposed method builds a low-rank representation of the data which can reconstruct both the predictor variables and the target variable via nonlinear functions. The target variable is utilized in guiding the reduction process, which in comparison to unsupervised methods, help filtering only those features which boosts the target prediction accuracy [9, 10]. The proposed method operates by constructing latent nonlinear projections, opposing techniques guiding the reconstruction linearly [9]. Therefore we extend supervised matrix factorization into non-linear capabilities. The nonlinear matrix factorization belongs to the family of models known as Gaussian Process Latent Variable Modeling [11]. Our stance on nonlinear projections is further elaborated in Section 3.

The *modus operandi* of our paper is defined as a joint nonlinear reconstruction of the predictors and target variable by optimizing the regression quality over the training data. The nonlinear functions are defined as regression weights in a mapped data space, which are expressed and learned in the dual-form using the kernel theory. In addition, a stochastic gradient descent algorithm is introduced for updating the latent data based on the learned dual regression weights. In the context of semi-supervision our model can operate with very few labeled instances. Detailed explanation of the method and all necessary derivations are described in Section 4.

No previous paper has attempted to compare factorization approaches against the state of the art in manifold regularization, regarding semi-supervised regression problems. In order to demonstrate the superiority of the presented method we implemented and compared against five strong state-of-art methods. A battery of experiments over eleven real life datasets at varying number of labeled instances is conducted. Our method clearly outperforms five state-of-art baselines in the vast majority of the experiments as discussed in Section 5. The main contributions of this study are:

- Formulated a supervised nonlinear factorizations model
- Developed a learning algorithm in the dual formulation

– Conducted a throughout empirical analysis against the state of the art (manifold regularization)

2 Related Work

Even though a plethora of **regression** models have been proposed, yet Support Vector Machines (SVMs) are among the strongest general purpose learning models. A particular implementation of SVMs tailored for approximating square error loss is called Least Square SVMs (LS-SVM) [12], and is shown to perform equivalently to the epsilon loss regression SVMs [13]. This study empirically compares against LS-SVM, in order to demonstrate the additive gain of incorporating unlabeled information.

The **semi-supervised regression** research was boosted by the elaboration of the unlabeled instances' structure into the regression models. A major stream explored the cluster notion in utilizing high density unlabeled instances' regions for predicting the target values [3, 4]. The other stream, called Manifold Regularization, assumes the data lie on a low-dimensional manifold and that the structure of the manifold should be respected in regressing target values of the unlabeled instances [5]. A discretized variant of the regularization was proposed to include the graph Laplacian representation of the unlabeled data as a penalty term [6]. The regularization of the manifold surfaces have been extended to involve second-order Hessian energy regularization [7], while a formalization of the vector field theory was employed in the so-called Parallel Field Regularization (PFR) [8]. In addition, a recent elaboration of surface smoothing included energy minimizations called total variation and Euler's elastica [14]. Another study attempts to discover eigenfunctions of the integral operator derived from both labeled and unlabeled instances [15], while efforts have extended to incorporate kernel theory to manifold regularization [16]. In this study we compare against three of the strongest baselines, the Laplacian regularization, the Hessian Regularization and the PFR regularization. In contrast to these existing approaches, our novel method explores hidden data structures via latent nonlinear reconstructions of both predictors and target variable.

Supervised Dimensionality Reduction involves label information as a guidance for dimensionality reduction. The Linear Discriminant Analysis is the pioneer of supervised decomposition [17]. SVMs were adjusted to high dimensional data through reducing the dimensionality via kernel matrix decomposition [18]. Generalized linear models [19] and Bayesian mixture modeling [10] have also been combined with supervised dimensionality reduction. Furthermore convolutional and sampling layers of convolutional networks are functioning as supervised decomposition [20]. The field of Gaussian Process Latent Variable Models (GPLVM) aims at detecting latent variables through a set of functions having a joint Gaussian distribution [21]. A similar model to ours has utilized GPLVM for pose estimation in images [22].

Due to its empirical success, matrix factorization has been employed in detecting latent features, while supervised matrix factorization is engineered to

emphasize the target variable [9]. For the sake of clarity, methods that reduce the dimensionality in a nonlinear fashion such as the kernel PCA [23], or kernel non-negative matrix factorization [24], should not be confused with the proposed method, because such methods are unsupervised in terms of target variable. Our method offers novelty compared to state-of-art techniques in proposing joint nonlinear reconstruction of both predictors and target variables, in a semi-supervised fashion, from a minimalistic latent decomposition through dual-form nonlinearity.

3 Elaborated Principle

The majority of machine learning methods expect a target variable to be a consequence of, or directly related to, the predictor variables. This study operates over the hypothesis that both the predictors and the target variables are *observed* effects of other hidden *original* factors/variables which are not recorded/known. Our method extracts original variables which can jointly approximate both predictors and target variables in a nonlinear fashion. The current study claims that original variables contain less noise and therefore better predict the target variable, while empirical results of Section 5 demonstrate its validity.

Let us assume the unknown original data to be composed of D-many hidden variables in N training and N' testing instances and denoted as $Z \in \mathbb{R}^{(N+N')\times D}$. Assume we could observe M-many predictor variables $X \in \mathbb{R}^{(N+N')\times M}$ and one target variable $Y \in \mathbb{R}^N$, with the aim of accurately predicting the test targets $Y_t \in \mathbb{R}^{N'}$. Semi-supervised scenarios where $N' > N$ are taken into consideration. Our method learns the original variables Z and nonlinear functions $g_j, h \in \mathbb{R}^D \rightarrow \mathbb{R}$ which can *jointly* approximate X, Y. Equation 1 describes the idea, while we included natural Gaussian noise with variance σ_X, σ_Y in the process. We introduce a syntactic notation $\mathbb{N}_a^b = \{a, a+1, \ldots, b-1, b\}$.

$$X_{i,j} = g_j(Z_{i,:}) + \mathcal{N}(0, \sigma_X) \; ; \quad Y_i = h(Z_{i,:}) + \mathcal{N}(0, \sigma_Y) \qquad (1)$$
$$i \in \mathbb{N}_1^{N+N'} \; , \; j \in \mathbb{N}_1^M$$

4 Supervised Nonlinear Factorizations (SNF)

As aforementioned, our novelty relies on learning a latent low-rank representation Z from observed data X, Y, such that the predictor variables and the target variable are *jointly* reconstructible from the low-rank data via *nonlinear* functions. Nonlinearity is achieved by expanding the low-rank data Z to a (probably much) higher-dimensional space $\mathbb{R}^F$ space via a mapping $\psi : \mathbb{R}^D \rightarrow \mathbb{R}^F$. Linear hyperplanes $V \in \mathbb{R}^{F\times M}, W \in \mathbb{R}^F$ with bias terms $V^0 \in \mathbb{R}^M, W^0 \in \mathbb{R}$ can therefore approximate X, Y in the mapped space as described in Equation 2.

$$\hat{X}_{i,j} = \langle \psi(Z_{i,:}), V_{:,j} \rangle + V_j^0 \; ; \quad \hat{Y}_i = \langle \psi(Z_{i,:}), W \rangle + W^0 \qquad (2)$$
$$i \in \mathbb{N}_1^{N+N'} \; , \; j \in \mathbb{N}_1^M$$

4.1 Maximum Aposteriori Optimization

Consecutively the objective is to maximize the joint likelihood of the predictors X, target Y and the maximum aposteriori estimators V, V^0, W, W^0 as shown in Equation 3. The hyperplanes parameters incorporate normal priors $V \sim \mathcal{N}(0, \lambda_V^{-1}), W \sim \mathcal{N}(0, \lambda_W^{-1})$. The distribution of the observed variables is also assumed normal $X \sim \mathcal{N}(\langle \psi(Z), V \rangle, \sigma_X)$ and $Y \sim \mathcal{N}(\langle \psi(Z), W \rangle, \sigma_Y)$ and independently distributed. The logarithmic likelihood, depicted in Equation 4 converts the objective to a summation of terms.

$$\underset{\psi(Z),V,V^0,W,W^0}{\operatorname{argmax}} \prod_{i=1}^{N+N'} p(X_{i,:}|\psi(Z_{i,:}), V, V^0)\; p(V) \prod_{l=1}^{N} p(Y_l|\psi(Z_{l,:}), W, W^0)\; p(W) \tag{3}$$

$$\underset{\psi(Z),V,V^0,W,W^0}{\operatorname{argmax}} \sum_{i=1}^{N+N'} \log\left(p(X_{i,:}|\psi(Z_{i,:}), V, V^0)\right) + \sum_{l=1}^{N} \log\left(p(Y_l|\psi(Z_{l,:}), W, W^0)\right) + \sum_{j=1}^{M} \log\left(p(V_{:,j})\right) + \log\left(p(W)\right) \tag{4}$$

Inserting the normal probability into Equation 4 converts logarithmic likelihoods into L2 norms with Tikhonov regularization terms as shown in Equation 5 and the variance terms σ_X, σ_Y drop out as constants. An additional biased regularization term $\lambda_Z \langle Z, Z \rangle$ is included in order to help the latent data avoid over-fitting.

$$\begin{aligned} &\underset{Z,V,V^0,W,W^0}{\operatorname{argmin}} \left(\sum_{j=1}^{M} \langle \xi_{:,j}, \xi_{:,j} \rangle + \lambda_V \langle V_{:,j}, V_{:,j} \rangle \right) + \langle \phi, \phi \rangle + \lambda_W \langle W, W \rangle + \lambda_Z \langle Z, Z \rangle \\ &\text{subject to: } \xi_{i,j} = X_{i,j} - \langle \psi(Z_{i,:}), V_{:,j} \rangle - V_j^0,\ i \in \mathbb{N}_1^{N+N'},\ j \in \mathbb{N}_1^M \\ &\qquad\qquad \phi_l = Y_l - \langle \psi(Z_{l,:}), W \rangle - W^0, \qquad l \in \mathbb{N}_1^N \end{aligned} \tag{5}$$

Computing the $\psi(Z)$ directly is intractable, therefore we will derive the dual-form representation in the next Section 4.2, where the kernel trick will be utilized to compute Z in the original space $\mathbb{R}^D$.

4.2 Dual-Form Solution - Learning the Nonlinear Regression Weights

The optimization of Equation 5 is carried on in an alternated fashion. Hyperplane weights V, V^0, W, W^0 are converted to dual variables and then solved by keeping Z fixed, while in a second step Z is solved keeping the dual weights fixed. This section is dedicated to learning the nonlinear weights in the dual-form. Each of the M-many predictors loss terms from Equation 5, (one per each

predictor variable $X_{:,j}$) can be learned isolated as described in the sub-objective function J_j of Equation 6. To facilitate forthcoming derivations we multiplied the objective function by $\frac{1}{2\lambda_V}$.

$$\underset{V_{:,j},V_j^0}{\operatorname{argmin}} \; J_j = \frac{1}{2\lambda_V}\langle \xi_{:,j}, \xi_{:,j} \rangle + \frac{1}{2}\langle V_{:,j}, V_{:,j} \rangle \tag{6}$$
$$\xi_{i,j} = X_{i,j} - \langle \psi(Z_{i,:}), V_{:,j} \rangle - V_j^0$$

In order to optimize Equation 6, the equality conditions are added to the objective function through Lagrange multipliers $\alpha_{i,j}$. The inner minimization objective is solved by computing stationary solution points $V_{:,j}, V_j^0, \xi_{:,j}$ and eliminating out the first derivatives ($\frac{\partial L_j}{\partial V_{:,j}} = 0, \frac{\partial L_j}{\partial \xi_{:,j}} = 0, \frac{\partial L_j}{\partial V_j^0} = 0$) as shown in Equation 7.

$$\begin{aligned} \underset{\alpha_{:,j},V_j^0}{\operatorname{argmax}} \; \underset{V_{:,j},V_j^0,\xi_{:,j}}{\operatorname{argmin}} \; L_j &= \frac{1}{2\lambda_V}\langle \xi_{:,j}, \xi_{:,j} \rangle + \frac{1}{2}\langle V_{:,j}, V_{:,j} \rangle \\ &+ \sum_{i=1}^{N+N'} \alpha_{i,j} \left(X_{i,j} - \langle \psi(Z_{i,:}), V_{:,j} \rangle - V_j^0 - \xi_{i,j} \right) \\ \rightarrow \; V_{:,j} &= \sum_{i=1}^{N+N'} \alpha_{i,j} \psi(Z_{i,:}) \\ \xi_{:,j} &= \lambda_V \alpha_{:,j} \\ \sum_{i=1}^{N+N'} \alpha_{i,j} &= 0 \end{aligned} \tag{7}$$

Replacing the stationary point solution of $V_{:,}, \xi{:}, j, V_j^0$ back into the objective function 7, we get rid of the variables $V_{:,j}, \xi_{:,j}$, yielding Equation 8.

$$\begin{aligned} \underset{\alpha_{:,j},V_j^0}{\operatorname{argmax}} \; &- \sum_{i=1,l=1}^{N+N^t} \alpha_{i,j}\alpha_{l,j} \langle \psi(Z_{i,:}), \psi(Z_{l,:}) \rangle - \lambda_V \langle \alpha_{:,j}, \alpha_{:,j} \rangle \\ &+ 2 \sum_{i=1}^{N+N^t} \alpha_{i,j} \left(X_{i,j} - V_j^0 \right) \end{aligned} \tag{8}$$

The solution of the dual maximization is given through eliminating the derivative of $\frac{\partial L}{\partial \alpha_{:,j}} = 0$ as presented in Equation 9. The kernel notation is introduced as $K_{i,l} = \langle \psi(Z_{i,:}), \psi(Z_{l,:}) \rangle$.

$$2\lambda_V \alpha_{:,j} + 2K \cdot \alpha_{:,j} + 2\langle \mathbf{1}, V_j^0 \rangle = 2X_{:,j} \; \rightarrow \; (K + \lambda_V I)\, \alpha_{:,j} + \langle \mathbf{1}, V_j^0 \rangle = X_{:,j} \tag{9}$$

Combining Equation 9 and the constraint of Equation 8, the final nonlinear reconstruction solution is given through the closed-form formulation of $\alpha_{:,j}, V_j^0$ as depicted in Equation 10.

$$\begin{bmatrix} V_j^0 \\ \alpha_{:,j} \end{bmatrix} = \begin{bmatrix} 0 & \mathbf{1}^T \\ \mathbf{1} & K + \lambda_V I \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ X_{:,j} \end{bmatrix} \tag{10}$$

Symmetrical to the predictors case, a dual-form maximization objective function is created and a *mot-a-mot* procedure like Section 4.2 can be trivially adopted in solving the nonlinear regression for the target variable. The derived solution is shown in Equation 11. Instead of using the symbol α we denote the dual-weights of the target regression dual problem using the symbol ω.

$$\begin{bmatrix} W^0 \\ \omega \end{bmatrix} = \begin{bmatrix} 0 & \mathbf{1}^T \\ \mathbf{1} & K^Y + \lambda_W I \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ Y \end{bmatrix}; \tag{11}$$
$$\text{where } K^Y_{i,l} = \langle \psi(Z_{i,:}), \psi(Z_{l,:}) \rangle; \quad i, l \in \mathbb{N}_1^N$$

A prediction of the target value of a test instance $t \in \mathbb{N}_{N+1}^{N+N'}$ is conducted using the learned dual weights as shown in Equation 12.

$$\hat{Y}_t = \sum_{i=1}^{N} \omega_i \, K^Y(Z_{i,:}, Z_{t,:}) + W^0 \tag{12}$$

Stochastic Gradient Descent - Learning the Low Dimensionality Representation A novel algorithm is applied to learn Z for optimizing Equation 8. Sub-losses composing only of $\alpha_{i,j}, \alpha_{l,j}$ are defined for all combinations $\forall i \in \mathbb{N}_1^{N+N^t}, \forall l \in \mathbb{N}_1^{N+N^t}$ and Z is updated in order to optimize each sub-loss in a stochastic gradient descent fashion as presented in Equation 13. The addition of penalty terms controlled by the hyper-parameter λ_Z which controls the regularization of Z as described in Equation 5. Our model called Supervised Nonlinear Factorizations (SNF) utilizes polynomial kernels with the derivatives needed for gradient descent represented in Equation 13.

$$\begin{aligned} Z_{i,k} &\leftarrow Z_{i,k} + \eta \left(\alpha_{i,j} \alpha_{l,j} \frac{\partial K_{i,l}}{\partial Z_{i,k}} - \lambda_Z Z_{i,k} \right) \\ Z_{l,k} &\leftarrow Z_{l,k} + \eta \left(\alpha_{i,j} \alpha_{l,j} \frac{\partial K_{i,l}}{\partial Z_{l,k}} - \lambda_Z Z_{l,k} \right) \end{aligned} \tag{13}$$

$$K_{i,l} = \left(\langle Z_{i,:}, Z_{l,:} \rangle + 1 \right)^d \;\rightarrow\; \frac{\partial K_{i,l}}{\partial Z_{r,k}} = d \left(\langle Z_{i,:}, Z_{l,:} \rangle + 1 \right)^{d-1} \times \begin{cases} Z_{l,k} & \text{if } r = i \\ Z_{i,k} & \text{if } r = l \\ 0 & \text{else} \end{cases}$$

Algorithm 1 combines all the steps of the proposed method. During each epoch all predictors' non-linear weights are solved and the latent data Z is updated. The target model is updated multiple times after each predictor model to boost convergence. The learning algorithm makes use of two different learning rates in updating Z, one for the predictors' loss (η_X) and one for the target loss (η_Y).

Algorithm 1 Learn SNF

Require: Data $X \in \mathbb{R}^{(N+N^t)\times M}, Y \in \mathbb{R}^{N^l}$, Latent Dimension: D, Learn Rates: η_X, η_Y, Number of Iterations: NumIter, Regularization parameters: $\lambda_Z, \lambda_V, \lambda_W$

Randomly set: $Z \in R^{(N+N')\times D}, V^0 \in \mathbb{R}^M, W^0 \in \mathbb{R}, \alpha \in \mathbb{R}^{(N+N')\times M}, \omega \in \mathbb{R}^{N^l}$,
for $1 \dots$ NumIter **do**
  for $j \in \{1 \dots M\}$ **do**
    Compute $K_{i,l} = K(Z_{i,:}, Z_{l,:}),\ i,l \in \mathbb{N}_1^{N+N'}$
    Solve $\begin{bmatrix} V_j^0 \\ \alpha_{:,j} \end{bmatrix} = \begin{bmatrix} 0 & \mathbf{1}^T \\ \mathbf{1} & K + \lambda_V I \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ X_{:,j} \end{bmatrix}$
    for $i \in \mathbb{N}_1^{N+N'},\ l \in \mathbb{N}_1^{N+N'},\ k \in \mathbb{N}_1^D$ **do**
      $Z_{i,k} \leftarrow Z_{i,k} + \eta_X \left(\alpha_{i,j}\alpha_{l,j} \frac{\partial K_{i,l}}{\partial Z_{i,k}} - \lambda_Z Z_{i,k} \right)$
      $Z_{l,k} \leftarrow Z_{l,k} + \eta_X \left(\alpha_{i,j}\alpha_{l,j} \frac{\partial K_{i,l}}{\partial Z_{l,k}} - \lambda_Z Z_{l,k} \right)$
    end for
    Compute $K^Y_{i,l} = K(Z_{i,:}, Z_{l,:}),\ i,l \in \mathbb{N}_1^{N}$
    Solve $\begin{bmatrix} W^0 \\ \omega \end{bmatrix} = \begin{bmatrix} 0 & \mathbf{1}^T \\ \mathbf{1} & K^Y + \lambda_W I \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ Y \end{bmatrix}$
    for $i \in \mathbb{N}_1^{N+N'},\ l \in \mathbb{N}_1^{N+N'},\ k \in \mathbb{N}_1^D$ **do**
      $Z_{i,k} \leftarrow Z_{i,k} + \eta_Y \left(\omega_i\omega_l \frac{\partial K^Y_{i,l}}{\partial Z_{i,k}} - \lambda_Z Z_{i,k} \right)$
      $Z_{l,k} \leftarrow Z_{l,k} + \eta_Y \left(\omega_i\omega_l \frac{\partial K^Y_{i,l}}{\partial Z_{l,k}} - \lambda_Z Z_{l,k} \right)$
    end for
  end for
end for
return $Z, \alpha, V^0, \omega, W^0$

5 Empirical Results

Five strong state of the art baselines and empirical evidence over eleven datasets are mainly the outline of our experiments, which will be detailed in this section, together with the results and their interpretation.

5.1 Baselines

The proposed method Supervised Nonlinear Factorizatoins (SNF) is compared against the following five baselines:

- **Least Square Support Vector Machines (LS-SVM)** [12] is a strong general purpose regression model and the comparisons against it will show the gain of incorporating unlabeled instances.
- **Laplacian Manifold Regularization (Laplacian) [5], Hessian Energy Regularization (Hessian) [7], Parallel Field Regularization (PFR)[8]** are strong state-of-art baselines belonging to the popular field of manifold regularization. Comparing against them gives an insight into the state-of-art quality of our results.

– **Linear Latent Reconstructions (LLR)** [9] offers the possibility to understand the addiditive benefits of exploring nonlinear projections compared to plain linear ones.

5.2 Reproducibility

All our experiments were run in a three fold cross-validation mode and the hyper-parameters of our model were tuned using only train and validation data. The evaluation metric used in all experiments is the Mean Square Error (MSE).

SNF requires the tuning of seven hyper-parameters: the regularization weights $\lambda_Z, \lambda_V, \lambda_W$, the learning rates η_X, η_Y, the number of latent dimensions D and the degree of the polynomial kernel d. The search ranges of hyper-parameters are: $\lambda_Z, \lambda_V, \lambda_W \in \{10^{-6}, 10^{-5} \dots 1, 10\}; \eta_X, \eta_Y \in \{10^{-5}, 10^{-4} \dots 0.1\}; d \in \{1, 2, 3, 4\}$ while the latent dimensionality was set to one of $50\%, 75\%$ of the original dimensions. The maximum number of epochs was set to 1000. A grid search methodology was followed in finding the best combination of hyper-parameters. Please note that we followed exactly the same fair principle in computing the hyper-parameters of all baselines.

We selected eleven popular regression datasets in a random fashion from dataset repository websites. The selected datasets are AutoPrice, ForestFires, BostonHousing, MachineCPU, Mpg, Pyrimidines, Triazines, WisconsinBreastCancer from UCI[1] ; Baseball, BodyFat from StatLib[2]; Bears[3]. All the datasets were normalized between [-1,1] before usage.

5.3 Results

The experiments comparing the accuracy of our method SNF against the five strong baselines were conducted in scenarios with few labeled instances, as typically encountered in semi-supervised learning situations.

In the first experiment 5% labeled instances were selected randomly, while all other instances left unlabeled. Therefore, the competing methods had 5% target visibility and all methods, except LS-SVM, utilized the predictor variables of all the unlabeled instances. The results of the experiments are shown in Table 1. The metric of evaluation is the Mean Square Error (MSE), while both the mean MSE and the standard deviation are shown in each dataset-method cell. The winning method of each baseline is highlighted in bold. As it is distinguishable from the sum of wins, SNF outperforms the baselines in the majority of datasets (six in total), while the closest competing baseline wins in only two datasets. Furthermore in datasets such as AutoPrice, BodyFat and Mpg the improvement is significant. Even when SNF is not the winning method, the margin to the first method is not significant, as it occurs in the Baseball, ForestFires and WisconsinBreastCancer datasets.

[1] archive.ics.uci.edu
[2] lib.stat.cmu.edu
[3] people.sc.fsu.edu/∼jburkardt/datasets/triola/

Table 1. Results - MSE - Real-life Datasets (5 % Labeled Instances)

Dataset	LS-SVM	Laplacian	Hessian	PFR	LLR	SNF
A.Price	0.075 ± 0.012	0.156 ± 0.059	0.102 ± 0.036	0.067 ± 0.024	0.090 ± 0.026	**0.049 ± 0.026**
B.ball	0.093 ± 0.019	0.131 ± 0.020	**0.072 ± 0.015**	0.082 ± 0.029	0.086 ± 0.022	0.073 ± 0.017
Bears	0.205 ± 0.046	0.407 ± 0.212	0.399 ± 0.218	0.380 ± 0.220	0.292 ± 0.146	**0.130 ± 0.032**
B.Fat	0.013 ± 0.007	0.081 ± 0.002	0.039 ± 0.009	0.019 ± 0.007	0.017 ± 0.006	**0.008 ± 0.002**
F.Fires	0.020 ± 0.004	**0.0137 ± 0.015**	0.0141 ± 0.015	**0.0137 ± 0.015**	0.018 ± 0.017	0.0139 ± 0.014
B.Hous.	0.087 ± 0.027	0.123 ± 0.027	0.104 ± 0.042	0.071 ± 0.027	0.110 ± 0.031	**0.069 ± 0.024**
M.Cpu	0.061 ± 0.041	0.053 ± 0.043	0.027 ± 0.016	0.018 ± 0.012	0.030 ± 0.021	**0.015 ± 0.007**
Mpg	0.044 ± 0.006	0.074 ± 0.009	0.055 ± 0.007	0.052 ± 0.008	0.062 ± 0.015	**0.041 ± 0.006**
Pyrim	0.131 ± 0.020	0.106 ± 0.043	**0.097 ± 0.050**	0.101 ± 0.055	0.152 ± 0.065	0.102 ± 0.053
Triaz.	0.191 ± 0.014	**0.160 ± 0.026**	0.165 ± 0.031	0.166 ± 0.035	0.196 ± 0.026	0.175 ± 0.039
WiscBC.	0.608 ± 0.074	0.356 ± 0.039	0.356 ± 0.042	**0.3499 ± 0.035**	0.429 ± 0.074	0.3504 ± 0.028
Wins	**0**	**1.5**	**2**	**1.5**	**0**	**6**

Table 2. Results - MSE - Real-life Datasets (10 % Labeled Instances)

Dataset	LS-SVM	Laplacian	Hessian	PFR	LLR	SNF
A.Price	0.051 ± 0.007	0.115 ± 0.054	0.084 ± 0.042	0.048 ± 0.018	0.062 ± 0.005	**0.037 ± 0.014**
B.ball	0.127 ± 0.035	0.092 ± 0.021	**0.068 ± 0.011**	0.080 ± 0.022	0.084 ± 0.014	0.072 ± 0.010
Bears	0.160 ± 0.053	0.182 ± 0.067	0.202 ± 0.018	0.156 ± 0.051	**0.067 ± 0.021**	0.071 ± 0.020
B.Fat	0.006 ± 0.001	0.058 ± 0.006	0.015 ± 0.007	0.008 ± 0.004	0.011 ± 0.004	**0.003 ± 0.002**
F.Fires	0.018 ± 0.001	0.0146 ± 0.014	0.0140 ± 0.015	0.0142 ± 0.014	0.016 ± 0.015	**0.0139 ± 0.015**
B.Hous.	0.067 ± 0.015	0.115 ± 0.026	0.095 ± 0.024	0.067 ± 0.004	0.071 ± 0.015	**0.061 ± 0.015**
M.Cpu	0.030 ± 0.007	0.039 ± 0.025	0.041 ± 0.030	0.020 ± 0.011	0.024 ± 0.015	**0.012 ± 0.003**
Mpg	0.071 ± 0.047	0.062 ± 0.009	0.048 ± 0.011	**0.038 ± 0.008**	0.066 ± 0.016	0.040 ± 0.004
Pyrim	0.076 ± 0.005	0.086 ± 0.042	0.076 ± 0.041	**0.069 ± 0.021**	0.107 ± 0.060	0.076 ± 0.030
Triaz.	0.265 ± 0.068	0.173 ± 0.041	**0.162 ± 0.033**	0.163 ± 0.038	0.169 ± 0.012	0.175 ± 0.009
WiscBC.	0.624 ± 0.044	**0.283 ± 0.027**	0.287 ± 0.017	0.284 ± 0.024	0.344 ± 0.079	0.307 ± 0.020
Wins	**0**	**1**	**2**	**2**	**1**	**5**

For the sake of completeness we repeated the experiments with another degree of randomly re-drawn labeled instances (10 % labeled instances). Table 2 presents the details of experiments over the selected eleven real-life datasets. The accuracy of SNF is prolonged even in this experiment. The sum of the winning methods (depicted in bold) shows that SNF wins in five of the datasets against the only two wins of the closest baseline. In particular the cases of BodyFat and MachineCpu demonstrate significant improvements in terms of MSE. As shown by the results, even in cases where our method is not the first, still it is close to the winner.

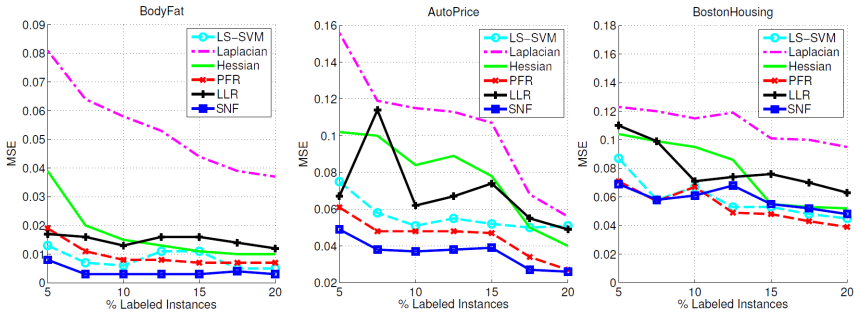

Fig. 1. Scale-up Experiments

In addition to the aforementioned results we extend our empirical analysis by conducting more fine-grained scale-up experiments with varying degree of labeled training instances. Figure 1 demonstrates the performance of all competing methods on a subset of datasets with a range of present labels varying from 5% up to 20%. SNF is seen to win in the earliest labeled percentages of the BostonHousing dataset (up to 10 %) while following in the later stages. On the contrary, we observe that our method dominates in all levels of label presence in the scaled-up experiments involving the BodyFat and AutoPrice datasets.

The accuracy of our method is grounded on a couple of reasons/observations. First of all, we would like to emphasize that each mentioned method is based on a different principle and *modus operandi*. Consequently, the dominance of a method compared to baselines depends on whether (or not) the datasets follow the principle of that particular method. Arguably the domination of SNF over manifold regularization baselines is due to the fact that our principle of mining hidden latent variables is likely (as results show) more present in general real-life datasets, therefore SNF is suited to the detection of those relations.

6 Conclusions

Throughout the present paper, a novel method that addresses the task of semi-supervised regression was proposed. The proposed method constructs a low-rank representation which jointly approximates the observed data via nonlinear functions which are learned in their dual formulation. A novel stochastic gradient descent technique is applied to learn the low-rank data using the obtained dual weights. Detailed experiments are conducted in order to compare the performance of the proposed method against five strong baselines over eleven real-life datasets. Empirical evidence over experiments in varying degrees of labeled instances demonstrate the efficiency of our method. The supervised nonlinear factorizations outperformed the manifold regularization state-of-art methods in the majority of experiments.

Acknowledgement

This research has been co-funded by the EU in FP7 via REDUCTION (#288254) and iTalk2Learn (#318051) projects.

References

1. Hastie, T., Tibshirani, R., Friedman, J.H.: The Elements of Statistical Learning. Corrected edn. Springer (July 2003)
2. Zhu, X.: Semi-supervised learning literature survey. Technical Report 1530, Computer Sciences, University of Wisconsin-Madison (2008)
3. Singh, A., Nowak, R.D., Zhu, X.: Unlabeled data: Now it helps, now it doesn't. In Koller, D., Schuurmans, D., Bengio, Y., Bottou, L., eds.: NIPS, Curran Associates, Inc. (2008) 1513–1520

4. Sinha, K., Belkin, M.: Semi-supervised learning using sparse eigenfunction bases. In: NIPS. (2009) 1687–1695
5. Belkin, M., Niyogi, P., Sindhwani, V.: Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. Journal of Machine Learning Research **7** (2006) 2399–2434
6. Melacci, S., Belkin, M.: Laplacian support vector machines trained in the primal. Journal of Machine Learning Research **12** (2011) 1149–1184
7. Kim, K.I., Steinke, F., Hein, M.: Semi-supervised regression using hessian energy with an application to semi-supervised dimensionality reduction. In: NIPS. (2009) 979–987
8. Lin, B., Zhang, C., He, X.: Semi-supervised regression via parallel field regularization. In: NIPS. (2011) 433–441
9. Menon, A.K., Elkan, C.: Predicting labels for dyadic data. Data Min. Knowl. Discov. **21**(2) (2010) 327–343
10. Mao, K., Liang, F., Mukherjee, S.: Supervised dimension reduction using bayesian mixture modeling. Journal of Machine Learning Research - Proceedings Track **9** (2010) 501–508
11. Lawrence, N.: Probabilistic non-linear principal component analysis with gaussian process latent variable models. The Journal of Machine Learning Research **6** (2005) 1783–1816
12. Suykens, J.A.K., Vandewalle, J.: Least squares support vector machine classifiers. Neural Processing Letters **9**(3) (1999) 293–300
13. Ye, J., Xiong, T.: Svm versus least squares svm. Journal of Machine Learning Research - Proceedings Track **2** (2007) 644–651
14. Lin, T., Xue, H., Wang, L., Zha, H.: Total variation and euler's elastica for supervised learning. In: ICML. (2012)
15. Ji, M., Yang, T., Lin, B., Jin, R., Han, J.: A simple algorithm for semi-supervised learning with improved generalization error bound. In: ICML. (2012)
16. Nilsson, J., Sha, F., Jordan, M.I.: Regression on manifolds using kernel dimension reduction. In: ICML. (2007) 697–704
17. Ye, J.: Least squares linear discriminant analysis. In: ICML. (2007) 1087–1093
18. Pereira, F., Gordon, G.: The support vector decomposition machine. In: Proceedings of the 23rd international conference on Machine learning. ICML '06, New York, NY, USA, ACM (2006) 689–696
19. Rish, I., Grabarnik, G., Cecchi, G., Pereira, F., Gordon, G.J.: Closed-form supervised dimensionality reduction with generalized linear models. In: Proceedings of the 25th international conference on Machine learning. ICML '08, New York, NY, USA, ACM (2008) 832–839
20. Ciresan, D.C., Meier, U., Gambardella, L.M., Schmidhuber, J.: Convolutional neural network committees for handwritten character classification. In: ICDAR, IEEE (2011) 1135–1139
21. Urtasun, R., Darrell, T.: Discriminative gaussian process latent variable models for classification. In: In International Conference in Machine Learning. (2007)
22. Navaratnam, R., Fitzgibbon, A., Cipolla, R.: The joint manifold model for semi-supervised multi-valued regression. In: Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on. (2007) 1–8
23. Hoffmann, H.: Kernel pca for novelty detection. Pattern Recogn. **40**(3) (March 2007) 863–874
24. Lee, H., Cichocki, A., Choi, S.: Kernel nonnegative matrix factorization for spectral eeg feature extraction. Neurocomputing **72**(13-15) (2009) 3182–3190