Super-Sparse Regression for Fast Age Estimation From Faces at Test Time

Ambra Demontis, Battista Biggio, Giorgio Fumera, and Fabio Roli

Dept. of Electrical and Electronic Engineering, University of Cagliari
Piazza d'Armi, 09123 Cagliari, Italy
{ambra.demontis,battista.biggio,fumera,roli}@diee.unica.it
WWW home page: http://prag.diee.unica.it

Abstract. Age estimation from faces is a challenging problem that has recently gained increasing relevance due to its potentially multi-faceted applications. Many current methods for age estimation rely on extracting computationally-demanding features from face images, and then use nonlinear regression to estimate the subject's age. This often requires matching the submitted face image against a set of face prototypes, potentially including all training face images, as in the case of kernel-based methods. In this work, we propose a super-sparse regression technique that can reach comparable performance with respect to other nonlinear regression techniques, while drastically reducing the number of reference prototypes required for age estimation. Given a similarity measure between faces, our technique learns a sparse set of virtual face prototypes, whose number is fixed a priori, along with a set of optimal weight coefficients to perform linear regression in the space induced by the similarity measure. We show that our technique does not only drastically reduce the number of reference prototypes without compromising estimation accuracy, but it can also provide more interpretable decisions.

1 Introduction

Human faces naturally convey information about people's identity, age, gender, health, and emotional state. Earlier approaches have exploited this information for improving face recognition systems, as well as for law enforcement and statistical analysis [9, 10]. With the advent of the Internet of Things, many other application scenarios have emerged, due to the increase of applications based on human-computer interactions. In particular, smart devices may change their behavior depending on the inferred user's age from their face image; *e.g.*, televisions may deny watching adult programs to children, web browsers may deny access to adult websites, and medicine cabinets may not open [9, 11, 21]. However, age estimation is a very difficult problem, as each individual ages differently, and the aging process does not only depend on one's genes, but also on various factors such as health state, stress, living environment, smoke and living style. Generally, the face aging process consists of two main phases: during childhood, the head grows significantly, changing the *geometry* of the face and distances

between fiducial points like eyes, nose, *etc.*; during adulthood, instead, the aging process involves the formation of wrinkles, changing the image *texture* [9, 10].

Current approaches for age estimation from faces have addressed this problem by exploiting and combining several feature representations, as well as classification and regression algorithms, to successfully cope with the intrinsic complexity and nonlinearity of this problem (Sect. 2); in particular, kernel methods like Support Vector Regression (SVR) and Support Vector Machines (SVM) [12, 7]. Despite these techniques may naturally yield a sparse solution, *i.e.*, they require matching the submitted face image only against a small subset of the training face images (*e.g.*, the *support vectors* in SVR and SVM), the retrieved solutions may not be sparse enough. In particular, the number of support vectors tends to grow linearly with the training set size [19, 5], causing an increase of the computational complexity at test time, and hindering the suitability of these approach for real-time, *online* age estimation [13, 9]. Some methods that exhibit a reduced complexity have also been proposed, based on the exploitation of manifold learning algorithms to map input images onto more compact, reduced feature spaces, and then applying classification or regression more efficiently in that space. However, they are mostly devoted to improve prediction accuracy rather than speed at test time [17, 12]. Furthermore, there are two other related issues: (i) the proposed systems are often very complicated, and it is thus difficult to interpret their decisions and understand why they predict a given age value (*e.g.*, it is not easy to understand which characteristics from face images are exploited by the system to discriminate between young and old people); and (ii) their inherent complexity poses a serious risk of overfitting to specific datasets.

In this paper, we propose an approach aimed to overcome these limitations (Sect. 3). It is inspired to a well-principled SVM reduction method that we recently proposed in [1] to reduce the number of face templates (*i.e.*, prototypes) in face verification systems, through the creation of a super-sparse, budgeted set of virtual vectors, *i.e.*, a fixed number of artificially-generated face templates. In this work, we extend that method and propose a super-sparse regression technique that can reach comparable performance with respect to other nonlinear regression techniques, while requiring a much smaller number of reference prototypes for age estimation at test time. Assuming that a similarity measure between faces is given, our technique jointly learns a very sparse set of virtual face prototypes, whose number is fixed a priori, and a set of optimal weight coefficients to linearly combine the similarities computed between the submitted face images (from which we aim to estimate the subject's age) and the set of reference prototypes. This allows our technique to drastically reduce the number of required prototypes without almost compromising estimation accuracy, while also providing more interpretable decisions. Our experimental analysis shows that our approach can achieve comparable performances to other state-of-the-art techniques on two well-known benchmark datasets for age estimation, while reducing of more than one order of magnitude the complexity at test time (Sect. 4). We conclude the paper by discussing contributions, limitations, and future developments of this work (Sect. 5).

2 Age Estimation from Faces

During the last years, the problem of age estimation from faces has received an increasing interest. Several works have been published since the pioneering work by Kwon and Lobo [15], in which an anthropometric model and wrinkle analyses were exploited to discriminate babies from adults. Most of the recent work nowadays exploits complex feature representation that account for geometry, shapes and textures of face images, like appearance active models [8, 12, 4], local binary patterns [22, 7], and several other visual descriptors [6, 16, 12, 3, 13]. Few works have also shown that suitable projection techniques (*e.g.*, principal component analysis, or locality preserving projection) in conjunction with well-principled learning techniques (*e.g.*, SVM for classification, and SVR for regression) can even infer useful information for age estimation directly from the raw, gray-level pixel values [10, 12]. In a similar fashion, very recent approaches have applied deep-learning methods to the same end [21].

Limitations and open issues. In the majority of the cases, the aforementioned work has focused on improving the performance of age estimation on benchmark datasets, without considering constraints deriving from the application of the proposed methods in real-time applications. Notably, in [13], a sparse kernel-based projection and learning technique have been proposed *also* to speed up age estimation at test time. Besides the issue of computational complexity at test time, the proposed methods outputs decisions which are difficult to interpret. Even for sparse methods, that rely on computing tens of features or prototype matchings, as in [13], this can be a problem. Last but not least, there is a third issue, motivated by the continuously-observed improvement of performances on benchmark datasets, *i.e.*, the need of verifying whether current methods tend to overfit on such specific datasets.

To overcome these three issues, we propose a super-sparse regression method, that can reliably estimate a person's age by matching his or her face image against a very small, budgeted set of reference prototypes, without significantly affecting estimation accuracy. Given its super-sparsity, our method provides more interpretable decisions to end-users and system administrators, and, for the same reason, it should also be less prone to overfit on the specific datasets.

3 Super-sparse Regression for Fast Age Estimation

In this section, we illustrate our approach. We assume that a set $\mathcal{D} = \{\boldsymbol{x}_i, y_i\}_{i=1}^n \in \mathcal{X}^n \times \mathbb{R}_+^n$ of n face images $\boldsymbol{x}_i$ is given, along with the corresponding ages y_i. We do not set any constraint on the representation of faces: we consider them to be objects in an abstract space $\mathcal{X}$, *i.e.*, they can be represented either as feature vectors or as structured objects (*e.g.*, graphs, or bags of visual descriptors). We further assume that a similarity function $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ between two faces is given. For instance, it can be a kernel function (which is positive semi-definite), or any other similarity function, including functions that are not Mercer kernels.

Within this setting, the underlying idea of our approach is to estimate the age of a subject $\boldsymbol{x}$ as a function $f(\boldsymbol{x})$ defined as a *sparse* linear combination of

similarities between $\boldsymbol{x}$ and a *small* set of face prototypes $\boldsymbol{z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m) \in \mathcal{X}^m$, whose number m is *budgeted*, *i.e.*, fixed a priori:

$$f(\boldsymbol{x}) = \sum_{j=1}^{m} \beta_j k(\boldsymbol{x}, \boldsymbol{z}_j) + b \ , \tag{1}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_m) \in \mathbb{R}^m$ is the vector of coefficients and $b \in \mathbb{R}$ the bias, which have to be learnt *together with* the prototypes $\boldsymbol{z}$. This can be formulated as the following optimization problem:

$$\min_{\boldsymbol{\beta}, b, \boldsymbol{z}} \Omega = \sum_{i=1}^{n} u_i \left(f(\boldsymbol{x}_i) - y_i \right)^2 + \lambda \boldsymbol{\beta}^\top \boldsymbol{\beta} \ , \tag{2}$$

where the scalars $u_1, \ldots, u_n$ balance the contribution of each sample $\boldsymbol{x}_k$ to the empirical loss (*e.g.*, if the distribution of samples per age in the training set is not uniform), the quadratic regularizer $\boldsymbol{\beta}^\top \boldsymbol{\beta}$ controls overfitting, and λ is a regularization parameter. Due to the presence of a quadratic regularizer, as in ridge regression, we name our approach *Super-Sparse Ridge* ($\mathrm{S}^2\mathrm{R}$) regression. This approach is clearly linear in the space induced by the similarity function, but not necessarily in the input space $\mathcal{X}$, *i.e.*, the similarity mapping can be used to induce nonlinearity as in kernel methods. Problem 2 turns out to be very similar to that resulting from the novel face verification approach that we recently developed in [1], except for the fact that we are considering an explicit bias term b here, and performing regression on a generic target variable (*i.e.*, not on an SVM's discriminant function to reduce its support vectors). We thus exploit a similar algorithm to compute its solution, as described below.

The objective function of problem (2) can be rewritten in matrix form:

$$\Omega\left(\boldsymbol{\beta}, b, \boldsymbol{z}\right) = \left(\boldsymbol{f}^\top \mathbf{U} \boldsymbol{f} - 2 \boldsymbol{y}^\top \mathbf{U} \boldsymbol{f} + \boldsymbol{y}^\top \mathbf{U} \boldsymbol{y} \right) + \lambda \boldsymbol{\beta}^\top \boldsymbol{\beta} \ , \tag{3}$$

where the column vectors $\boldsymbol{f}, \boldsymbol{y} \in \mathbb{R}^n$ contain the values of f and y for the training samples, $\mathbf{U} \in \mathbb{R}^{n \times n}$ is a diagonal matrix with $\mathrm{diag}(\mathbf{U}) = (u_1, \ldots, u_n)$, and $\boldsymbol{f} = \mathbf{K}_{\boldsymbol{xz}} \boldsymbol{\beta} + b$, being $\mathbf{K}_{\boldsymbol{xz}} \in \mathbb{R}^{n \times m}$ the similarity matrix computed between $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ and the prototypes $\boldsymbol{z}$. The objective function (3) can be iteratively minimized by modifying $\boldsymbol{\beta}$, b and $\boldsymbol{z}$, by first randomly initializing the initial prototypes $\{\boldsymbol{z}_j^{(0)}\}_{j=1}^m$ with m training samples from $\mathcal{D}$, and then alternating the two steps described in the following.

(1) $\boldsymbol{\beta}$-step. We compute the optimal coefficients $\boldsymbol{\beta}$ by keeping the prototypes $\boldsymbol{z}$ fixed. This amounts to a standard ridge regression problem, which can be analytically solved by deriving Eq. (3) with respect to $\boldsymbol{\beta}$ and b (with $\boldsymbol{z}$ constant), and then setting the corresponding gradients to zero:

$$\underbrace{\begin{bmatrix} \mathbf{K}_{\boldsymbol{xz}}^\top \mathbf{U} \mathbf{K}_{\boldsymbol{xz}} + \lambda \mathbb{I} & \mathbf{K}_{\boldsymbol{xz}}^\top \mathbf{U} \mathbf{1} \\ \mathbf{1}^\top \mathbf{U} \mathbf{K}_{\boldsymbol{xz}} & \mathbf{1}^\top \mathbf{U} \mathbf{1} \end{bmatrix}}_{\mathbf{M}} \begin{bmatrix} \boldsymbol{\beta} \\ b \end{bmatrix} = \begin{bmatrix} \mathbf{K}_{\boldsymbol{xz}}^\top \\ \mathbf{1}^\top \end{bmatrix} \mathbf{U} \boldsymbol{y} \ , \tag{4}$$

where $\mathbb{I} \in \mathbb{R}^{m \times m}$ is the identity matrix. The system (4) can be iteratively solved without necessarily inverting $\mathbf{M}$, *e.g.*, using stochastic gradient descent [23].

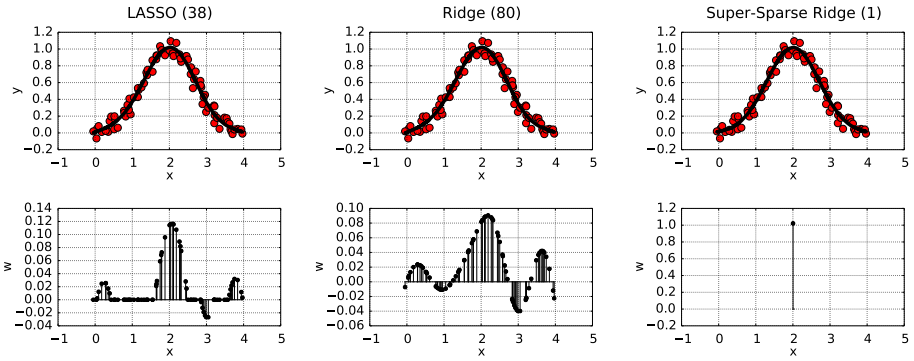

Fig. 1. A simple mono-dimensional, nonlinear regression problem. Plots in the top row show the estimated regression function (solid black line) for LASSO, Ridge, and our Super-Sparse Ridge regression, all trained on the RBF kernel matrix computed between the displayed (red) points, to yield linear functions in the kernel space. The number of selected prototypes is reported in parentheses. The non-zero weight coefficients assigned to the selected prototypes are reported in the plots in the bottom row. As one may reasonably expect, our method only requires one prototype to reliably estimate the given Gaussian-like function, thus significantly reducing complexity at test time.

(2) $\boldsymbol{z}$-step. We update $\boldsymbol{z}$ by iteratively minimizing (3) through gradient descent (no closed-form solution exists). Deriving with respect to a given $\boldsymbol{z}_j$, and using the numerator-layout convention for matrix derivatives, we obtain:

$$\frac{\partial \Omega}{\partial \boldsymbol{z}_j} = 2\,(\boldsymbol{h}-\boldsymbol{y})^\top \mathbf{U}\left(\beta_j \frac{\partial \mathbf{K}_{\boldsymbol{x}\boldsymbol{z}_j}}{\partial \boldsymbol{z}_j} + \mathbf{K}_{\boldsymbol{x}\boldsymbol{z}}\frac{\partial \boldsymbol{\beta}}{\partial \boldsymbol{z}_j} + \mathbf{1}\frac{\partial b}{\partial \boldsymbol{z}_j}\right) + 2\lambda\boldsymbol{\beta}^\top \frac{\partial \boldsymbol{\beta}}{\partial \boldsymbol{z}_j} \quad , \tag{5}$$

where $\mathbf{K}_{\boldsymbol{x}\boldsymbol{z}_j}$ is the j-th column of $\mathbf{K}_{\boldsymbol{x}\boldsymbol{z}}$. Accordingly, all the derivatives with respect to $\boldsymbol{z}_j$ are vectors or matrices with the same number of columns as the dimensionality of $\boldsymbol{z}_j$. In Eq. (5) we need to compute $\frac{\partial \boldsymbol{\beta}}{\partial \boldsymbol{z}_j}$ and $\frac{\partial b}{\partial \boldsymbol{z}_j}$, which can be obtained by deriving Eq. (4). The final gradient is thus given as:

$$\begin{bmatrix} \frac{\partial \boldsymbol{\beta}}{\partial \boldsymbol{z}_j} \\ \frac{\partial b}{\partial \boldsymbol{z}_j} \end{bmatrix} = -\mathbf{M}^{-1}\left(\beta_j \begin{bmatrix} \mathbf{K}_{\boldsymbol{x}\boldsymbol{z}}^\top \\ \mathbf{1}^\top \end{bmatrix} + \begin{bmatrix} \mathbf{S}^\top \\ \mathbf{0}^\top \end{bmatrix}\right)\mathbf{U}\frac{\partial \mathbf{K}_{\boldsymbol{x}\boldsymbol{z}_j}}{\partial \boldsymbol{z}_j} \quad , \tag{6}$$

where $\mathbf{S} \in \mathbb{R}^{n\times m}$ is a matrix consisting of all zeros except for the j^{th} column which is equal to $(\boldsymbol{f}-\boldsymbol{y})$, and $\mathbf{0}, \mathbf{1} \in \mathbb{R}^n$ are column vectors of n zeros and n ones, respectively.

Derivative of $\mathbf{K}_{\boldsymbol{x}\boldsymbol{z}_j}$. In (6), the computation of the derivative of $k(\boldsymbol{x}_1, \boldsymbol{z}_j), \ldots, k(\boldsymbol{x}_n, \boldsymbol{z}_j)$, with respect to the corresponding $\boldsymbol{z}_j$, depends on the given similarity measure k. If k has an analytical representation, as in the case of kernels, the derivative can be easily computed; *e.g.*, for the RBF kernel, $k(\boldsymbol{x}_i, \boldsymbol{z}_j) = \exp(-\gamma||\boldsymbol{x}_i - \boldsymbol{z}_j||^2)$, and thus $\frac{\partial k(\boldsymbol{x}_i, \boldsymbol{z}_j)}{\partial \boldsymbol{z}_j} = -2\gamma\exp(-\gamma||\boldsymbol{x}_i - \boldsymbol{z}_j||^2)(\boldsymbol{x}_i - \boldsymbol{z}_j)$. Otherwise, numerical optimization techniques can be used.

The above optimization procedure is shown as Algorithm 1. We also report a simple graphical example to show how our algorithm works in Fig. 1, in compar-

Algorithm 1 Super-Sparse Ridge (S^2R) Regression (adapted from [1])

Input: the training set $\mathcal{D} = \{\boldsymbol{x}_i, y_i\}_{i=1}^n$; the similarity function $k(\cdot,\cdot)$; the regularization parameter λ; the initial prototypes $\{\boldsymbol{z}_j^{(0)}\}_{j=1}^m$; the step size η; a small number ϵ.
Output: The coefficients $\boldsymbol{\beta}$, b and the virtual prototypes $\{\boldsymbol{z}_j\}_{j=1}^m$.

Set the iteration count $q \leftarrow 0$.
Compute $\boldsymbol{\beta}^{(0)}$ and $b^{(0)}$ for $\boldsymbol{z}_1^{(0)}, \ldots, \boldsymbol{z}_m^{(0)}$ and $\mathcal{D}$ using Eq. (4).
repeat
  Set $j \leftarrow \mathrm{mod}(q, m) + 1$ to index a virtual face prototype.
  Compute $\frac{\partial \Omega}{\partial \boldsymbol{z}_j}$ using Eq. (5).
  Increase the iteration count $q \leftarrow q + 1$
  Set $\boldsymbol{z}_j^{(q)} \leftarrow \boldsymbol{z}_j^{(q-1)} + \eta \frac{\partial \Omega}{\partial \boldsymbol{z}_j^{(q-1)}}$.
  if $\boldsymbol{z}_j^{(q)} \notin \mathcal{X}$, **then** project $\boldsymbol{z}_j^{(q)}$ onto $\mathcal{X}$.
  Set $\boldsymbol{z}_i^{(q)} \leftarrow \boldsymbol{z}_i^{(q-1)}$, $\forall i \neq j$.
  Compute $\boldsymbol{\beta}^{(q)}$ and $b^{(q)}$ for $\boldsymbol{z}_1^{(q)}, \ldots, \boldsymbol{z}_m^{(q)}$ and $\mathcal{D}$ using Eq. (4).
until $\left| \Omega\left(\boldsymbol{\beta}^{(q)}, b^{(q)}, \boldsymbol{z}^{(q)}\right) - \Omega\left(\boldsymbol{\beta}^{(q-1)}, b^{(q-1)}, \boldsymbol{z}^{(q-1)}\right) \right| < \epsilon$
return: $\boldsymbol{\beta} = \boldsymbol{\beta}^{(q)}$, $b = b^{(q)}$ and $\boldsymbol{z} = \boldsymbol{z}^{(q)}$.

ison with popular regression methods like LASSO [20] (which induces sparsity through ℓ_1 regularization) and ridge regression [14].

4 Experiments

In this section we report an experimental analysis to show how the proposed super-sparse regression can: (i) drastically reduce the number of reference prototype, speeding up age estimation at test time; (ii) provide more interpretable decisions; and (iii) avoid the risk of overfitting on cross-database evaluations.

Datasets. For our empirical evaluation, we have used two well-known, public benchmark databases, described below.

FG-Net Aging. This database consists of 1002 images of 82 subjects, including from 6 to 18 images per subject. Images exhibit a variable resolution; some of them are grayscale, many are blurred. For each image, 68 manually-labeled face landmark points are provided to facilitate face normalization and feature extraction. However, we decided not to use them, to simulate a more realistic experimental setting (in which, of course, test images are not manually labeled).

FRGC. This database includes 49,228 frontal images of 568 different subjects (241 female and 327 male subjects) belonging to 7 different ethnicity groups. The minimum and the maximum age values reported in this database are respectively 17 and 69, and there are multiple images per subject. Images were acquired during different time sessions (one per month). On average, 6 images per user were acquired during a single session. To keep the complexity of our experiments manageable, we randomly select a subset of about 5,000 images.

For the sake of completeness, we report the distributions of ages for the two considered datasets in Fig. 2.

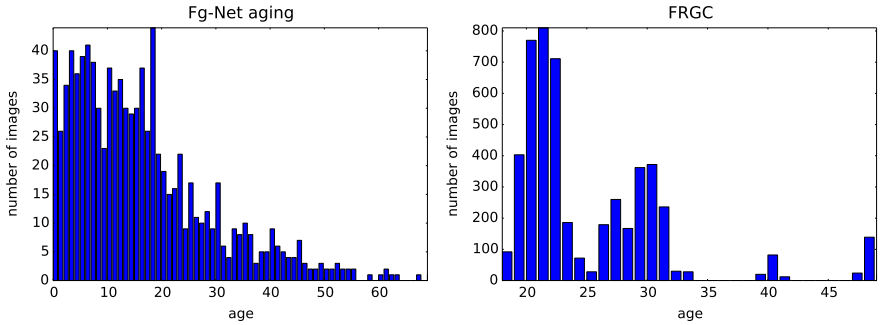

Fig. 2. Number of images per age for Fg-Net (*left*) and FRGC (*right*).

Experimental Setup. We normalize face images by first detecting the eye positions with a (trained) detector originally proposed in [24]; then, we align faces to have eyes into a fixed position, normalize illumination, and use an ellipse-based normalization to eliminate irrelevant background information, as described in [2]. Similarly to previous work in age estimation, our results are averaged using a 5-fold cross-validation procedure (using different subjects in each fold), and reported in terms of the Mean Absolute Error (MAE), given as:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|f(\boldsymbol{x}_i) - y_i\right| \tag{7}$$

where $f(\boldsymbol{x}_i)$ is the regression estimate of the true subject's age y_i for the test image $\boldsymbol{x}_i$, and n the test set size (*i.e.*, the size of the test fold).

We adopt three different feature representations for face images: (i) the grey-level values of pixels (pixel values), (ii) principal component analysis (PCA-based mapping), and (iii) linear discriminant analysis (LDA-based mapping). To obtain (ii) and (iii), we start from the pixel representation of images and apply the two aforementioned projection techniques, retaining the first 100 components for the PCA-based mapping (approximately capturing 90% of the variance in the data), and $K-1$ components for the LDA-based mapping, being K the number of distinct ages in the training data.

We use these three representations as features, and, after linearly rescaling them in $[-1, 1]$, we apply LASSO, SVR (with linear kernel), and Ridge regression. We then consider the RBF kernel (with $\gamma = 1/d$, being d the number of dimensions in the input space) and use the same methods trained using the kernel values as input features. The regularization parameter for LASSO, SVR and Ridge regression were set through an inner 2-fold cross-validation, to optimize the expected MAE value. We consider the proposed approach for super-sparse ridge (S^2R) regression with $m = 5$ and $m = 15$ virtual reference prototypes, for which we set the gradient step size η at each iteration by selecting the best value $\eta \in \{0.3, 0.1, 0.01, 0.001\}$. In general, the number m of prototypes could also be tuned through a cross-validation procedure. We finally evaluate all the aforementioned approaches on a cross-database scenario, where we replace the test fold of the dataset used for training with the full dataset used for testing.

Regressor	**Fg-Net/Fg-Net**	**Fg-Net/FRGC**	**FRGC/Fg-Net**	**FRGC/FRGC**
Pixel values				
LASSO	9.16 ± 1.41	13.63 ± 0.23	14.05 ± 0.4	6.51 ± 0.6
SVR	9.1 ± 1.39	12.45 ± 0.38	13.6 ± 0.41	6.02 ± 0.79
Ridge	9.21 ± 1.39	13.63 ± 0.23	14.05 ± 0.4	6.53 ± 0.62
LASSO RBF	**6.92 ± 1.62**	8.76 ± 0.65	12.87 ± 0.33	4.14 ± 1.09
SVR RBF	6.98 ± 1.79	9.01 ± 0.49	**12.58 ± 0.29**	**3.87 ± 1.33**
Ridge RBF	6.94 ± 1.62	8.74 ± 0.66	12.87 ± 0.32	4.13 ± 1.09
S^2R RBF (5)	10.09 ± 1.07	6.93 ± 0.8	15.96 ± 1.11	5.48 ± 1.22
S^2R RBF (15)	8.92 ± 1.58	**6.89 ± 0.21**	13.01 ± 0.16	5.44 ± 2.30
PCA-based mapping				
LASSO	7.83 ± 1.68	10.93 ± 0.77	13.46 ± 0.61	4.9 ± 0.68
SVR	7.21 ± 1.96	8.98 ± 0.63	12.76 ± 0.29	**3.76 ± 1.28**
Ridge	9.06 ± 1.07	11.09 ± 1.27	13.41 ± 0.81	5.25 ± 0.46
LASSO RBF	8.22 ± 1.65	8.69 ± 1.5	12.75 ± 0.32	4.33 ± 1.13
SVR RBF	**7.05 ± 1.96**	8.95 ± 0.72	**12.48 ± 0.31**	3.93 ± 1.54
Ridge RBF	8.23 ± 1.65	8.67 ± 1.51	12.76 ± 0.31	4.32 ± 1.13
S^2R RBF (5)	7.92 ± 1.3	9.09 ± 0.84	14.02 ± 0.39	5.37 ± 0.89
S^2R RBF (15)	7.86 ± 1.3	**7.76 ± 0.40**	12.87 ± 0.34	5.90 ± 2.49
LDA-based mapping				
LASSO	8.72 ± 1.6	11.98 ± 0.91	13.78 ± 0.47	5.79 ± 0.73
SVR	8.63 ± 1.38	11.48 ± 0.37	13.77 ± 0.47	5.79 ± 0.73
Ridge	8.71 ± 1.6	11.98 ± 0.91	13.78 ± 0.47	5.78 ± 0.73
LASSO RBF	**7.92±1.87**	10.92 ± 0.74	13.55±0.32	5.93 ± 0.9
SVR RBF	8.06 ± 1.78	10.57 ± 0.7	13.7 ± 0.38	**5.08 ± 0.92**
Ridge RBF	7.92 ± 1.87	10.92 ± 0.75	13.55 ± 0.32	5.93 ± 0.9
S^2R RBF (5)	8.49 ± 1.62	11.56 ± 0.35	16.84 ± 2.06	9.56 ± 2.19
S^2R RBF (15)	8.37 ± 1.62	**9.93 ± 0.40**	**13.28 ± 0.5**	7.30 ± 1.89

Table 1. Average MAE and standard deviation for the given configurations. Each column reports training/test sets, including cross-database evaluations (*e.g.*, Fg-Net/FRGC means that we train the regression function on Fg-Net, and test on FRGC). Best results are highlighted in bold.

Results. Results are shown in Table 1, for the given configurations (pixel values, PCA- and LDA-based mappings), linear methods (LASSO, SVR, and Ridge), RBF-based methods (LASSO RBF, SVR RBF, Ridge RBF), and S^2R regression with 5 and 15 prototypes. First, note that linear methods exhibit lower performances, confirming that the problem of age estimation is better tackled by nonlinear approaches. It is then clear that S^2R regression can achieve comparable performances with the other nonlinear regressors, although using only a very small number of prototypes (*cf.* Table 2). SVR RBF and LASSO RBF outperform occasionally the other techniques. As expected, due to its super-sparsity, S^2R regression also exhibits good performance on the cross-database evaluations (*cf.* second and third column in Table 1). Sometimes, cross-database learning may be difficult due to the presence of unbalanced distributions of ages between the two considered datasets. Consequently, one method may tend to

Regressor	Fg-Net	FRGC
Pixel values		
LASSO RBF	758.4	3977.8
SVR RBF	759.0	3974.0
Ridge RBF	760.8	4000.0
PCA-based mapping		
LASSO RBF	758.8	3928.2
SVR RBF	759.4	3975.4
Ridge RBF	760.8	4000.0
LDA-based mapping		
LASSO RBF	758.2	3975.8
SVR RBF	753.4	1584.8
Ridge RBF	760.8	4000.0

Table 2. Average number of prototypes (*i.e.*, required matchings for testing) and standard deviation for nonlinear regression methods, for the different considered configurations. Note how the number of matchings is always significantly higher than 5 or 15, *i.e.*, the number of prototypes used by our S^2R regression.

overfit and predict the age value with the highest prior probability in the training set. Our method naturally allows to compensate for this problem by using non-uniform values in $\mathbf{U}$ (see Eq. 2), *e.g.*, by assigning a value that is inversely-proportional to the prior probability of observing the corresponding age value in the training set. We leave however a more detailed cross-database analysis to future work, also from a more theoretically-sound perspective.

Interpretability. As mentioned before, interpretability of decisions is another important property to understand whether the regression algorithm has properly learned some aging pattern, and, thus, if it may correctly predict images from different datasets. In Fig. 3, we report two examples of the prototypes learned by S^2R regression on Fg-Net and FRGC, respectively. As one may appreciate, our method assigns correctly lower age values to "smaller" faces (*i.e.*, faces of children), and higher values to images which exhibit textures characterized by higher frequencies, *i.e.*, by the presence of wrinkles.

5 Conclusions and Future Work

We have proposed a novel super-sparse regression method, inspired by our recently-proposed SVM reduction technique in the context of face verification [1]. That method was in turn inspired by existing reduction methods [18], and capable of outperforming them, for the following reasons: (i) it is not greedy (*i.e.*, it iteratively modifies each prototype during the minimization process), and (ii) it can also be trained with similarity functions that are not necessarily positive semi-definite (*i.e.*, Mercer) kernels. Our results have shown that the proposed method can achieve comparable performances to other popular regression techniques, either sparse (*e.g.*, SVR, LASSO) or not (*e.g.*, Ridge), while reducing the number of prototypes required for age estimation at test time of orders of

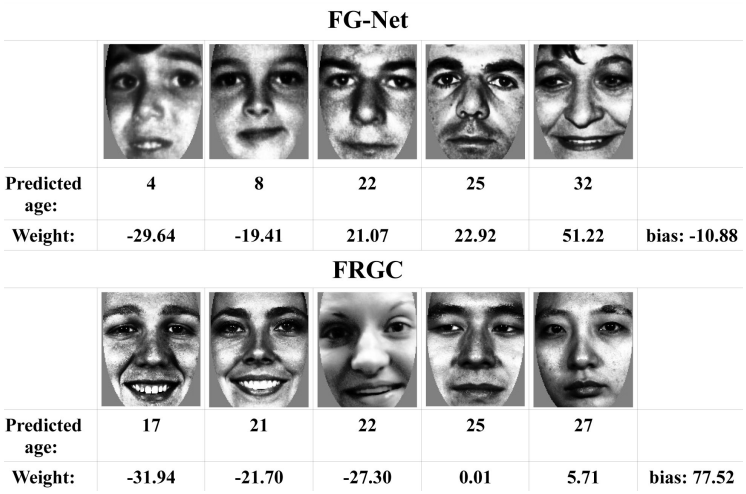

Fig. 3. Virtual reference prototypes learned by our S^2R regression method for FG-Net and FRGC, along with the corresponding weights $\boldsymbol{\beta}$, bias b, and predicted age $f(\boldsymbol{x})$.

magnitude. Furthermore, thanks to its super-sparsity property, the proposed approach provides decisions which are easier to interpret for system administrators and end-users of the age estimation system, and it has also exhibited promising results on cross-database evaluations. Future research directions therefore include finding a theoretically-sound explanation to this behavior, to gain more interesting insights on this property. Another interesting line of work consists of evaluating the proposed method on more complex feature representations and visual descriptors, in order to empirically validate whether age estimation accuracy can be improved on benchmark datasets while keeping a super-sparse solution, *i.e.*, while dramatically reducing complexity at test time.

Acknowledgments. This work has been supported by the project "Computational quantum structures at the service of pattern recognition: modeling uncertainty" (CRP-59872) funded by Regione Autonoma della Sardegna (RAS), L.R. 7/2007, Bando 2012, and by the project "Security of pattern recognition systems in future internet" (CRP-18293) funded by RAS, L.R. 7/2007, Bando 2009.

References

1. Biggio, B., Melis, M., Fumera, G., Roli, F.: Sparse support faces. In: Int'l Conf. on Biometrics (ICB). pp. 1–6 (2015)
2. Bolme, D., Ross Beveridge, J., Teixeira, M., Draper, B.: The CSU face identification evaluation system: Its purpose, features, and structure. In: Crowley, J., Piater, J., Vincze, M., Paletta, L. (eds.) Computer Vision Systems, LNCS, vol. 2626, pp. 304–313. Springer Berlin Heidelberg (2003)
3. Chang, K.Y., Chen, C.S.: A learning framework for age rank estimation based on face images with scattering transform. IEEE Trans. Image Processing 24(3), 785–798 (2015)

4. Chao, W.L., Liu, J.Z., Ding, J.J.: Facial age estimation based on label-sensitive learning and age-oriented regression. Patt. Rec. 46(3), 628–641 (2013)
5. Chapelle, O.: Training a support vector machine in the primal. Neural Comput. 19(5), 1155–1178 (2007)
6. Chen, Y.L., Hsu, C.T.: Subspace learning for facial age estimation via pairwise age ranking. IEEE Trans. Inf. Forensics and Security 8(12), 2164–2176 (2013)
7. Choi, S.E., Lee, Y.J., Lee, S.J., Park, K.R., Kim, J.: Age estimation using a hierarchical classifier based on global and local facial features. Patt. Rec. 44(6), 1262–1281 (2011)
8. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. In: IEEE Trans. Patt. Anal. Mach. Intell. 23(6), 681–685 (2001)
9. Fu, Y., Guo, G., Huang, T.: Age synthesis and estimation via faces: A survey. IEEE Trans. Patt. Anal. Mach. Intell. 32(11), 1955–1976 (2010)
10. Fu, Y., Huang, T.: Human age estimation with regression on discriminative aging manifold. IEEE Trans. Multimedia 10(4), 578–584 (2008)
11. Guo, G., Fu, Y., Dyer, C., Huang, T.: Image-based human age estimation by manifold learning and locally adjusted robust regression. IEEE Trans. Image Processing 17(7), 1178–1188 (2008)
12. Guo, G., Fu, Y., Dyer, C., Huang, T.: A probabilistic fusion approach to human age prediction. In: Computer Vision and Pattern Recognition Workshops, 2008. CVPRW '08. IEEE CS. pp. 1–6 (2008)
13. Guo, G., Mu, G.: Simultaneous dimensionality reduction and human age estimation via kernel partial least squares regression. In: IEEE Conf. Computer Vision and Pattern Recognition (CVPR). pp. 657–664 (2011)
14. Hoerl, A.E., Kennard, R.W.: Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12(1), 55–67 (1970)
15. Kwon, Y.H., Lobo, N.D.V.: Age classification from facial images. In: IEEE Conf. Computer Vision and Pattern Recognition (CVPR). pp. 762–767 (1999)
16. Li, C., Liu, Q., Dong, W., Zhu, X., Liu, J., Lu, H.: Human age estimation based on locality and ordinal information. IEEE Trans. Cybernetics PP(99), 1–1 (2014)
17. Li, Z., Park, U., Jain, A.: A discriminative model for age invariant face recognition. IEEE Trans. Inf. Forensics and Security 6(3), 1028–1037 (2011)
18. Schölkopf, B., Mika, S., Burges, C.J.C., Knirsch, P., Muller, K.R., Rätsch, G., Smola, A.J.: Input space versus feature space in kernel-based methods. IEEE Trans. Neural Networks 10(5), 1000–1017 (1999)
19. Steinwart, I.: Sparseness of support vector machines. J. Mach. Learn. Res. 4(11), 1071–1105 (2003)
20. Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society (Series B) 58, 267–288 (1996)
21. Wang, X., Guo, R., Kambhamettu, C.: Deeply-learned feature for age estimation. In: IEEE Winter Conf. Applications Computer Vision (WACV). pp. 534–541 (2015)
22. Ylioinas, J., Hadid, A., Hong, X., Pietikäinen, M.: Age estimation using local binary pattern kernel density estimate. In: Petrosino, A. (ed.) Int'l Conf. Image Analysis and Processing (ICIAP), LNCS, vol. 8156, pp. 141–150. Springer Berlin Heidelberg (2013)
23. Zhang, T.: Solving large scale linear prediction problems using stochastic gradient descent algorithms. In: 21st Int'l Conf. Mach. Learning (ICML). Omnipress. pp. 919–926 (2004)
24. Zhu, X., Ramanan, D.: Face detection, pose estimation, and landmark localization in the wild. In: IEEE Conf. Computer Vision and Pattern Recognition (CVPR). pp. 2879–2886. IEEE (2012)