Abundant Inverse Regression using Sufficient Reduction and its Applications

THE UNIVERSITY of WISCONSIN MADISON

Hyunwoo J. Kim*, Brandon M. Smith*, Nagesh Adluru, Charles R. Dyer, Sterling C. Johnson, Vikas Singh

http://cs.wisc.edu/~hwkim/projects/air/

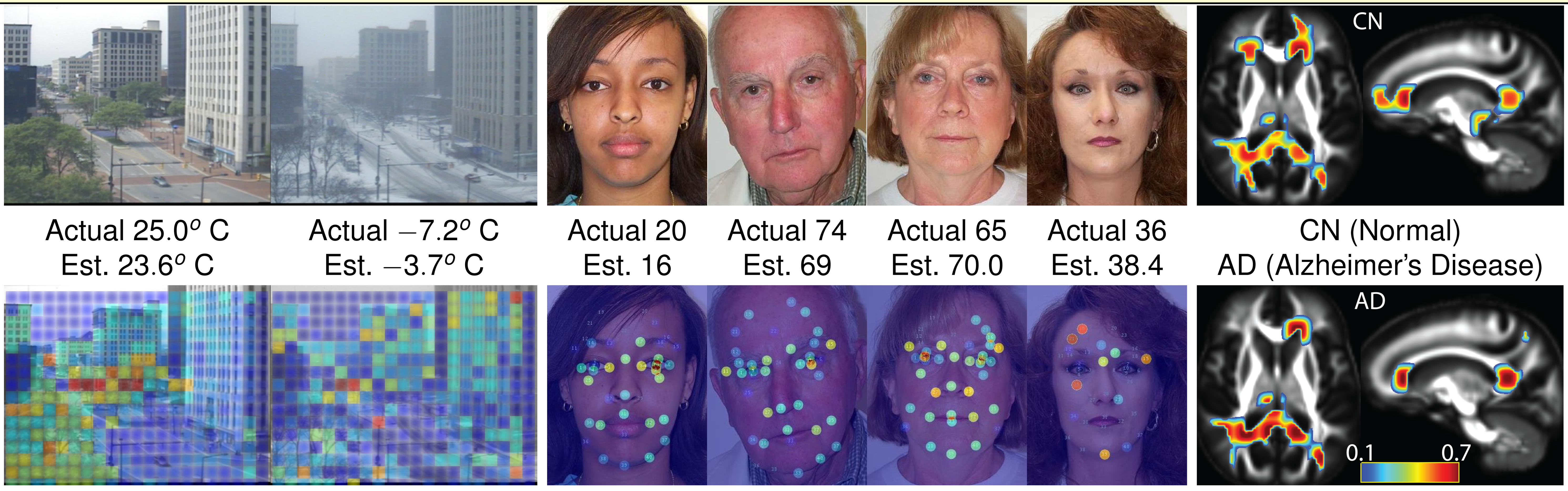

Figure: Dynamic feature weights for three tasks: ambient temperature prediction (left), age estimation (middle) on Lifespan database (Meredith et al., 2004, Guo, et al., 2012.) , and AD classification (right). AD classification accuracy of 86.17% by simple thresholding of continuous prediction by AIR comparing to SVM+PCA (80%-85%) (Hwang et al., 2015). AIR provides a way to determine, at test time, which features are most important to the prediction. Our results are competitive, which demonstrates that we achieve this capability without sacrificing accuracy.

Objective

Goal: Develop a regression model explaining why a particular prediction was made at the level of specific examples/samples
Strategy: Inverse Regression and Sufficient Reduction in the "abundant" feature setting.

Main Idea

Desired: Relevance of individual covariates at the level of specific samples for a given regression task.

Challenge ("Chicken-or-egg problem"): Relevance/confidence score to individual covariates x^i should condition the estimate based on knowledge of all other (uncorrupted) covariates x^{-j}.
$f(x^1|x^2, x^3, \dots, y)$ is high-dimensional requiring large amount of data.
$f(x^1|x^2, y), f(x^1|x^3, y), f(x^1|x^4, y), \dots f(x^1|x^p, y)$ too many cases.
Solution: sufficient reduction

$$f(x^i|\phi(X)), \text{ where } x^i|X, \phi(X) \sim x^i|\phi(X)$$

Desired: Robust regression model which allows missing or randomly corrupted covariates with their dynamic weights.
Solution: distance measure with dynamic weights.

Sufficient Reduction and Inverse Regression

- Given a regression model $h : X \to Y$, a reduction $\phi : \mathbf{R}^p \to \mathbf{R}^q, q \leq p$, is a **sufficient reduction** if it satisfies one of the following conditions:
  1) inverse reduction, $X|(Y, \phi(X)) \sim X|\phi(X)$,
  2) forward reduction, $Y|X \sim Y|\phi(X)$,
  3) joint reduction, $X \perp\!\!\!\perp Y|\phi(X)$,
  where $\perp\!\!\!\perp$ indicates independence, $\sim$ means identically distributed, and $A|B$ refers to the random vector A given the vector B.
- Forward regression $\mathbb{E}[Y|X]$: $f : X \to Y$
- Inverse regression $\mathbb{E}[X|Y]$: $f : Y \to X$
- Sliced Inverse Regression (Li, 1991) estimates $\phi(X)$ by PCA over $\mathbb{E}[X|Y]$.
- Relevance (dynamic weights) in our model:

$$\mathbb{E}_j[f(x^i|\phi^j(X))], \text{ where } x^i|X, \phi^j(X) \sim x^i|\phi^j(X) \quad (1)$$

Distance Measure with Relevance

$$\mathrm{d}_w(x_1, x_2, w_1, w_2) := \sqrt{\frac{\sum_j w_{x_1^j} w_{x_2^j} \left[\mathrm{d}(x_1^j, x_2^j)^2 - 2\sigma^2_{x^j|z^j}\right]}{\sum_j w_{x_1^j} w_{x_2^j}}}. \quad (2)$$

Relevance of covariates $w_{x^j} := \sum_l w_{\phi^l} f(x^j|\hat{y}^l) / \sum_l w_{\phi^l}$
Global weight, $w_{\phi^l} := \mathbb{E}[(y - \phi^l(x^l))^2]^{-1}$, $\sigma^2_{x^j|z^j} := \mathbb{E}[(x^j - \mathbb{E}[x^j|x^{-j}])^2]$

Algorithm

1: **Training**
2: Estimate a joint distribution for each covariate, $f(x^j, y)$
3: Find sufficient reduction $\phi^l : x^l \to y$ for each subset of features x^l
4: Estimate the prior/weight for $\phi_l(\cdot)$ as $w_{\phi^l} = \mathbb{E}[(y - \phi^l(x^l))^2]^{-1}$
5: Estimate cond. confidence of feature $w_{x^j} := \sum_l w_{\phi^l} f(x^j|\hat{y}^l) / \sum_l w_{\phi^l}$
6: Fit a feature confidence aware regressor $h : [\{x^j\}_{j=1}^K, \{w_{x^j}\}_{j=1}^K] \to y$
7: **Prediction**
8: Evaluate $w_{x^j} := \mathbb{E}f\left(x^j|\phi^l(x^l)\right)$ by lines 3 and 5, with learned w_{ϕ^l}.
9: $\hat{y} = h(\{x^j\}_{j=1}^K, \{w_{x^j}\}_{j=1}^K)$

Lemma 1: Optimal Global Weights for ϕ^l

Suppose we have K random variables (sufficient reduction),

$$\phi^1(x_1) \sim \mathcal{N}(y, \sigma_1^2), \dots, \phi^K(x_K) \sim \mathcal{N}(y, \sigma_K^2), \quad (3)$$

where $\sigma_l^2 > 0, \forall l \in \{1, \dots, K\}$. Consider a convex combination of ϕ^l. Its expectation is y. Assuming that the errors of all sufficient reductions are independent, the problem to find the optimal weights for the convex combination with the minimal variance can be formulated as

$$\min_w \sum_{l=1}^K \sigma_l^2 (w_l)^2 \text{ s.t.} |w|_1 = 1 \text{ and } w_l \geq 0, \text{ for all } l \in 1, \dots, K. \quad (4)$$

The unique global optimum of Eq. (4) is $w_l = \sigma_l^{-2} / \sum_{k=1}^K \sigma_k^{-2}$.

Experiments: Ambient Temperature Prediction

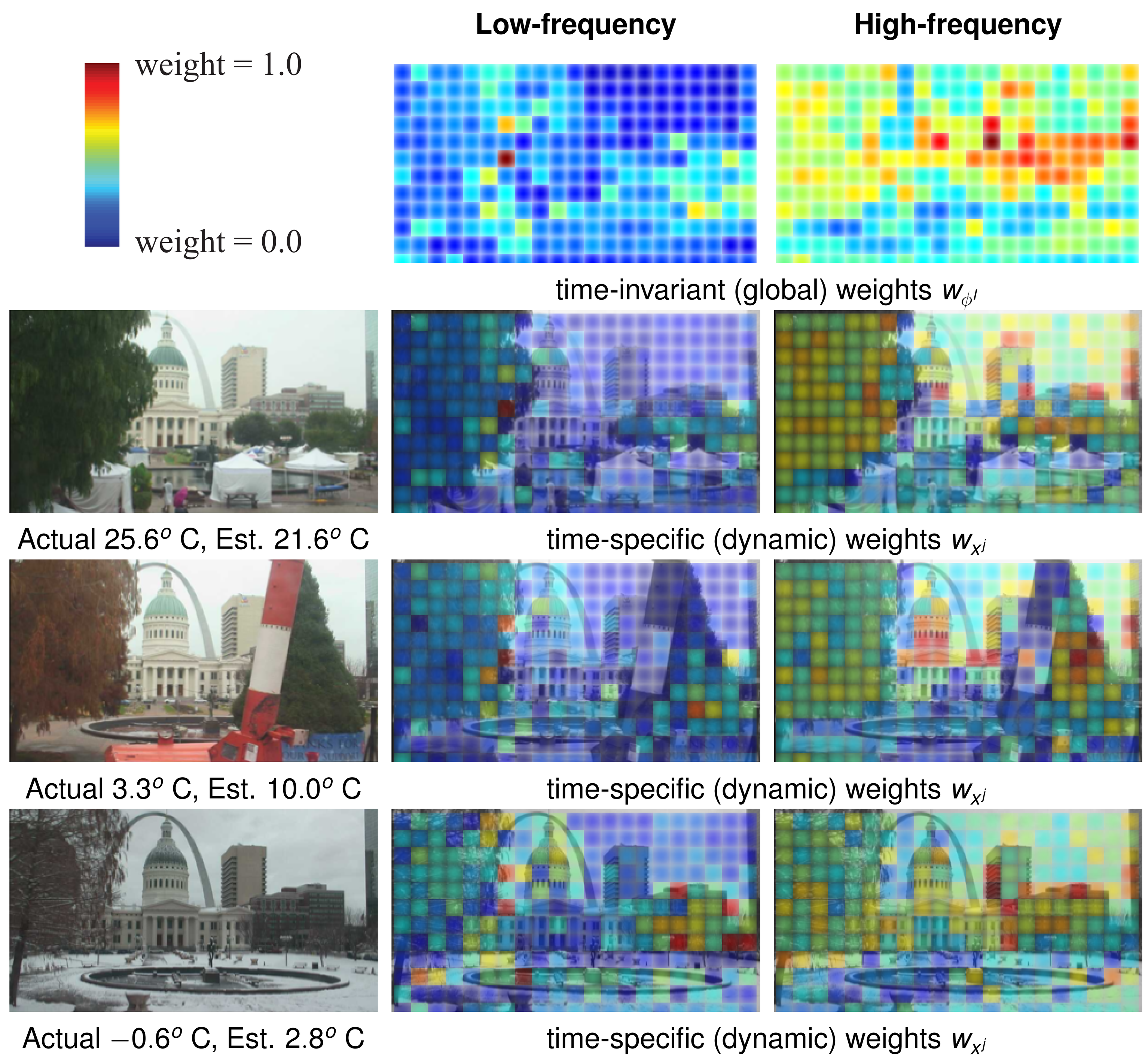

Figure: Qualitative results on scene (a) from the *Hot or Not* dataset (Glasner et al., 2015) .

ECCV 2016

Research supported by NIH AG040396, NSF CAREER award 1252725, UW ADRC AG033514
UW ICTR 1UL1RR025011, UW CPCP AI117924 and Waisman Core Grant P30 HD003352-45