Techniques for Learning Multiple Related Tasks

Massimiliano Pontil
Department of Computer Science
University College London

PASCAL Pump Priming Project:
Multitask Learning: Optimization Methods and Applications
joint with T. Evgeniou (INSEAD), M. Herbster (UCL),
G. Bakir (MPI-BK), J.P. Vert (Ecole des Mines)

thanks to: Andreas Argyriou and Charles Micchelli

Project Goals

- Optimization methods for multi-task learning
- Theoretical investigations (convergence, error analysis, approximation)
- Implementation and demonstration
- Develop at least one real application

  (conjoint analysis, bioinformatics, robot learning)
- Lecture notes for a course on convex optimization

Achieved Results

- Method for learning shared features across tasks
- Conjoint analysis application
- Matlab implementation
- Analysis of the method (convergence, non-linear extensions, approximation)

Learning Multiple Tasks Simultaneously

- By a task we mean a real-valued function (for regression / classification)
- Learning multiple related tasks vs. learning independently
- Few data per task; pooling data across related tasks

  Example 1: predict users' preferences to products

  Example 2: object detection in computer vision

Approach

- Learn each task by L_2-norm regularization

$$\operatorname*{Min}_{w \in \mathbb{R}^d} \sum_{i=1}^{m} (w^\top x_i - y_i)^2 + \gamma w^\top D^{-1} w, \qquad \gamma > 0$$

- Further minimize over 'structure matrix' D:

$$\operatorname*{Min}_{D \in \mathcal{D}} \sum_{t=1}^{T} \left(\operatorname*{Min}_{w \in \mathbb{R}^d} \sum_{i=1}^{m} (w^\top x_{ti} - y_{ti})^2 + \gamma w^\top D^{-1} w \right)$$

- $\mathcal{D}$: subset of positive definite matrices with bounded trace

Alternate Minimization Algorithm

- Alternating minimization over W (supervised learning) and D (unsupervised "correlation" of tasks).

Initialization: set $D = \frac{1}{d} I_{d \times d}$
while convergence condition is not true **do**
  for $t = 1, \ldots, T$

$$\text{set } w_t = \arg\min_{w \in \mathbb{R}^d} \sum_{i=1}^{m} (w^\top x_{ti} - y_{ti})^2 + \gamma\, w^\top D^{-1} w$$

  end for

$$\text{set } D = \frac{(WW^\top)^{\frac{1}{2}}}{\operatorname{trace}(WW^\top)^{\frac{1}{2}}}, \quad \text{where } W = [w_1, \ldots, w_T]$$

end while

Conjoint Analysis Experiment

- Consumers' ratings of products [Lenk et al. 1996]

- 180 persons (tasks)

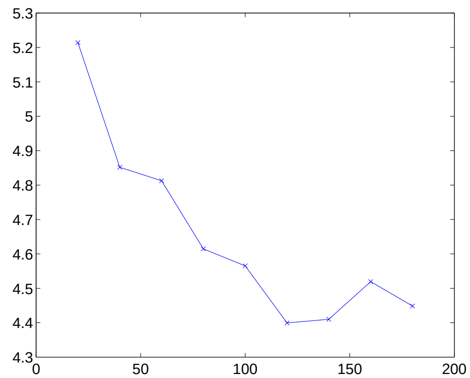
- 8 PC models (training examples); 4 PC models (test examples)

- 13 binary input variables (RAM, CPU, price etc.) $+$ bias term

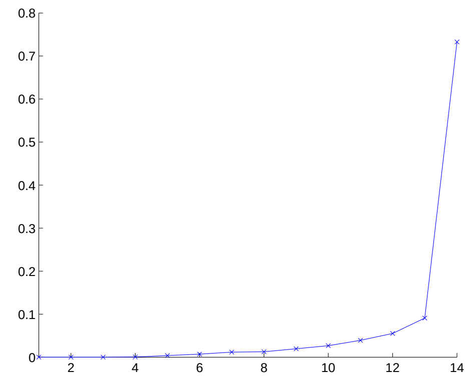
- Integer output in $\{0, \ldots, 10\}$ (likelihood of purchase)

Conjoint Analysis Experiment

Test error



Eig(D)



\#tasks

- Performance improves with more tasks (for independent tasks, error $= 16.53$)
- A single most important feature shared by all persons

Interpretation 1: Spectral Regularization

$$\operatorname*{Min}_{D\in\mathcal{D}} \sum_{t=1}^{T} \left(\operatorname*{Min}_{w\in\mathbb{R}^d} \sum_{i=1}^{m} (w^\top x_{ti} - y_{ti})^2 + \gamma w^\top D^{-1} w \right)$$

Rewrite above problem as a matrix regularization one:

$$\operatorname*{Minimize}_{W\in\mathbb{R}^{d\times T},\ D\in\mathcal{D}} \qquad \mathrm{Error}(W) + \gamma\,\mathrm{trace}(W^\top D^{-1} W)$$

where $W = \begin{pmatrix} | & & | \\ w_1 & \dots & w_T \\ | & & | \end{pmatrix}$, $\mathrm{Error}(W) = \sum_{t=1}^{T}\sum_{i=1}^{m}(w_t^\top x_{ti} - y_{ti})^2$

Spectral Regularization

Lemma: if $\mathcal{D} = \{D \succ 0, \operatorname{trace} D \leq 1\}$ then

$$\inf_{D \in \mathcal{D}} \operatorname{trace}(W^\top D^{-1} W) = \|W\|_1^2$$

with $\|W\|_1$ the L_1 norm of the singular values of W

- Extension: if F is a *spectral function*, then $\inf_{D \in \mathcal{D}} \operatorname{trace}(W^\top F(D) W)$ is a spectral function of the covariance matrix $WW^\top$

- Infimizer may be analytically computed

Interpretation 2: Learning Common Features

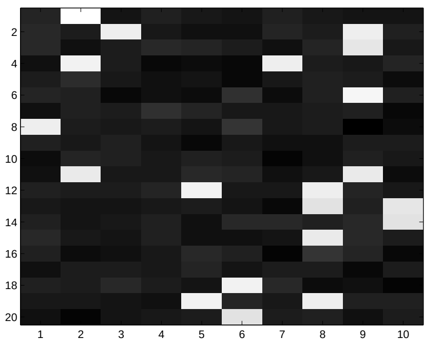
Writing $D = U\Lambda U^\top$, with U orthogonal and $A = U^\top W$ and minimizing over Λ, our original approach reduces to

$$\underset{A,\,U^\top U=I}{\text{Minimize}} \quad \text{Error}(UA) + \gamma \|A\|_{2,1}^2$$

- Interpretation: learn a small set of common features shared by the tasks
- if $U = I$ (fixed), method selects important variables shared by tasks

Effect of $(2,1)$-Norm

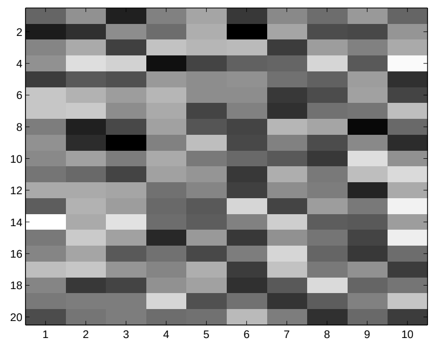
- Compare matrices favoured by different norms:



$\|A\|_{1,1}$

(Vector 1-norm)

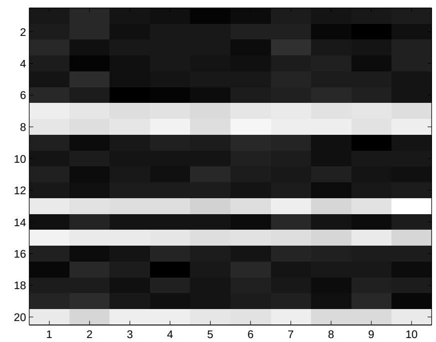
Sparsity



$\|A\|_{2,2}$

(Frobenius norm)

Uniformity



$\|A\|_{2,1}$

(Mixed norm)

Structured sparsity

Equivalent problem

In summary, the following problems are equivalent (here $p \in (0,2]$)

$$\underset{W, D \in \mathcal{D}}{\text{Minimize}} \quad \text{Error}(W) + \gamma \, \text{trace}(W D^{1-\frac{2}{p}} W) \tag{1}$$

$$\underset{W}{\text{Minimize}} \quad \text{Error}\,(W) + \gamma \, \|W\|_p^2 \tag{2}$$

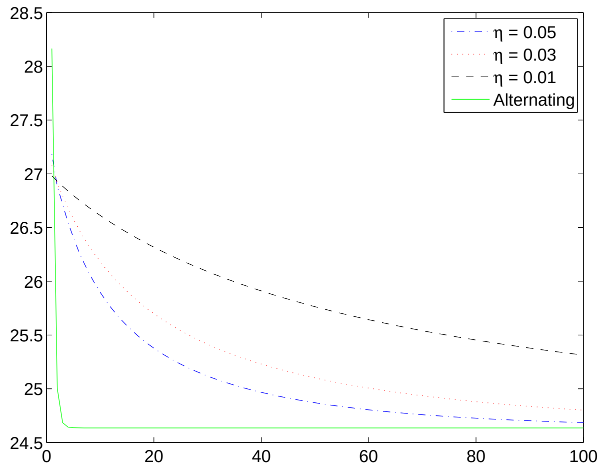
$$\underset{A,\, U^\top U = I}{\text{Minimize}} \quad \text{Error}(UA) + \gamma \, \|A\|_{2,p}^2 \tag{3}$$

- (1) is our original proposal and is jointly convex
- (2) is also convex but may be more difficult to solve (next slide)
- (3) helps us gain intuition on our proposal but is non convex

Computational Cost

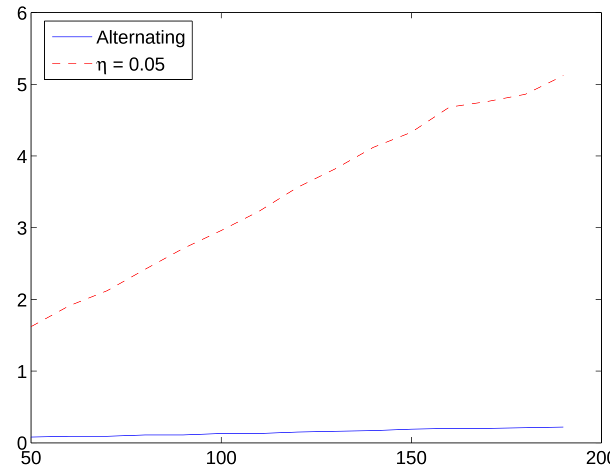
Reg.

error



#iterations

secs.

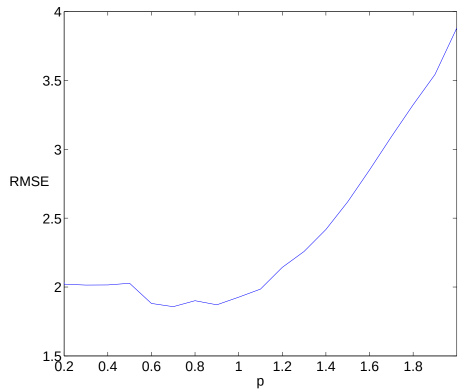


#tasks

- Compare computational cost of alternating minimization vs. gradient descent (on problem (2)), for $p = 1.5$
- Curves for different learning rates are shown

Computer Survey Experiment

RMSE



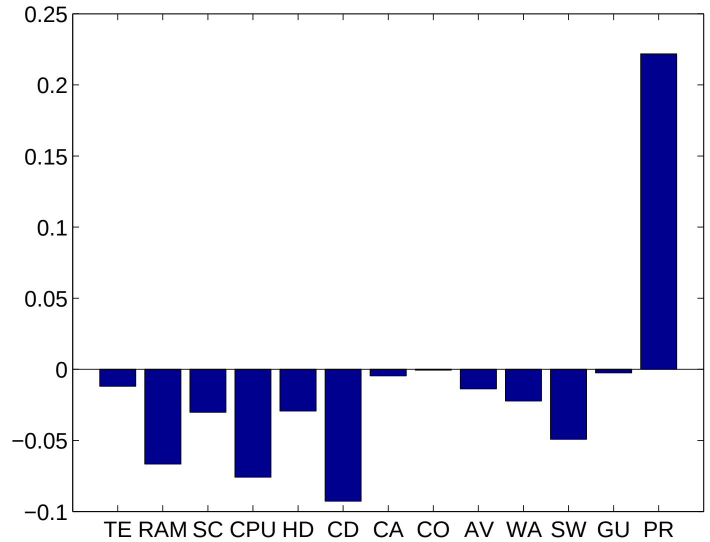
p

Method	RMSE
p = 2	3.88
p = 1	1.93
p = 0.7	1.86

- Performance using L_p spectral regularizers
- Trace norm ($p = 1$) is best among the *norms*
- A non-convex regularizer ($p < 1$) does even better

Computer Survey Experiment

u_1



- The eigenvectors of D are the features U solving problem (3)
- The most important feature weighs *technical characteristics* vs. *price*

Additional Results

- Algorithm (with some perturbation) *converges* to the optimal solution [AEP]
- Conditions for joint convexity [AMPY]
- Nonlinear extension via kernels [AEP]
- Can be used for transfer learning
- Extension to tasks with attributes [ABEV]

Additional Results (cont.)

- Improves over hierarchical Bayes (which also learns a matrix D using Bayesian inference but with more elaborate priors) [EPT-08]

- More general regularizers can be considered, e.g.

$$\sum_{t=1}^{T}(w_t - \bar{w})^\top D^{-1}(w_t - \bar{w})$$

- Universal multi-task kernels [CMPY-08]

Future work

Begun new projects on this topics (ongoing EPSRC grant)

- Different tasks' domains
- Generalization error bounds
- Consider temporal data
- Online learning

Published papers

[AEP] A. Argyriou, T. Evgeniou and M. Pontil. Convex multi-task feature learning. *Machine Learning*, to appear.

[AMPY] A. Argyriou, C.A. Micchelli, M. Pontil and Y. Ying. Spectral regularization for multi-task structure learning, *NIPS 2008*.

[ABEV] J. Abernethy, F. Bach, T. Evgeniou and J.-P. Vert. Low-rank matrix factorization with attributes. Technical report N24/06/MM, Ecole des Mines de Paris,2006.

[CMPY] A. Caponnetto, C.A. Micchelli, M. Pontil, Y. Ying. Universal multi-task kernels. Preprint, 2007.

[EPT] T. Evgeniou, M. Pontil and O. Toubia. A convex optimization approach to modeling heterogeneity in conjoint estimation. *Marketing Science*, to appear.