Various Formulations for Learning the Kernel and Structured Sparsity

Massimiliano Pontil
University College London

(in collaboration with Andreas Argyriou, Luca Baldassarre, Charles Micchelli, Jean Morales and Yiming Ying)

NIPS Workshop on "New Directions in Multiple Kernel Learning"
11/12/2010

Plan of the talk

- Problem formulation and relation to the Group Lasso
- Convex combinations of continuously parameterized kernels
- Learning the kernel in multi-task learning
- Constrained multiple kernel learning and structured sparsity

Learning the kernel

Let $K : X \times X \to \mathbb{R}$ be a positive semidefinite (psd) kernel and H_K the corresponding Hilbert space (RKHS) with norm $\| \cdot \|_K$
Learning method:

$$E(K) = \min_{f \in H_K} \sum_{i=1}^{m} L(y_i, f(x_i)) + \gamma \|f\|_K^2 \qquad (*)$$

How do we choose K?
Prescribed a set $\mathcal{K}$ of psd kernels and consider the problem

$$\min_{K \in \mathcal{K}} E(K) \qquad (**)$$

Related work: *[Argyriou et al. 05, Bach et al. 04; Chapelle et al. 02, Cortes et al. 09, Lanckriet et al. 04, Micchelli & P. 05, Rakotomamonjy et al. 08, Sonnenburg et al. 06, Wu et al. 07, Zien & Ong, 07,...]*

Convexity

By the Representer Theorem, solution of $(*)$ has the form $\sum_{i=1}^{m} c_i K(x_i, \cdot)$, yielding the finite dimensional optimization problem

$$E(K) = \min_{c \in \mathbb{R}^m} \sum_{i=1}^{m} L(y_i, (Kc)_i) + \gamma \langle c, Kc \rangle$$

Assume $K := (K(x_i, x_j))_{i,j=1}^m$ is invertible and let $z = Kc$:

$$E(K) = \min_{z \in \mathbb{R}^m} \sum_{i=1}^{m} L(y_i, z_i) + \gamma \langle z, K^{-1} z \rangle$$

Function $(z, K) \mapsto \langle z, K^{-1} z \rangle$ is jointly convex, thus E is convex

Relation to mixed norm regularization

Proposition *[Micchelli & P. 05]* If $\mathcal{K} = \textit{conv-hull}\{K_1, \ldots, K_n\}$ then $(**)$ is equivalent to the problem

$$\min \left\{ \sum_{i=1}^{m} L\Big(y_i, \sum_{j=1}^{n} f_j(x_i)\Big) + \gamma \left(\sum_{j=1}^{n} \|f_j\|_{K_j} \right)^2 : f_j \in H_{K_j} \right\}$$

If $\hat{f}_1, \ldots, \hat{f}_n$ is a solution, then $\sum_j \hat{\lambda}_j K_j$ is a solution to $(**)$, with

$$\hat{\lambda}_j = \frac{\|\hat{f}_j\|_{K_j}}{\sum_\ell \|\hat{f}_\ell\|_{K_\ell}}$$

Related work: *[Bach et al. 04, Lin and Zhang 06, Ravikumar et al. 09]*

Relation to mixed norm regularization (cont.)

$$\min \left\{ \sum_{i=1}^{m} L\Big(y_i, \sum_{j=1}^{n} f_\ell(x_i)\Big) + \gamma \left(\sum_{j=1}^{n} \|f_j\|_{K_j} \right)^2 : f_j \in H_{K_j} \right\}$$

The above observation uses *[Aronszajn, 50]*:

- $\|f\|^2_{\sum_j K_j} = \inf \left\{ \sum_j \|f_j\|^2_{K_j} : f_j \in H_{K_j},\ \sum_j f_j = f \right\}$
- $\|f\|^2_{\lambda_j K_j} = \frac{1}{\lambda_j} \|f\|^2_{K_j}$

and a variational form for the ℓ_1 norm

- $\|\beta\|_1^2 = \inf \left\{ \sum_j \frac{\beta_j^2}{\lambda_j} : \lambda > 0,\ \sum_j \lambda_j \leq 1 \right\}$

Note: can be extended to ℓ_p, $p \in (0, 2)$ (see also *[Kloft et al. 09]*)

Lasso and Group Lasso

Two important "parametric" versions of MKL:

- **Lasso:** Choose $f_j(x) = \beta_j x_j$, $K_j(x,t) = x_j t_j$

$$\sum_{i=1}^{m} L(y_i, \langle \beta, x_i \rangle) + \gamma \left(\sum_{j=1}^{d} |\beta_j| \right)^2$$

- **Group Lasso** Choose $f_j(x) = \sum_{j \in J_\ell} \beta_j x_j$, $K_j(x,t) = \langle x_{|J_\ell}, t_{|J_\ell} \rangle$, where $\{J_\ell\}_{\ell=1}^{n}$ is a partition of $\mathbb{N}_d$

$$\sum_{i=1}^{m} L(y_i, \langle \beta, x_i \rangle) + \gamma \left(\sum_{\ell=1}^{n} \|\beta_{|J_\ell}\|_2 \right)^2$$

Plan of the talk

- Problem formulation and relation to the Group Lasso
- Convex combinations of continuously parameterized kernels
- Learning the kernel in multi-task learning
- Constrained multiple kernel learning and structured sparsity

Continuous Parameterization of Kernels

Let $\mathcal{P}(\Omega)$ be the set of *probability measures* on $\Omega \subseteq \mathbb{R}^p$ and G a *kernel mapping* such that $G(\omega)(x,x) \leq \kappa$ for all ω and x. Choose

$$\mathcal{K} := \left\{ K \ : \ K(x,t) = \int_\Omega G(\omega)(x,t) dp(\omega) : p \in \mathcal{P}(\Omega) \right\}$$

- Example: $\Omega \subseteq [0,\infty)$ and $G(\omega)(x,t) = e^{-\omega \|x-t\|^2}$
  Schoenberg's Theorem yields the set of *radial kernels*
- Higher dimensional Ω: $G(\omega) = e^{-\sum_i \omega_i (x_i - t_i)^2}$, etc.
- If Ω is a finite set we recover standard MKL

Note: related work *[Gehler & Nowozin, 08]*; connection to mixed norm regularization *[Micchelli and P. 07]*;

Convex Optimization

Recall: $E(K) = \min_{c \in \mathbb{R}^m} \sum_{i=1}^{m} L(y_i, (Kc)_i) + \gamma \langle c, Kc \rangle$

Let $Q(z) = \sum_{i=1}^{m} L(y_i, z_i)$ and define its *conjugate function*:

$$Q^*(v) = \sup_{z \in \mathbb{R}^m} \langle z, v \rangle - Q(z)$$

Using Fenchel duality, problem $(**)$ reduces to the **minimax problem**

$$\min_{K \in \mathcal{K}} E(K) = - \max_{K \in \mathcal{K}} \min_{c \in \mathbb{R}^m} \frac{1}{4\gamma} \langle c, Kc \rangle + Q^*(c)$$

Characterization of a Solution

$$\max_{K \in \mathcal{K}} \min_{c \in \mathbb{R}^m} \frac{1}{4\gamma} \langle c, Kc \rangle + Q^*(c)$$

Theorem *[Argyriou et al. 05]* $\hat{c}$ and $\hat{K} = \int_\Omega G(\omega) d\hat{p}(\omega)$ solve the above problem *iff*

1. $\langle \hat{c}, G(\omega)\, \hat{c} \rangle = \max_{\omega \in \Omega} \langle \hat{c}, G(\omega)\, \hat{c} \rangle \quad \forall\, \omega \in \mathrm{supp}(\hat{p})$

2. $\hat{c}$ is a solution to $\min_{c \in \mathbb{R}^m} \; Q(\hat{K} c) + \gamma \langle c, \hat{K} c \rangle$

Moreover there exists a solution involving **at most** $m + 1$ kernels:

$$\hat{K} = \sum_{i=1}^{m+1} \lambda_i\, G(\omega_i)$$

Algorithm

Choose $K^{(1)} \in \mathcal{K}$ and incrementally build an estimate of the solution:

1. Compute $\hat{c} = \underset{c \in \mathbb{R}^m}{\operatorname{argmin}} \; Q(K^{(t)}c) + \gamma \langle c, K^{(t)}c \rangle$
2. Find $\hat{\omega} \in \operatorname{argmax}_{\omega \in \Omega} \langle \hat{c}, G(\omega)\hat{c} \rangle$
3. Let $K^{(t+1)}$ be the optimal convex comb. of $G(\hat{\omega})$ and $K^{(t)}$ and return to Step 1

Theorem *[Argyriou 07]* There exists a limit point of the algorithm and any limit point is a solution of $(**)$

MNIST Experiments

- Gaussian basic kernels with two parameters (associated with left and right halves of images)
- Compared with varying finite grids of basic kernels

Task	LTK	5×5	10×10	LTK	5×5	10×10	LTK	5×5	10×10
	$\omega \in [75, 25000]^2$			$\omega \in [100, 10000]^2$			$\omega \in [500, 5000]^2$		
odd-even	5.8	15.8	11.2	5.8	10.1	6.2	5.8	6.8	5.8
3 vs. 8	2.7	6.5	5.1	2.5	4.6	2.5	2.6	3.5	2.5
4 vs. 7	1.8	3.9	2.9	1.7	2.7	2.0	1.8	2.0	1.8

- More robust than MKL wrt. parameter range
- Limitation: Step 2 is a non-convex optimization problem, difficult when ω is a high dimensional vector

Plan of the talk

- Problem formulation
- Convex combinations of continuously parameterized kernels
- Learning the kernel in multi-task learning
- Constrained multiple kernel learning and structured sparsity

Learning the kernel in multi-task learning

Choose $x = (z, t) \in \mathbb{R}^d \times \{1, \ldots, T\}$

$$E(K) = \min_{f \in H_K} \sum_{i=1}^{m} L(y_i, f(z_i, t_i)) + \gamma \|f\|_K^2$$

Learning the kernel $\equiv$ learning task relatedness

Linear case *[Evgeniou et al. 05]* $f(z, t) = \langle v, B_t z \rangle$. We choose

$$\mathcal{K} = \left\{ \langle B_t z, B_{t'} z' \rangle, \ B_t \in \mathcal{B}_t \subseteq \mathbb{R}^{N \times d} \right\}$$

and consider

$$\min_{K \in \mathcal{K}} E(K)$$

Classes of linear multi-task kernels (I)

Example: convex combinations of a "common task" and "independent tasks" kernels:

$$\mathcal{K} = \{\lambda n \langle z, z' \rangle \delta_{tt'} + (1-\lambda)\langle z, z' \rangle \ : \ \lambda \in [0,1]\}$$

Equivalent problem with quadratic regularizer relating the tasks: [*Evgeniou & P. 04*]

$$\min_{w_0, w_1, \ldots, w_T} \sum_{i=1}^{m} L(y_i, \langle w_{t_i}, z_i \rangle) + \gamma \left(\frac{1}{\lambda n} \sum_{t=1}^{T} \|w_t - w_0\|^2 + \frac{1}{1-\lambda} \|w_0\|^2 \right)$$

Extension: combine multiple models of task relatedness

Nonlinear classes: *[Baldassarre et al. 10, Caponnetto et al. 08, Castro 07, De Vito et al. 10, Micchelli and P. 05, Reisert & Burkhardt 07,...]*

Classes of linear multi-task kernels (II)

Common linear kernel:

$$\mathcal{K} = \{\langle z, Dz' \rangle \delta_{tt'} : D \in \mathbf{S}_+^d,\ \mathrm{tr}(D) \leq 1\}$$

Equivalent problems:

- Learning few common orthogonal features [*Argyriou et al. 06]*

$$\min_{U,A} \sum_{i=1}^{m} L\left(\langle a_{t_i}, U^\top z_i \rangle, y_i\right) + \gamma \left(\sum_{j=1}^{d} \|(a_{1j}, \ldots, a_{Tj})\|_2 \right)^2$$

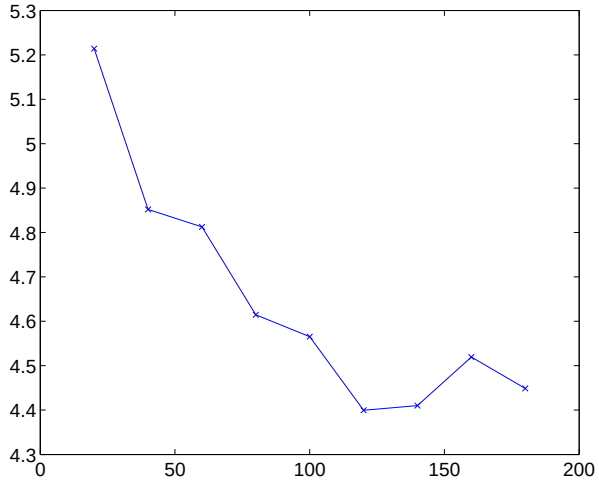
- Trace norm regularization [*Srebro et al. 05]*

$$\sum_{i=1}^{m} L\left(\langle w_{t_i}, z_i \rangle, y_i\right) + \gamma \|\sigma(W)\|_1^2$$

More examples: further structure on matrix D *[Argyriou et al. 07]*, task clustering kernel *[Evgeniou et al. 05, Jacob et al. 08]*, extension to multiple views etc.

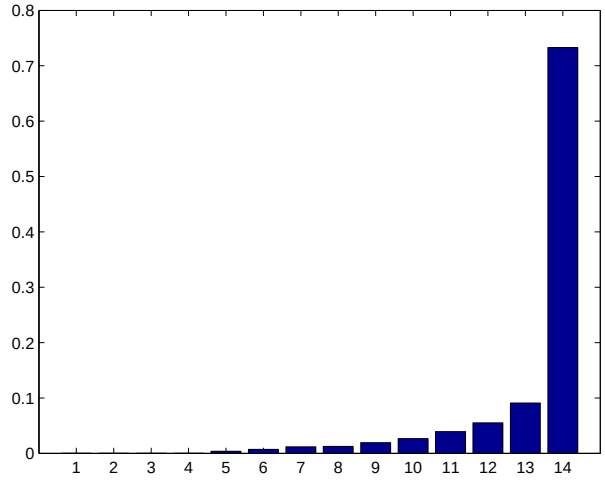
Experiment (Computer Survey) *[Argyriou et al. 08]*

Test error



#tasks

Eig(D)



- Performance improves with more tasks
- D approx. rank one (a single most important feature shared by the tasks)

Plan of the talk

- Problem formulation and relation to the Group Lasso
- Convex combinations of continuously parameterized kernels
- Learning the kernel in multi-task learning
- Constrained multiple kernel learning and structured sparsity

Structured sparsity [*Morales et al. 10*]

Let $\Lambda \subseteq \mathbb{R}^n_{++}$. We consider constrained convex combinations of kernels

$$\mathcal{K} = \left\{ \sum_j \lambda_j K_j : \lambda \in \Lambda \right\}$$

Problem $\min_{K \in \mathcal{K}} E(K)$ is now equivalent to

$$\min \left\{ \sum_{i=1}^m L\Big(y_i, \sum_{j=1}^n f_j(x_i)\Big) + \gamma \Omega\Big(\|f_1\|_{K_1}, \ldots, \|f_n\|_{K_n}\Big) : f_j \in H_j \right\}$$

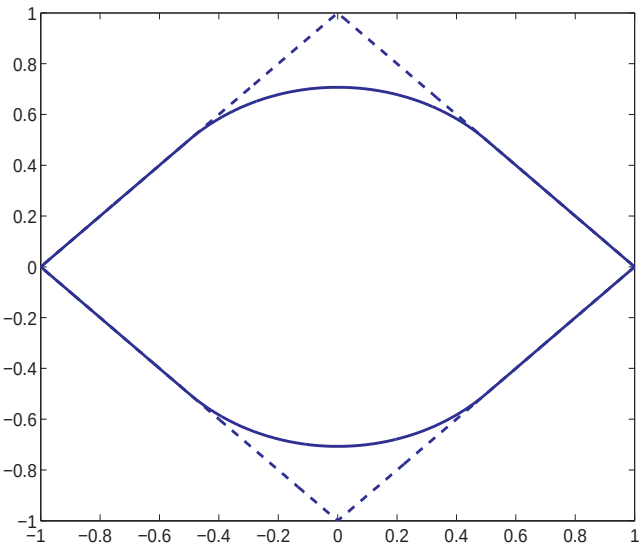
where

$$\Omega(\beta|\Lambda) = \frac{1}{2} \inf_{\lambda \in \boldsymbol{\Lambda}} \sum_{j=1}^n \left(\frac{\beta_j^2}{\lambda_j} + \lambda_j \right)$$

Constrained variational form

$$\Omega(\beta|\Lambda) = \frac{1}{2} \inf_{\lambda \in \boldsymbol{\Lambda}} \sum_{j=1}^{n} \left(\frac{\beta_j^2}{\lambda_j} + \lambda_j \right)$$

- Intuition: favor vectors β such that $|\beta| := (|\beta_1|, \ldots, |\beta_n|) \in \Lambda$



Example: $\Lambda = \{\lambda \in \mathbb{R}^2 : \lambda_1 \geq \lambda_2 > 0\}$

$$\Omega(\beta|\Lambda) = \begin{cases} \|\beta\|_1 & \text{if } |\beta_1| > |\beta_2| \\ \sqrt{2}\|\beta\|_2 & \text{otherwise} \end{cases}$$

Some properties

$$\Omega(\beta|\Lambda) = \frac{1}{2} \inf_{\lambda \in \Lambda} \sum_{j=1}^{n} \left(\frac{\beta_j^2}{\lambda_j} + \lambda_j \right)$$

- For any β and Λ, it holds that $\Omega(\beta|\Lambda) \geq \|\beta\|_1$ and equality holds iff $|\beta| \in \Lambda$
- If Λ is a convex set then the function Ω is convex as well
- If Λ is a convex cone then Ω is a norm and

$$\Omega(\beta|\Lambda) \leq \|\beta\|_1 \max_{k=1}^{n} \Omega(e_k|\Lambda)$$

Some examples of set Λ

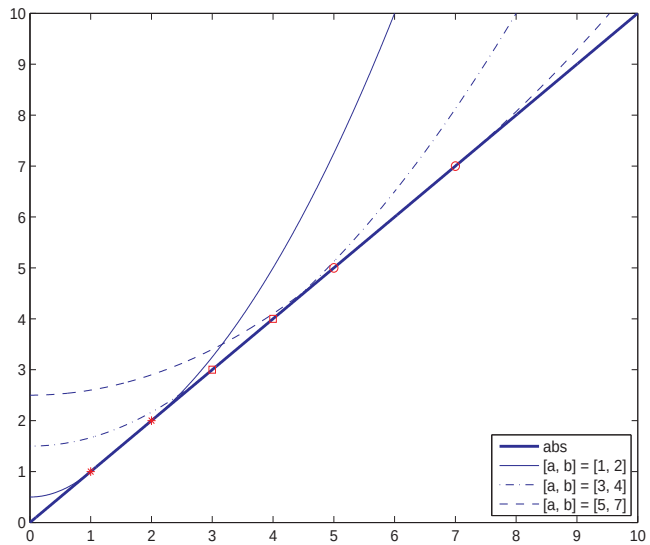
- Box: $[a, b]^n$
- Ordering relationships
- Sparsity pattern consists of few "connected regions"
- Hierarchical constraints: compute a vector formed by ℓ_1-norm across prescribed groups, then use the variational form

Example 1: box penalty

If $\beta \in \mathbb{R}$, we find

$$\Omega(\beta|[a,b]) = |\beta| + \frac{1}{2a}(a - |\beta|)_+^2 \frac{1}{2b}(|\beta| - b)_+^2$$

prevents $|\beta|$ to be too large and/or too small *[Jacob 09, Owen 07]*

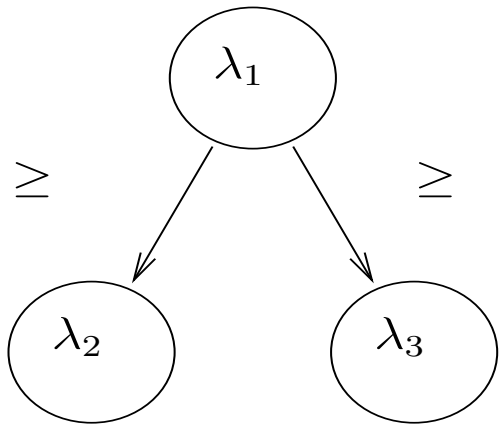


immediate extension to higher dimensions

Example 2: graph penalty

Let A be the incidence matrix of a directed acyclic graph (DAG) and choose

$$\Lambda = \left\{\lambda \in \mathbb{R}^n_{++} : A\lambda \geq 0\right\}$$



$$A = \begin{bmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \end{bmatrix}$$

- The constraints $A\lambda \geq 0$ model ordering relationships between pairs of components of λ
- Useful in the context of ANOVA models

Related work: [*Rocha et al., 09*]

Example 3: k-wedge

The wedge is a first order difference constraint. Let

$$\Lambda = \left\{ \lambda \in \mathbb{R}^n_{++} : D^k(\lambda) \geq 0 \right\},$$

where D^k is the k-th order difference operator:

$$D^k(\lambda) = \left(\sum_{\ell=0}^{k} (-1)^\ell \binom{k}{\ell} \lambda_{j+k-\ell} : j = 1, \ldots, n-k \right)$$

- If $k = 1$, we have a "reversed" wedge (increasing coordinates)
- When $k > 1$, the penalty encourages vectors whose sparsity pattern is concentrated on few **contiguous regions**

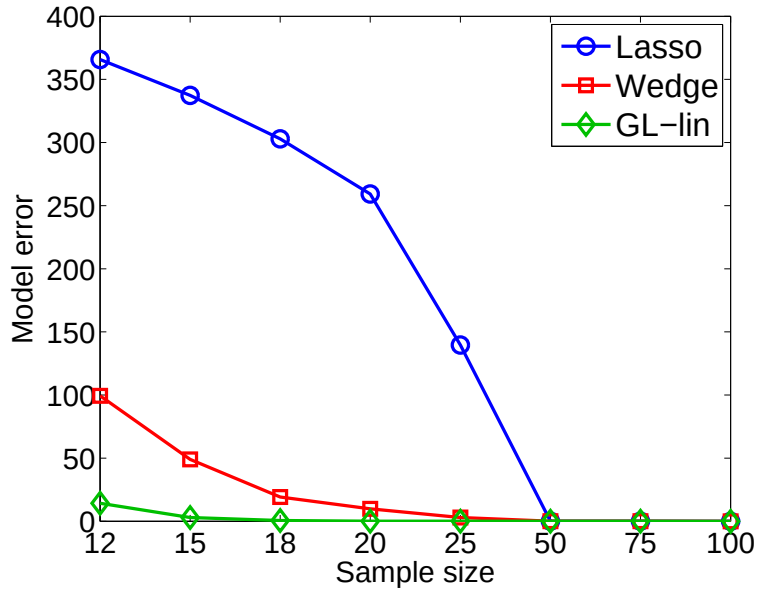
Optimization method

$$\min \left\{ \|y - X\beta\|_2^2 + \frac{\gamma}{2} \sum_{i=1}^{n} \left(\frac{\beta_i^2}{\lambda_i} + \lambda_i \right) : \beta \in \mathbb{R}^n, \lambda \in \Lambda \right\}$$

- This is a **convex optimization problem** which we solve by alternating minimization
- β-step is a simple least squares problem
- λ-step requires a solver (e.g. CVX) but in special cases (tree graph penalty) can be solved efficiently (e.g. $O(n)$ time)
- Convergence proof requires a perturbation of the penalty term (in order to ensure that the λ_i are bounded away from zero)

Experiment 1: line graph

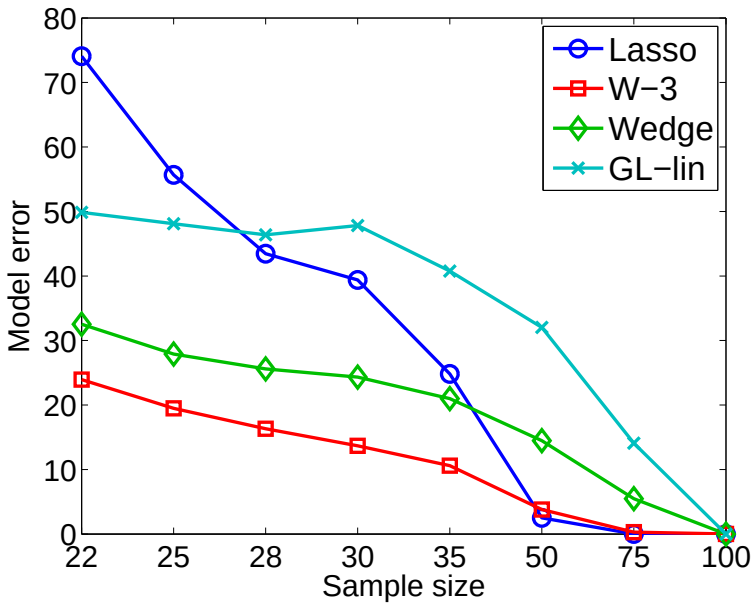
$\beta^* \in \mathbb{R}^{100}$, first 10 values decreasing from 10 to 1



Average 200 trials, comparing line-penalty with Lasso and Hierarchical Group Lasso

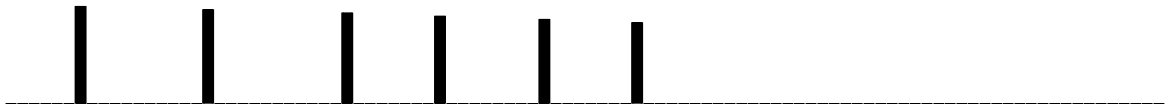
Experiment 2: two contiguous regions

$\beta^* \in \mathbb{R}^{100}$ with 5 initial and 15 in the middle nonzero values, generated from a cubic polynomial

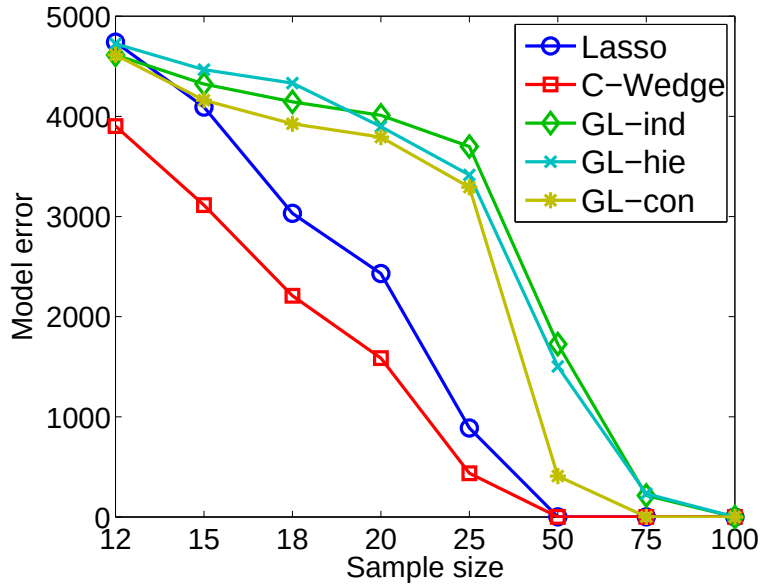


Average 200 trials, comparing line-penalty with Lasso and Hierarchical Group Lasso

Experiment 3: composite wedge

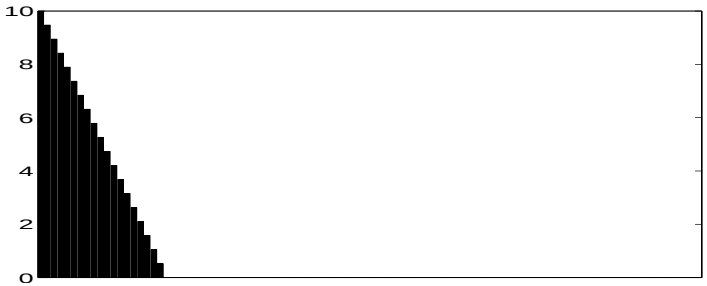


$\beta^* \in \mathbb{R}^{100}$ with 6 nonzero values 30, . . . , 25
each in a random position in a successive group of 10 positions

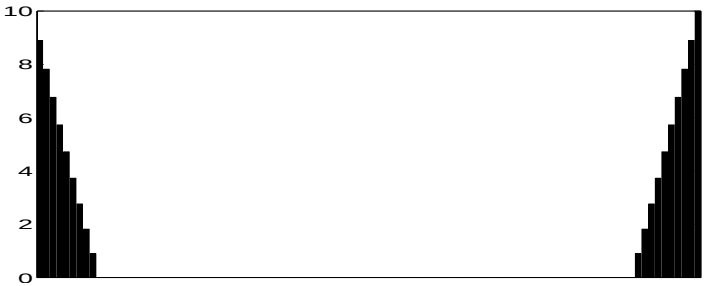


Comparing the C-wedge with the Lasso and group Lasso variants

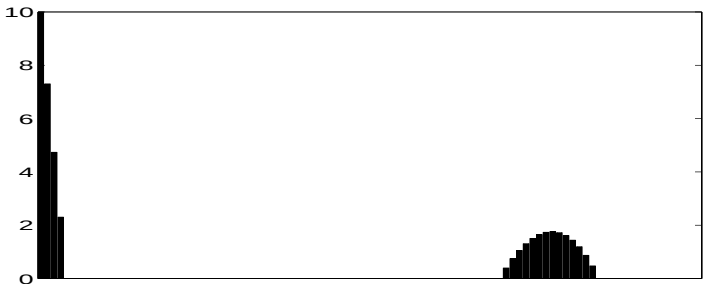
Experiment 4: k-wedge



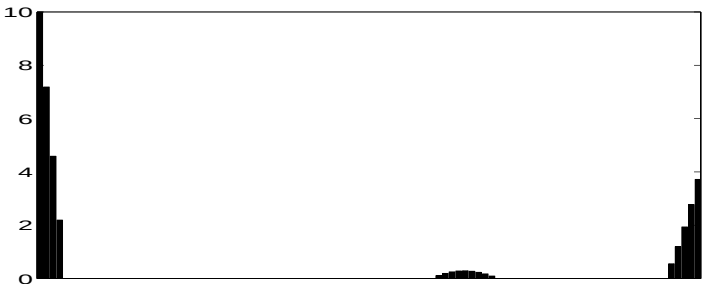
wedge



$k = 2$



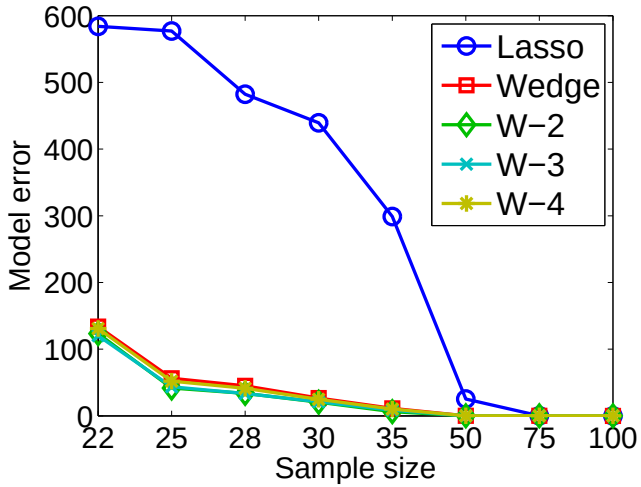
$k = 3$



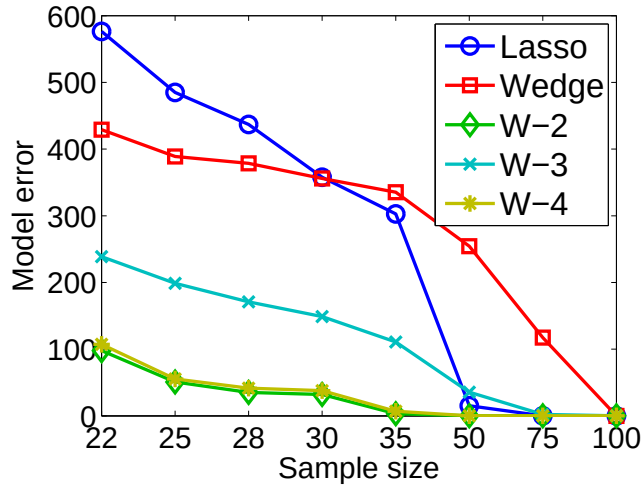
$k = 4$

Silhouette of the polynomials by number of degree k

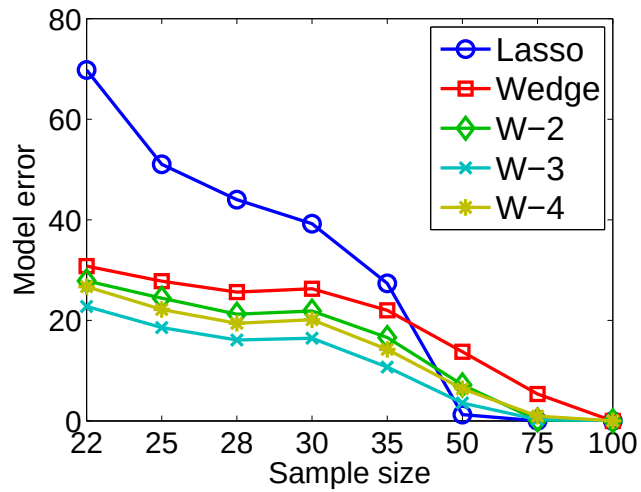
Experiment 4 (contd.)



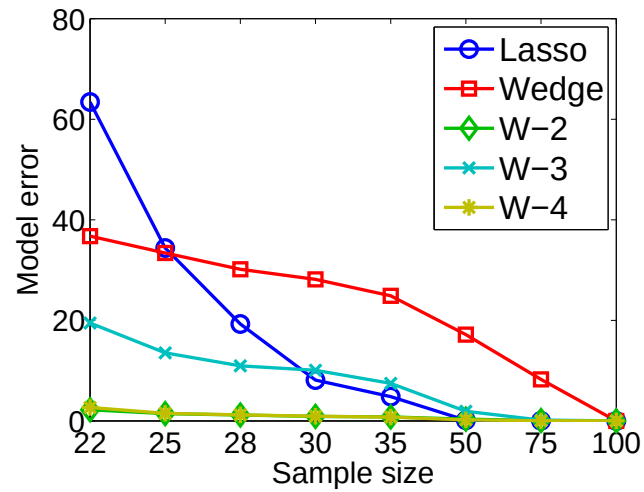
wedge



$k = 2$



$k = 3$



$k = 4$

Comparison between $\Omega(\beta|W^k)$, $k = 1, 2, 3$, Wedge and Lasso

Conclusions

- Continuous classes may prove advantageous over finite convex combinations; require solving a hard optimization problem
- Learning the kernel in multi-task learning as a means to model task relatendness; richer kernel classes require further explorations
- Family of convex regularizers for structured sparsity: improvement over Lasso and Group Lasso