Kernel machines and sparsity

2 juillet, 2009
ENBIS'09, Saint Etienne

Laboratoire d'Informatique, de Traitement de l'Information et des Systèmes

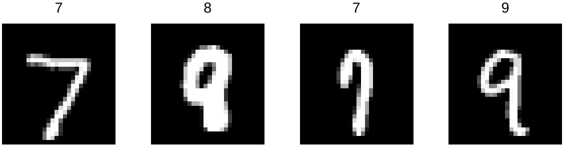
Stéphane Canu & Alain Rakotomamonjy
stephane.canu@litislab.eu

Roadmap

Optical character recognition

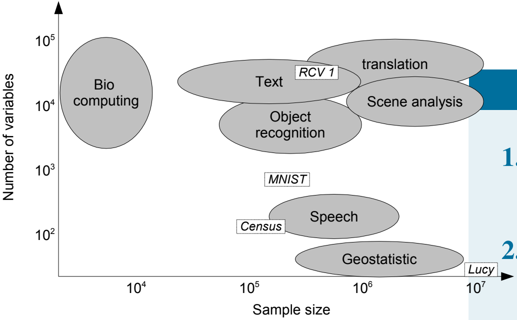
Example (The MNIST database)

- MNIST[1], data = « image-label »
- $n = 60,000$; $d = 700$; classes $= 10$
- Kernel error rate = 0.56 %,
- Best error rate = 0.4 % .



[1] http://yann.lecun.com/exdb/mnist/index.html

Learning challenges: the size effect



3 key issues

1. learn any problem:
   - functional universality
2. from data:
   - statistical consistency
3. with large data sets:
   - computational efficency

kernel machines adress these three issues
(up to a certain point regarding efficency)

L. Bottou, 2006

Kernel machines

Definition (Kernel machines)

$$\mathcal{A}\big((\mathbf{x}_i, y_i)_{i=1,n}\big)(\mathbf{x}) = \psi\Big(\sum_{i=1}^{n} \alpha_i k(\mathbf{x}, \mathbf{x}_i) + \sum_{j=1}^{p} \beta_j q_j(\mathbf{x})\Big)$$

$\boldsymbol{\alpha}$ et $\boldsymbol{\beta}$: parameters to be estimated.

Exemples

$\mathcal{A}(x) = \sum_{i=1}^{n} \alpha_i (x - x_i)_+^3 + \beta_0 + \beta_1 x$	splines
$\mathcal{A}(\mathbf{x}) = \text{sign}\Big(\sum_{i \in I} \alpha_i \exp^{-\frac{\lVert \mathbf{x} - \mathbf{x}_i \rVert^2}{b}} + \beta_0\Big)$	SVM
$\mathbb{P}(y \mid \mathbf{x}) = \frac{1}{Z} \exp\Big(\sum_{i \in I} \alpha_i \mathbb{1}_{\{y = y_i\}} (\mathbf{x}^\top \mathbf{x}_i + b)^2\Big)$	exponential family

Roadmap

In the beginning was the kenrel...

Definition (Kernel)

a function of two variable k from $\mathcal{X} \times \mathcal{X}$ to $\mathbb{R}$

Definition (Positive kernel)

A kernel $k(s,t)$ on $\mathcal{X}$ is said to be positive

- if it is symetric: $k(s,t) = k(t,s)$
- an if for any finite positive interger n:

$$\forall \{\alpha_i\}_{i=1,n} \in \mathbb{R}, \forall \{\mathbf{x}_i\}_{i=1,n} \in \mathcal{X}, \quad \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j) \geq 0$$

it is strictly positive if for $\alpha_i \neq 0$

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j) > 0$$

Examples of positive kernels

the linear kernel: $\mathbf{s}, \mathbf{t} \in \mathbb{R}^d, \quad k(\mathbf{s}, \mathbf{t}) = \mathbf{s}^\top \mathbf{t}$

symetric: $\mathbf{s}^\top \mathbf{t} = \mathbf{t}^\top \mathbf{s}$

positive:
$$\sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j) = \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j \mathbf{x}_i^\top \mathbf{x}_j = \left(\sum_{i=1}^n \alpha_i \mathbf{x}_i\right)^\top \left(\sum_{j=1}^n \alpha_j \mathbf{x}_j\right) = \left\|\sum_{i=1}^n \alpha_i \mathbf{x}_i\right\|^2$$

the product kernel: $k(\mathbf{s}, \mathbf{t}) = g(\mathbf{s}) g(\mathbf{t})$ for some $g : \mathbb{R}^d \to \mathbb{R}$,

symetric by construction

positive:
$$\sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j) = \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j g(\mathbf{x}_i) g(\mathbf{x}_j) = \left(\sum_{i=1}^n \alpha_i g(\mathbf{x}_i)\right) \left(\sum_{j=1}^n \alpha_j g(\mathbf{x}_j)\right) = \left(\sum_{i=1}^n \alpha_i g(\mathbf{x}_i)\right)^2$$

k is positive $\Leftrightarrow$ (its square root exists) $\Leftrightarrow$ $k(\mathbf{s}, \mathbf{t}) = \langle \phi_{\mathbf{s}}, \phi_{\mathbf{t}} \rangle$

positive definite Kernel (PDK) algebra (closure)

if $k_1(\mathbf{s},\mathbf{t})$ and $k_2(\mathbf{s},\mathbf{t})$ are two positive kernels

- DPK are a convex cone: $\forall a_1 \in \mathbb{R}^+ \quad a_1 k_1(\mathbf{s},\mathbf{t}) + k_2(\mathbf{s},\mathbf{t})$
- product kernel $k_1(\mathbf{s},\mathbf{t})k_2(\mathbf{s},\mathbf{t})$

proofs

- by linearity:

$$\sum_{i=1}^n\sum_{j=1}^n \alpha_i\alpha_j\big(a_1k_1(i,j)k_2(i,j)\big) = a_1\sum_{i=1}^n\sum_{j=1}^n \alpha_i\alpha_j k_1(i,j) + \sum_{i=1}^n\sum_{j=1}^n \alpha_i\alpha_j k_2(i,j)$$

- by linearity: $\displaystyle\sum_{i=1}^n\sum_{j=1}^n \alpha_i\alpha_j\big(\psi(\mathbf{x}_i)\psi(\mathbf{x}_j)\big) = \Big(\sum_{i=1}^n \alpha_i\psi(\mathbf{x}_i)\Big)\Big(\sum_{j=1}^n \alpha_j\psi(\mathbf{x}_j)\Big)$

- assuming $\exists \psi_\ell$ s.t. $k_1(\mathbf{s},\mathbf{t}) = \sum_\ell \psi_\ell(\mathbf{s})\psi_\ell(\mathbf{t})$

$$\sum_{i=1}^n\sum_{j=1}^n \alpha_i\alpha_j\, k_1(\mathbf{x}_i,\mathbf{x}_j)k_2(\mathbf{x}_i,\mathbf{x}_j) = \sum_{i=1}^n\sum_{j=1}^n \alpha_i\alpha_j\Big(\sum_\ell \psi_\ell(\mathbf{x}_i)\psi_\ell(\mathbf{x}_j)k_2(\mathbf{x}_i,\mathbf{x}_j)\Big)$$
$$= \sum_\ell\sum_{i=1}^n\sum_{j=1}^n \big(\alpha_i\psi_\ell(\mathbf{x}_i)\big)\,\big(\alpha_j\psi_\ell(\mathbf{x}_j)\big)\, k_2(\mathbf{x}_i,\mathbf{x}_j)$$

N. Cristianini and J. Shawe Taylor, kernel methods for pattern analysis, 2004

Kernel engineering: building PDK

- for any polynomial with positive coef. ϕ from $\mathbb{R}$ to $\mathbb{R}$ $\qquad \phi\big(k(\mathbf{s},\mathbf{t})\big)$
- if $\Psi$is a function from $\mathbb{R}^d$ to $\mathbb{R}^d$ $\qquad k\big(\Psi(\mathbf{s}),\Psi(\mathbf{t})\big)$
- if φ from $\mathbb{R}^d$ to $\mathbb{R}^+$, is minimum in 0 $\qquad k(\mathbf{s},\mathbf{t}) = \varphi(\mathbf{s}+\mathbf{t}) - \varphi(\mathbf{s}-\mathbf{t})$
- convolution of two positive kernels is a positive kernel $\qquad K_1 \star K_2$

the Gaussian kernel is a PDK

$$\begin{aligned} \exp(-\|\mathbf{s}-\mathbf{t}\|^2) &= \exp(-\|\mathbf{s}\|^2 - \|\mathbf{t}\|^2 - 2\mathbf{s}^\top\mathbf{t}) \\ &= \exp(-\|\mathbf{s}\|^2)\exp(-\|\mathbf{t}\|^2)\exp(2\mathbf{s}^\top\mathbf{t}) \end{aligned}$$

- $\mathbf{s}^\top\mathbf{t}$ is a PDK and function exp as the limit of positive series expansion, so $\exp(2\mathbf{s}^\top\mathbf{t})$ is a PDK
- $\exp(-\|\mathbf{s}\|^2)\exp(-\|\mathbf{t}\|^2)$ is a PDK as a product kernel
- the product of two PDK is a PDK

some examples of PD kernels...

type	name	$k(s,t)$
radial	gaussian	$\exp\left(-\frac{r^2}{b}\right), \quad r = \|s-t\|$
radial	laplacian	$\exp(-r/b)$
radial	rationnal	$1 - \frac{r^2}{r^2+b}$
radial	loc. gauss.	$\max\left(0, 1 - \frac{r}{3b}\right)^d \exp(-\frac{r^2}{b})$
non stat.	χ^2	$\exp(-r/b), \; r = \sum_k \frac{(s_k-t_k)^2}{s_k+t_k}$
projective	polynomial	$(s^\top t)^p$
projective	affine	$(s^\top t + b)^p$
projective	cosine	$s^\top t / \|s\| \|t\|$
projective	correlation	$\exp\left(\frac{s^\top t}{\|s\|\|t\|} - b\right)$

Most of the kernels depends on a quantity b called the bandwidth

kernels for objects and structures

kernels on histograms and probability distributions

$$k(p(x), q(x)) = \int k_i(p(x), q(x)) \mathbb{P}(x) dx$$

kernel on strings

- spectral string kernel $k(\mathbf{s}, \mathbf{t}) = \sum_u \phi_u(\mathbf{s}) \phi_u(\mathbf{t})$
- using sub sequences
- similarities by alignements $k(\mathbf{s}, \mathbf{t}) = \sum_\pi \exp(\beta(\mathbf{s}, \mathbf{t}, \pi))$

kernels on graphs

- the pseudo inverse of the (regularized) graph Laplacian

  $L = D - A$ A is the adjency matrix$D$ the degree matrix

- diffusion kernels $\frac{1}{Z(b)} \exp^{bL}$
- subgraph kernel convolution (using random walks)

and kernels on heterogeneous data (image), HMM, automata...

Shawe-Taylor & Cristianini's Book, 2004 ; JP Vert, 2006

Roadmap

From kernel to functions

$$\mathcal{H}_0 = \left\{ f \,\middle|\, m_f < \infty; f_j \in \mathbb{R}; t_j \in \mathcal{X},\ f(\mathbf{x}) = \sum_{j=1}^{m_f} f_j k(\mathbf{x}, t_j) \right\}$$

let define the bilinear form $\left(g(\mathbf{x}) = \sum_{i=1}^{m_g} g_i k(\mathbf{x}, s_i) \right)$:

$$\forall f, g \in \mathcal{H}_0,\ \langle f, g \rangle_{\mathcal{H}_0} = \sum_{j=1}^{m_f} \sum_{i=1}^{m_g} f_j\, g_i\, k(t_j, s_i)$$

Evaluation functional: $\forall \mathbf{x} \in \mathcal{X}$

$$f(\mathbf{x}) = \langle f(.), k(\mathbf{x}, .) \rangle_{\mathcal{H}_0}$$

from k to $\mathcal{H}$

with any postive kernel, a hypothesis set $\mathcal{H} = \bar{\mathcal{H}_0}$ can be constructed with its metric

RKHS

Definition (reproducing kernel Hibert space (RKHS))

a Hilbert space $\mathcal{H}$ embeded with the inner product $\langle ., . \rangle_{\mathcal{H}}$ is said to be with reproduicing kernel if it exists a positive kernel k such that

$$\forall s \in \mathcal{X}, k(., s) \in \mathcal{H} \text{ et } \forall f \in \mathcal{H}, \quad f(s) = \langle f(.), k(s, .) \rangle_{\mathcal{H}}$$

positive kernel ⇔ RKHS

- any function is pointwise defined
- defines the inner product
- it defines the regularity (smoothness) of the hypothesis set

functional differentiation in RKHS

Let J be a functional

$$J: \begin{array}{ccc} \mathcal{H} & \rightarrow & \mathbb{R} \\ f & \mapsto & J(f) \end{array} \qquad \text{examples:} \qquad J_1(f) = \|f\|^2, J_2(f) = f(\mathbf{x}),$$

J directional derivative in direction g at point f

$$dJ(f,g) = \lim_{\varepsilon \rightarrow 0} \frac{J(f+\varepsilon g) - J(f)}{\varepsilon}$$

Gradient $\nabla_J(f)$

$$\nabla_J: \begin{array}{ccc} \mathcal{H} & \rightarrow & \mathcal{H} \\ f & \mapsto & \nabla_J(f) \end{array} \qquad \text{if} \qquad dJ(f,g) = \langle \nabla_J(f), g \rangle_{\mathcal{H}}$$

exercice: find out $\nabla_{J_1}(f)$ et $\nabla_{J_2}(f)$

other kernels (what realy matters)

- finite kernels
$$k(\mathbf{s},\mathbf{t}) = \left(\phi_1(\mathbf{s}), ..., \phi_p(\mathbf{s})\right)^\top \left(\phi_1(\mathbf{t}), ..., \phi_p(\mathbf{t})\right)$$
- Mercer kernels

  positive on a compact set $\Leftrightarrow$ $k(\mathbf{s},\mathbf{t}) = \sum_{j=1}^{p} \lambda_j \phi_j(\mathbf{s})\phi_j(\mathbf{t})$
- positive kernels
- positive semi-definite
- conditionnaly positive (for some functions p_j)
$$\forall \{\mathbf{x}_i\}_{i=1,n}, \forall \alpha_i, \sum_i^n \alpha_i p_j(\mathbf{x}_i) = 0; \quad j = 1, p, \quad \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j) \geq 0$$
- symetric non positive
$$k(\mathbf{s},\mathbf{t}) = \tanh(\mathbf{s}^\top \mathbf{t} + \alpha_0)$$
- non symetric – non positive

the key property: $\nabla_{J_t}(f) = k(t, .)$ holds

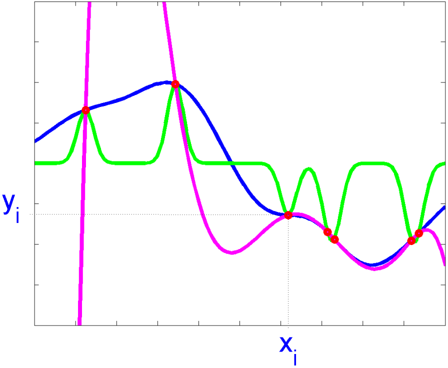
Let's summarize

- positive kernels $\Leftrightarrow$ RKHS $= \mathcal{H}$ $\Leftrightarrow$ regularity $\|f\|^2_{\mathcal{H}}$
- the key property: $\nabla_{J_t}(f) = k(t,.)$ holds not only for positive kernels
  $f(\mathbf{x}_i)$ exists (pointwise defined functions)
- universal consistency in RKHS
- the Gram matrix summarize the pairwise comparizons

Plan

Interpolation splines

find out $f \in \mathcal{H}$ such that $f(\mathbf{x}_i) = y_i, \qquad i = 1, ..., n$



It is an ill posed problem

Interpolation splines: minimum norm interpolation

$$\begin{cases} \min\limits_{f \in \mathcal{H}} & \frac{1}{2}\|f\|^2_{\mathcal{H}} \\ \text{such that} & f(\mathbf{x}_i) = y_i, \qquad i = 1, \ldots, n \end{cases}$$

The lagrangian (α_i Lagrange multipliers)

$$L(f, \boldsymbol{\alpha}) = \frac{1}{2}\|f\|^2 - \sum_{i=1}^{n} \alpha_i \big(f(\mathbf{x}_i) - y_i\big)$$

optimality for f: $\quad \nabla_f L(f, \boldsymbol{\alpha}) = 0 \quad \Leftrightarrow \quad f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i k(\mathbf{x}_i, \mathbf{x})$

dual formulation (remove f from the Lagrangian):

$$Q(\boldsymbol{\alpha}) = -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j) + \sum_{i=1}^{n} \alpha_i y_i \qquad \text{solution:} \quad \max_{\boldsymbol{\alpha} \in \mathbb{R}^n} Q(\boldsymbol{\alpha})$$

$$\boxed{K\boldsymbol{\alpha} = \mathbf{y}}$$

Interpolation splines: minimum norm interpolation

$$\begin{cases} \displaystyle\min_{f\in\mathcal{H}} & \frac{1}{2}\|f\|^2_{\mathcal{H}} \\ \text{such that} & f(\mathbf{x}_i) = y_i, \qquad i=1,\dots,n \end{cases}$$

The lagrangian (α_i Lagrange multipliers)

$$L(f,\boldsymbol{\alpha}) = \frac{1}{2}\|f\|^2 - \sum_{i=1}^{n} \alpha_i\big(f(\mathbf{x}_i) - y_i\big)$$

optimality for f: $\nabla_f L(f,\boldsymbol{\alpha}) = 0 \quad \Leftrightarrow \quad f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i k(\mathbf{x}_i, \mathbf{x})$

dual formulation (remove f from the Lagrangian):

$$Q(\boldsymbol{\alpha}) = -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j k(\mathbf{x}_i,\mathbf{x}_j) + \sum_{i=1}^{n}\alpha_i y_i \qquad \text{solution:} \quad \max_{\boldsymbol{\alpha}\in\mathbb{R}^n} Q(\boldsymbol{\alpha})$$

$$K\boldsymbol{\alpha} = \mathbf{y}$$

Interpolation splines: minimum norm interpolation

$$\begin{cases} \min_{f \in \mathcal{H}} & \frac{1}{2}\|f\|^2_{\mathcal{H}} \\ \text{such that} & f(\mathbf{x}_i) = y_i, \qquad i = 1, \dots, n \end{cases}$$

The lagrangian (α_i Lagrange multipliers)

$$L(f, \boldsymbol{\alpha}) = \frac{1}{2}\|f\|^2 - \sum_{i=1}^{n} \alpha_i \big(f(\mathbf{x}_i) - y_i\big)$$

optimality for f: $\quad \nabla_f L(f, \boldsymbol{\alpha}) = 0 \quad \Leftrightarrow \quad f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i k(\mathbf{x}_i, \mathbf{x})$

dual formulation (remove f from the Lagrangian):

$$Q(\boldsymbol{\alpha}) = -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j) + \sum_{i=1}^{n} \alpha_i y_i \qquad \text{solution:} \quad \max_{\boldsymbol{\alpha} \in \mathbb{R}^n} Q(\boldsymbol{\alpha})$$

$$\boxed{K\boldsymbol{\alpha} = y}$$

Representer theorem

Theorem (Representer theorem)

Let $\mathcal{H}$ be a RKHS with kernel $k(s,t)$. Let ℓ be a function from $\mathcal{X}$ to $\mathbb{R}$ (loss function) and Φ a non decreasing function from $\mathbb{R}$ to $\mathbb{R}$. If there exists a function f^*minimizing:

$$f^* = \underset{f \in \mathcal{H}}{argmin} \; \sum_{i=1}^{n} \ell\big(y_i, f(\mathbf{x}_i)\big) + \Phi\big(\|f\|^2_{\mathcal{H}}\big)$$

then there exists a vector $\boldsymbol{\alpha} \in \mathbb{R}^n$ such that:

$$f^*(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i k(\mathbf{x}, \mathbf{x}_i)$$

it can be generalized to the semi parametric case: $+\sum_{j=1}^{m} \beta_j \phi_j(\mathbf{x})$

Smoothing splines

introducing the error (the slack) $\xi = f(x_i) - y_i$

$$(\mathcal{S}) \quad \begin{cases} \displaystyle\min_{f \in \mathcal{H}} & \frac{1}{2}\|f\|^2_{\mathcal{H}} + \frac{1}{2\lambda}\sum_{i=1}^{n} \xi_i^2 \\ \text{such that} & f(x_i) = y_i + \xi_i, \qquad i = 1, n \end{cases}$$

three equivalents definitions

$$(\mathcal{S}') \quad \min_{f \in \mathcal{H}} \frac{1}{2}\sum_{i=1}^{n}\big(f(x_i) - y_i\big)^2 + \frac{\lambda}{2}\|f\|^2_{\mathcal{H}}$$

$$\begin{cases} \displaystyle\min_{f \in \mathcal{H}} & \frac{1}{2}\|f\|^2_{\mathcal{H}} \\ \text{such that} & \displaystyle\sum_{i=1}^{n}\big(f(\mathbf{x}_i) - y_i\big)^2 \le C' \end{cases} \qquad \begin{cases} \displaystyle\min_{f \in \mathcal{H}} & \displaystyle\sum_{i=1}^{n}\big(f(\mathbf{x}_i) - y_i\big)^2 \\ \text{such that} & \|f\|^2_{\mathcal{H}} \le C'' \end{cases}$$

using the representer theorem $(\mathcal{S}'') \quad \displaystyle\min_{\boldsymbol{\alpha} \in \mathbb{R}^n} \frac{1}{2}\|K\boldsymbol{\alpha} - \mathbf{y}\|^2 + \frac{\lambda}{2}\boldsymbol{\alpha}^\top K \boldsymbol{\alpha}$

$\Leftrightarrow (K + \lambda I)\boldsymbol{\alpha} = \mathbf{y}$

using $\displaystyle\min_{\alpha \in \mathbb{R}^n} \frac{1}{2}\|K\boldsymbol{\alpha} - \mathbf{y}\|^2 + \frac{\lambda}{2}\boldsymbol{\alpha}^\top \boldsymbol{\alpha} \Leftrightarrow \boldsymbol{\alpha} = (K^\top K + \lambda I)^{-1} K^\top \mathbf{y}$

From 0 to interpollation

- problem: $\min_{f \in \mathcal{H}} \frac{1}{2} \sum_{i=1}^{n} \big(f(x_i) - y_i\big)^2 + \frac{\lambda}{2} \|f\|_{\mathcal{H}}^2$
- solution: $\boldsymbol{\alpha}(\lambda) = (K + \lambda I)^{-1} \mathbf{y}$

- $\lambda = 0 : \rightarrow \boldsymbol{\alpha}(0) = K^{-1}\mathbf{y}$: interpollation
- $\lambda = \infty : \rightarrow \boldsymbol{\alpha}(\infty) = 0$:

Regularization path $\mathcal{S}$: set of solutions as a function of λ

$$\mathcal{S} = \{\boldsymbol{\alpha}(\lambda) \mid \lambda \in [0, \infty[\}$$

also called the *solution path*

1D Ridge Regression in the costs domain

the Loss term L as a function of the penalty term P

$$\min_{\alpha \in \mathbb{R}} \sum_{i=1}^{n} (x_i \alpha - y_i)^2 + \lambda \, \alpha^2$$

$$\begin{cases} L(\alpha) = \sum_{i=1}^{n} (x_i \alpha - y_i)^2 \\ P(\alpha) = \alpha^2 \end{cases}$$

$$\begin{cases} L(P) = aP \pm b\sqrt{P} + c \\ a, b \text{ and } c \in \mathbb{R} \end{cases}$$

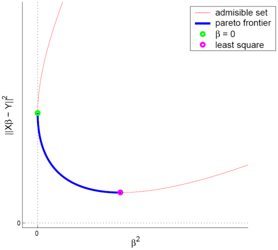


Figure: regularization path as a function of the criteria L and P.

How to tune the regularization parameter λ?

- brute force

  for each $\lambda_1 < \lambda_2 < \ldots < \lambda_k < \ldots < \lambda_K$

  compute $\boldsymbol{\alpha}_k = (K + \lambda_k I)^{-1}\, \mathbf{y}\,, \quad k = 1, K$

  $\mathcal{O}\big(Kn^3\big)$

- warm start

  $\boldsymbol{\alpha}_k = \Phi(\boldsymbol{\alpha}_{k-1})$ (using ℓ conjugate gradient iterations)

  $\mathcal{O}\big(K\ell n^2\big)$

- warm start + prediction step

  $\boldsymbol{\alpha}_k^{(p)} = \boldsymbol{\alpha}_{k-1} + \rho\nabla_{\boldsymbol{\alpha}}\big(L(\boldsymbol{\alpha}_{k-1}) + \lambda_k P(\boldsymbol{\alpha}_{k-1})\big)$ (prediction)

  $\boldsymbol{\alpha}_k = \Phi(\boldsymbol{\alpha}_k^{(p)})$ (correction step using CG)

  $\mathcal{O}\big(K\ell' n^2\big)$

- use only the prediction step!

  $\boldsymbol{\alpha}_k = \boldsymbol{\alpha}_{k-1} + \lambda_k \Psi(\boldsymbol{\alpha}_{k-1})$ (prediction step)

  to do so the regularization path has to be piecewise linear

  $\mathcal{O}\big(Kn^2\big)$

How to tune the regularization parameter λ?

- brute force

  for each $\lambda_1 < \lambda_2 < \ldots < \lambda_k < \ldots < \lambda_K$
  compute $\boldsymbol{\alpha}_k = (K + \lambda_k I)^{-1}\, \mathbf{y}\,, \quad k = 1, K$

  $\mathcal{O}\big(Kn^3\big)$

- warm start

  $\boldsymbol{\alpha}_k = \Phi(\boldsymbol{\alpha}_{k-1})$ (using ℓ conjugate gradient iterations)

  $\mathcal{O}\big(K\ell n^2\big)$

- warm start + prediction step

  $\boldsymbol{\alpha}_k^{(p)} = \boldsymbol{\alpha}_{k-1} + \rho\nabla_{\boldsymbol{\alpha}}(L(\boldsymbol{\alpha}_{k-1}) + \lambda_k P(\boldsymbol{\alpha}_{k-1}))$ (prediction)
  $\boldsymbol{\alpha}_k = \Phi(\boldsymbol{\alpha}_k^{(p)})$ (correction step using CG)

  $\mathcal{O}\big(K\ell' n^2\big)$

- use only the prediction step!

  $\boldsymbol{\alpha}_k = \boldsymbol{\alpha}_{k-1} + \lambda_k \Psi(\boldsymbol{\alpha}_{k-1})$ (prediction step)
  to do so the regularization path has to be piecewise linear

  $\mathcal{O}\big(Kn^2\big)$

How to tune the regularization parameter λ?

- brute force

  for each $\lambda_1 < \lambda_2 < \ldots < \lambda_k < \ldots < \lambda_K$
  compute $\boldsymbol{\alpha}_k = (K + \lambda_k I)^{-1}\, \mathbf{y}\ , \quad k = 1, K$

  $\mathcal{O}\big(Kn^3\big)$

- warm start

  $\boldsymbol{\alpha}_k = \Phi(\boldsymbol{\alpha}_{k-1})$ (using ℓ conjugate gradient iterations)

  $\mathcal{O}\big(K\ell n^2\big)$

- warm start + prediction step

  $\boldsymbol{\alpha}_k^{(p)} = \boldsymbol{\alpha}_{k-1} + \rho\nabla_{\boldsymbol{\alpha}}\big(L(\boldsymbol{\alpha}_{k-1}) + \lambda_k P(\boldsymbol{\alpha}_{k-1})\big)$ (prediction)
  $\boldsymbol{\alpha}_k = \Phi(\boldsymbol{\alpha}_k^{(p)})$ (correction step using CG)

  $\mathcal{O}\big(K\ell' n^2\big)$

- use only the prediction step!

  $\boldsymbol{\alpha}_k = \boldsymbol{\alpha}_{k-1} + \lambda_k \Psi(\boldsymbol{\alpha}_{k-1})$ (prediction step)
  to do so the regularization path has to be piecewise linear

  $\mathcal{O}\big(Kn^2\big)$

How to tune the regularization parameter λ?

- brute force

  for each $\lambda_1 < \lambda_2 < \ldots < \lambda_k < \ldots < \lambda_K$
  compute $\boldsymbol{\alpha}_k = (K + \lambda_k I)^{-1}\,\mathbf{y}\,, \quad k = 1, K$

  $\mathcal{O}\big(Kn^3\big)$

- warm start

  $\boldsymbol{\alpha}_k = \Phi(\boldsymbol{\alpha}_{k-1})$ (using ℓ conjugate gradient iterations)

  $\mathcal{O}\big(K\ell n^2\big)$

- warm start + prediction step

  $\boldsymbol{\alpha}_k^{(p)} = \boldsymbol{\alpha}_{k-1} + \rho\nabla_{\boldsymbol{\alpha}}(L(\boldsymbol{\alpha}_{k-1}) + \lambda_k P(\boldsymbol{\alpha}_{k-1}))$ (prediction)
  $\boldsymbol{\alpha}_k = \Phi(\boldsymbol{\alpha}_k^{(p)})$ (correction step using CG)

  $\mathcal{O}\big(K\ell' n^2\big)$

- use only the prediction step!

  $\boldsymbol{\alpha}_k = \boldsymbol{\alpha}_{k-1} + \lambda_k \Psi(\boldsymbol{\alpha}_{k-1})$ (prediction step)
  to do so the regularization path has to be piecewise linear

  $\mathcal{O}\big(Kn^2\big)$

How to choose L and P to get linear reg. path?

Solution path is linear $\Leftrightarrow$ one cost is piecewise quadratic and the other one piecewise linear

convex case [Rosset & Zhu, 07]

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^d} L(\boldsymbol{\alpha}) + \lambda P(\boldsymbol{\alpha})$$

1. Piecewise linearity: $\displaystyle\lim_{\varepsilon\to 0} \frac{\boldsymbol{\alpha}(\lambda+\varepsilon) - \boldsymbol{\alpha}(\lambda)}{\varepsilon} = \text{constant}$
2. optimality

$$\nabla L(\boldsymbol{\alpha}(\lambda)) + \lambda \nabla P(\boldsymbol{\alpha}(\lambda)) = 0$$
$$\nabla L(\boldsymbol{\alpha}(\lambda+\varepsilon)) + (\lambda+\varepsilon) \nabla P(\boldsymbol{\alpha}(\lambda+\varepsilon)) = 0$$

3. use Taylor expantion

$$\lim_{\varepsilon\to 0} \frac{\boldsymbol{\alpha}(\lambda+\varepsilon) - \boldsymbol{\alpha}(\lambda)}{\varepsilon} = \left[\nabla^2 L(\boldsymbol{\alpha}(\lambda)) + \lambda \nabla^2 P(\boldsymbol{\alpha}(\lambda))\right]^{-1} \nabla P(\boldsymbol{\alpha}(\lambda))$$

$\nabla^2 L(\boldsymbol{\alpha}(\lambda)) = \textbf{constant}$ **and** $\nabla^2 P(\boldsymbol{\alpha}(\lambda)) = 0$

standart formulation

- portfolio optimization (Markovitz, 1952)
  - return *vs.* risk $\begin{cases} \min_{\boldsymbol{\alpha}} & \frac{1}{2}\boldsymbol{\alpha}^\top Q\boldsymbol{\alpha} \\ \text{with} & \mathbf{e}^\top\boldsymbol{\alpha} = C \end{cases}$
  - *efficiency frontier*: piecewise linear (*Critical path Algo.*)

- Sensitivity analysis: standart formulation (Heller, 1954)

$$\begin{cases} \min_{\boldsymbol{\alpha}} & \frac{1}{2}\boldsymbol{\alpha}^\top Q\boldsymbol{\alpha} + (\mathbf{c} + \lambda\, \Delta\mathbf{c})^\top\boldsymbol{\alpha} \\ \text{with} & A\boldsymbol{\alpha} = \mathbf{b} + \mu\, \Delta\mathbf{b} \end{cases}$$

- Parametric programming (see T. Gal's book 1968)
  - in the general case of PQP: the reg. path is piecewise linear
  - PLP and multi parametric programming

Piecewise linear regularization path algorithms

L	P	*regression*	*classification*	*clustering*
L_2	L_1	Lasso/LARS	L1 L2 SVM	L1 PCA/SVD
L_1	L_2	SVR	SVM	OC SVM
L_1	L_1	L1 LAD	L1 SVM	
		Danzig Selector		

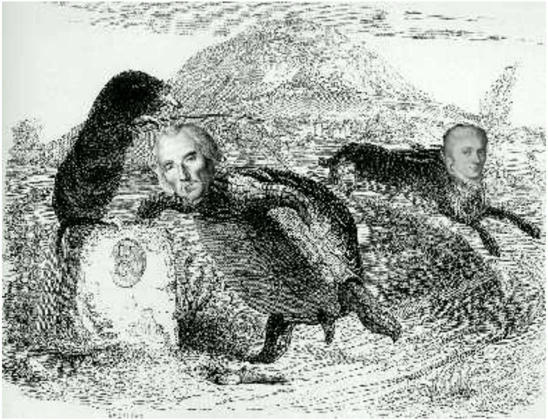
Table: example of piecewise linear regularization path algorithms.

$P: \quad L_p = \sum_{j=1}^{d} |\beta_j|^p$ $\qquad L: \quad L_p : |f(\mathbf{x}) - y|^p \quad \text{hinge } (yf(\mathbf{x}) - 1)_+^p$

$$\varepsilon\text{-insensitive} \qquad \begin{cases} 0 & \text{if } |f(\mathbf{x}) - y| < \varepsilon \\ |f(\mathbf{x}) - y| - \varepsilon & \text{else} \end{cases}$$

$$\text{Huber's loss:} \qquad \begin{cases} |f(\mathbf{x}) - y|^2 & \text{if } |f(\mathbf{x}) - y| < t \\ 2t|f(\mathbf{x}) - y| - t^2 & \text{else} \end{cases}$$

the world is changing



The Gaussian Hare and the Laplacian Tortoise
Computability of L_1 vs. L_2 Regression Estimators.

Portnoy & Koenker 1997

The consequence of having useful reg. path

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^d} L(\boldsymbol{\alpha}) + \lambda P(\boldsymbol{\alpha}) \quad \Leftrightarrow \quad \{\boldsymbol{\alpha}(\lambda) \mid \lambda \in [0, \infty]\}$$

- efficient computing

$\Longrightarrow$ piecewise linearity

$$\boldsymbol{\alpha}^{\text{NEW}} = \boldsymbol{\alpha}^{\text{OLD}} + (\lambda^{\text{NEW}} - \lambda^{\text{OLD}})\mathbf{u}$$

- piecewise linearity

$\Longrightarrow$ either L or P is L_1

- L_1 criteria

$\Longrightarrow$ sparsity: *a lot of* $\boldsymbol{\alpha}_j = 0$

sparsity and active constraints

why does L_1 provide sparsity?

Definition: strong homogeneity set (variables)

$$I_0 = \{ j \in \{1, ..., d\} \mid \alpha_j = 0 \}$$

Theorem

Regular if $L(\alpha) + \lambda P(\alpha)$ differentiable and if $I_0(\mathbf{y}) \neq \emptyset$

$$\forall \varepsilon > 0, \; \exists \mathbf{y}' \in \mathcal{B}(\mathbf{y}, \varepsilon) \text{ such that } I_0(\mathbf{y}') \neq I_0(\mathbf{y})$$

Singular if $L(\alpha) + \lambda P(\alpha)$ NONdifferentiable and if $I_0(\mathbf{y}) \neq \emptyset$

$$\exists \varepsilon > 0, \; \forall \mathbf{y}' \in \mathcal{B}(\mathbf{y}, \varepsilon) \text{ then } I_0(\mathbf{y}') = I_0(\mathbf{y})$$

singular criteria $\Longrightarrow$ sparsity — Nikolova, 2000

L_1 criteria are singular in 0

singularity provides sparsity

L1 Splines: introducing sparsity

The L1 error

$$(\mathcal{S}_1) \quad \begin{cases} \displaystyle\min_{f \in \mathcal{H}} & \frac{1}{2}\|f\|^2_{\mathcal{H}} + \frac{1}{\lambda}\sum_{i=1}^{n} |\xi_i| \\ \text{such that} & f(x_i) = y_i + \xi_i, \qquad i = 1, n \end{cases}$$

representer theorem:

$$f^*(x) = \sum_{i=1}^{n} (\alpha_+ - \alpha_-) k(x, x_i)$$

The dual:

$$(\mathcal{D}_1) \quad \begin{cases} \displaystyle\min_{\alpha_+ - \alpha_-} & \frac{1}{2}(\alpha_+ - \alpha_-)^\top K(\alpha_+ - \alpha_-) + (\alpha_+ + \alpha_-)^\top y \\ \text{such that} & 0 \le \alpha_i^+ \le \frac{1}{\lambda}, 0 \le \alpha_i^- \le \frac{1}{\lambda}, \qquad i = 1, n \end{cases}$$

Typical parametric quadratic program (pQP) but $\alpha_i \neq 0$

K-Lasso (Kernel Basis pursuit)

The Kernel Lasso

$$(\mathcal{S}_1) \quad \left\{ \min_{\boldsymbol{\alpha} \in \mathbb{R}^n} \quad \frac{1}{2}\|K\boldsymbol{\alpha} - \mathbf{y}\|^2 + \lambda \sum_{i=1}^{n} |\alpha_i| \right.$$

- Typical parametric quadratic program (pQP) with $\alpha_i = 0$
- Piecewise linear regularization path

The dual:

$$(\mathcal{D}_1) \quad \left\{ \begin{array}{ll} \min_{\boldsymbol{\alpha}} & \frac{1}{2}\|K\boldsymbol{\alpha}\|^2 \\ \text{such that} & K^\top(K\boldsymbol{\alpha} - \mathbf{y}) \leq t \end{array} \right.$$

- The K-Danzig selector can be treated the same way
- require to compute $K^\top K$ - no more function f!

Support vector regression (SVR)

Lasso's dual adaptation:

$$\begin{cases} \min\limits_{\boldsymbol{\alpha}} & \frac{1}{2}\|K\boldsymbol{\alpha}\|^2 \\ \text{s. t.} & K^\top(K\boldsymbol{\alpha} - \mathbf{y}) \leq t \end{cases} \qquad \begin{cases} \min\limits_{f \in \mathcal{H}} & \frac{1}{2}\|f\|^2_{\mathcal{H}} \\ \text{s. t.} & |f(\mathbf{x}_i) - y_i| \leq t, \quad i = 1, n \end{cases}$$

The support vector regression introduce slack variables

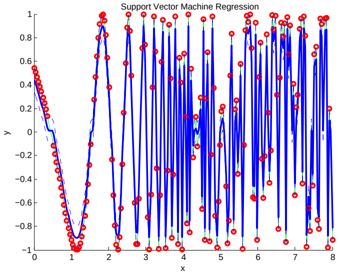
$$(SVR) \quad \begin{cases} \min\limits_{f \in \mathcal{H}} & \frac{1}{2}\|f\|^2_{\mathcal{H}} + C\sum|\xi_i| \\ \text{such that} & |f(\mathbf{x}_i) - y_i| \leq t + \xi_i \quad 0 \leq \xi_i \quad i = 1, n \end{cases}$$

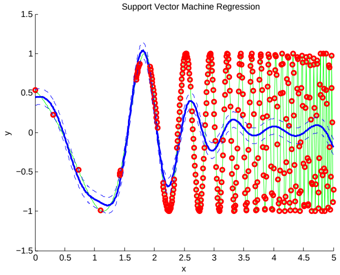
- a typical multi parametric quadratic program (mpQP)
- piecewise linear regularization path

$$\alpha(C, t) = \alpha(C_0, t_0) + (\frac{1}{C} - \frac{1}{C_0})\mathbf{u} + \frac{1}{C_0}(t - t_0)\mathbf{v}$$

- 2d Pareto's front (the tube width and the regularity)

Support vector regression illustration





C large | C small

- there exists other formulations such as LP SVR...

Large scale pQP

Not yet adapted

- reweighed LS
- interior points
- projected gradient

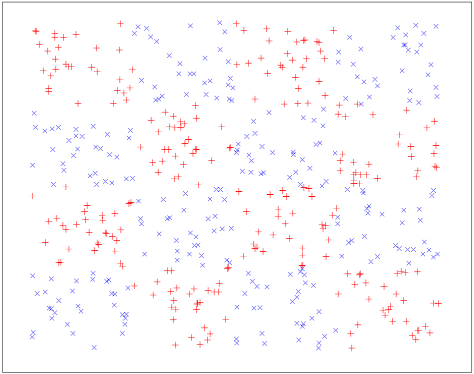
Adapted

- homotopy (regularization path) and other pQP
- active set
- decomposition
- coordinate wise (Gauss Seidel)

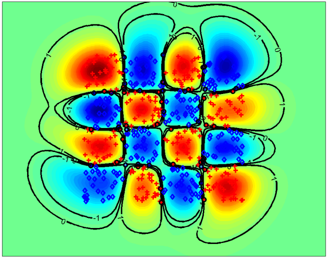
other: cutting plane, proximal...

checker board

- 2 classes
- 500 examples
- separable

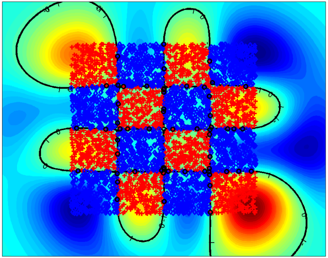


a separable case

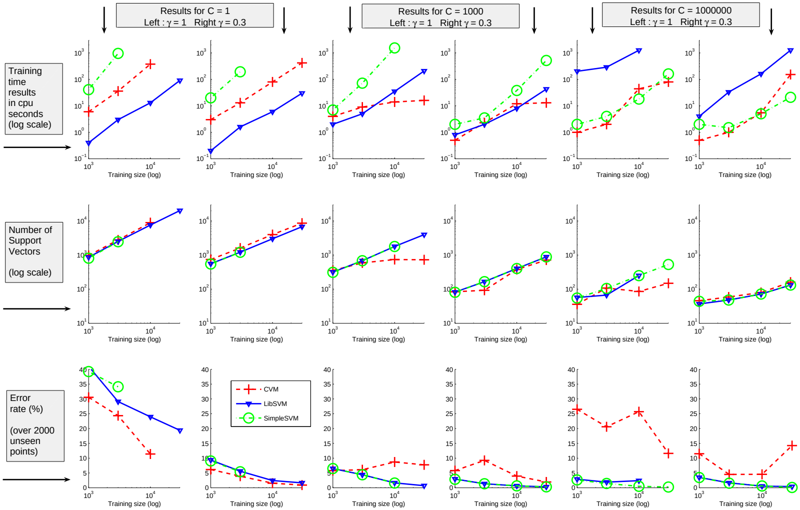


$n = 500$ data points

$n = 5000$ data points



empirical complexity



Plan

Multiple Kernel

The model

$$f(x) = \sum_{i=1}^{\ell} \alpha_i K(x, x_i) + b,$$

Given M kernel functions $K_1, \ldots, K_M$ that are potentially well suited for a given problem, find a positive linear combination of these kernels such that the resutling kernel K is “optimal”

$$K(\mathbf{x}, \mathbf{x}') = \sum_{m=1}^{M} d_m K_m(\mathbf{x}, \mathbf{x}'), \text{ with } d_m \geq 0, \sum_m d_m = 1$$

Need to learn together the kernel coefficients d_m and the SVR parameters α_i, b.

Multiple Kernel functional Learning

The problem (for given C and t)

$$\min_{\{f_m\},b,\xi,d} \quad \frac{1}{2}\sum_m \frac{1}{d_m}\|f_m\|^2_{\mathcal{H}_m} + C\sum_i \xi_i$$

$$\text{s.t.} \quad \Big|\sum_m f_m(x_i) + b - y_i\Big| \leq t + \xi_i \quad \forall i \xi_i \geq 0 \quad \forall i$$

$$\sum_m d_m = 1\ , \quad d_m \geq 0 \quad \forall m\ ,$$

regularization formulation

$$\min_{\{f_m\},b,d} \quad \frac{1}{2}\sum_m \frac{1}{d_m}\|f_m\|^2_{\mathcal{H}_m} + C\sum_i \max\Big(\Big|\sum_m f_m(x_i) + b - y_i\Big| - t, 0\Big)$$

$$\sum_m d_m = 1\ , \quad d_m \geq 0 \quad \forall m\ ,$$

Equivalently

$$\min_{\{f_m\},b,\xi,d} \quad \sum_i \max\Big(\Big|\sum_m f_m(x_i) + b - y_i\Big| - t, 0\Big) + \frac{1}{2C}\sum_m \frac{1}{d_m}\|f_m\|^2_{\mathcal{H}_m} + \mu\sum_m |d_m|$$

Multiple Kernel functional Learning

The problem (for given C and t)

$$\begin{aligned} \min_{\{f_m\},b,\xi,d} \quad & \frac{1}{2}\sum_m \frac{1}{d_m}\|f_m\|^2_{\mathcal{H}_m} + C\sum_i \xi_i \\ \text{s.t.} \quad & \Big|\sum_m f_m(x_i) + b - y_i\Big| \leq t + \xi_i \quad \forall i \xi_i \geq 0 \quad \forall i \\ & \sum_m d_m = 1 \;, \quad d_m \geq 0 \quad \forall m \;, \end{aligned}$$

Treated as a bi-level optimization task

$$\begin{aligned} \min_{d \in \mathbb{R}^M} \quad & \left\{ \begin{aligned} \min_{\{f_m\},b,\xi} \quad & \frac{1}{2}\sum_m \frac{1}{d_m}\|f_m\|^2_{\mathcal{H}_m} + C\sum_i \xi_i \\ \text{s.t.} \quad & \Big|\sum_m f_m(x_i) + b - y_i\Big| \geq t + \xi_i \quad \forall i \\ & \xi_i \geq 0 \quad \forall i \end{aligned} \right. \\ \text{s.t.} \quad & \sum_m d_m = 1 \;, \quad d_m \geq 0 \quad \forall m \;, \end{aligned}$$

Multiple Kernel Algorithm

Use a Reduced Gradient Algorithm[2]

$$\min_{d \in \mathbb{R}^M} \quad J(d)$$
$$\text{s.t.} \quad \sum_m d_m = 1 \ , \quad d_m \geq 0 \quad \forall m \ ,$$

SimpleMKL algorithm

set $d_m = \frac{1}{M}$ for $m = 1, \ldots, M$
while stopping criterion not met **do**
  compute $J(d)$ using an QP solver with $K = \sum_m d_m K_m$
  compute $\frac{\partial J}{\partial d_m}$, Hessian and descent direction D
  $\gamma \leftarrow$ compute optimal stepsize
  $d \leftarrow d + \gamma D$
end while

$\longrightarrow$ Recent improvement reported using the Hessian

[2] Rakotomamonjy et al. JMLR 08

Complexity

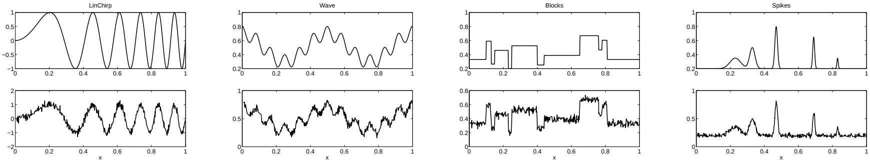
For each iteration:

- SVM training: $O(nn_{sv} + n_{sv}^3)$.
- Inverting $K_{sv,sv}$ is $O(n_{sv}^3)$, but might already be available as a by-product of the SVM training.
- Computing H: $O(Mn_{sv}^2)$
- Finding d: $O(M^3)$.

The number of iterations is usually less than 10.

$\longrightarrow$ When $M < n_{sv}$, computing d is not more expensive than QP.

Multiple Kernel experiments



	Single Kernel	Kernel *Dil*		Kernel *Dil-Trans*	
Data Set	Norm. MSE (%)	#Kernel	Norm. MSE	#Kernel	Norm. MSE
LinChirp	1.46 ± 0.28	7.0	1.00 ± 0.15	21.5	0.92 ± 0.20
Wave	0.98 ± 0.06	5.5	0.73 ± 0.10	20.6	0.79 ± 0.07
Blocks	1.96 ± 0.14	6.0	2.11 ± 0.12	19.4	1.94 ± 0.13
Spike	6.85 ± 0.68	6.1	6.97 ± 0.84	12.8	5.58 ± 0.84

Table: Normalized Mean Square error averaged over 20 runs.

Conclusion

- Kernels
- sparsity L_1
- efficent algorithm
- some limits
  - instability
  - large data sets
  - when to stop
  - non convexity

Perspectives

- more algorithms, more criteria, more applications

Conclusion

- Kernels
- sparsity L_1
- efficent algorithm
- some limits
  - instability — coupling with active set
  - large data sets — randomize
  - when to stop — derive relevant bounds
  - non convexity — iterative L1

Perspectives

- more algorithms, more criteria, more applications

Questions?

questions?

LAR(s) (and svmpath) in R — `www-stat.stanford.edu/~hastie`
Matlab — `asi.insa-rouen.fr/~vguigue/LARS.html`

kernlab — `cran.r-project.org/src/contrib/Descriptions/kernlab.html`
SVR Matlab — `asi.insa-rouen.fr/~arakotom`

Danzig selector — `www.l1-magic.org/`

biblio

- Least Angle Regression. Least angle regression, Bradley Efron, ; Annals of statistics, vol. 32 (2), pp.407-499, 2004
- Piecewise Linear Regularized Solution Paths, Saharon Rosset, Ji Zhu. Annals of Statistics, to appear, 2007
- Local strong homogeneity of a regularized estimator, Mila Nikolova, SIAM Journal on Applied Mathematics, vol. 61, no. 2, pp. 633-658, 2000.
- Two-Dimensional Solution Path for Support Vector Regression, Gang Wang, Dit-Yan Yeung, Frederick H. Lochovsky, ICML, 2006
- Algorithmic linear dimension reduction in the l_1 norm for sparse vectors, A. C. Gilbert, M. J. Strauss, J. A. Tropp, and R. Vershynin, 2006 (submitted).
- A tutorial on support vector regression, A. Smola and B. Scholkopf, Statistics and Computing 14(3), pp 199-222, 2004
- The Dantzig selector: Statistical estimation when p is much smaller than n, Emmanuel Candes and Terence Tao Submitted to IEEE Transactions on Information Theory, June 2005.
- An iterative thresholding algorithm for linear inverse problems with a sparsity constraint, I. Daubechies, M. Defrise and C. De Mol, Comm. Pure Appl. Math, 57, pp.1413-1541, 2004.
- Pathwise coordinate optimization, Jerome Friedman, Trevor Hastie and Robert Tibshirani, technical report, May 2007.