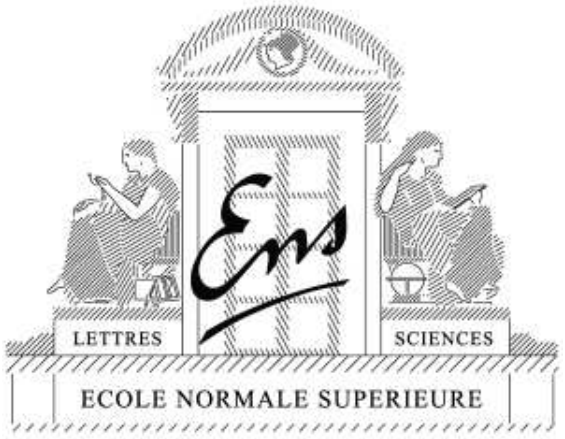
Multiple Kernel Learning

Francis Bach

Willow project, INRIA - Ecole Normale Supérieure, Paris



In collaboration with M. Jordan, R. Thibaux (U.C. Berkeley), G. Lanckriet (U.C. San Diego), A. Rakotomamonjy, S. Canu (INSA Rouen), Y. Grandvalet (UTC), Z. Harchaoui (Telecom Paris)

August 2008

Multiple kernel learning (MKL) - Outline

- Regularization for supervised learning
  - Kernel methods
- Multiple kernel learning framework
  - Formulation as non parametric group Lasso
  - Algorithms
  - Analysis of consistency of kernel selection
- Applications to computer vision

Supervised learning and regularization

- Data: $x_i \in \mathcal{X}$, $y_i \in \mathcal{Y}$, $i = 1, \ldots, n$
- Minimize with respect to function $f \in \mathcal{F}$:

$$\sum_{i=1}^{n} \ell(y_i, f(x_i)) \quad + \quad \frac{\lambda}{2}\|f\|^2$$

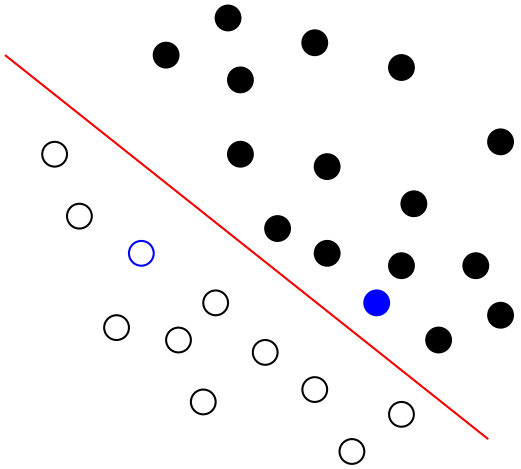
Error on data + Regularization

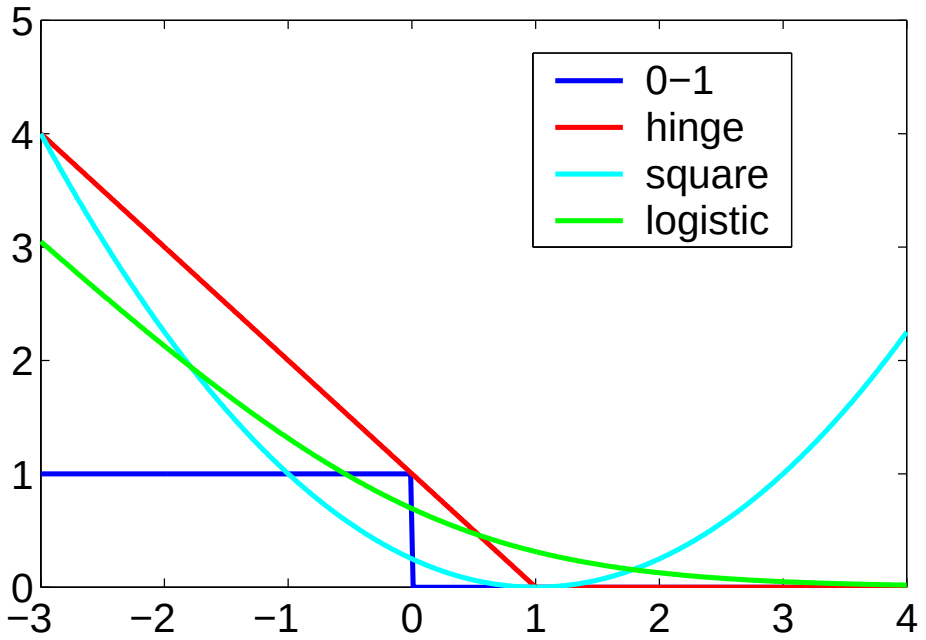
Loss & function space ? Norm ?

- Two issues:
  - Loss
  - Function space / norm

Usual losses

- **Regression**: $y \in \mathbb{R}$, prediction $\hat{y} = f(x)$,
  - quadratic cost $\ell(y, f(x)) = \frac{1}{2}(y - f(x))^2$
- **Classification** : $y \in \{-1, 1\}$ prediction $\hat{y} = \text{sign}(f(x))$
  - loss of the form $\ell(y, f(x)) = \ell(yf(x))$
  - "True" cost: $\ell(yf(x)) = 1_{yf(x)<0}$
  - Usual convex costs:





Regularizations

- Main goal: control the "capacity" of the learning problem
- Two main lines of work
  1. Use Hilbertian (RKHS) norms
     - Non parametric supervised learning and kernel methods
     - Well developped theory
  2. Use "sparsity inducing" norms
     - main example: ℓ_1-norm
     - Perform model selection as well as regularization
     - Often used heuristically
- **Group lasso / MKL : two types of regularizations**

Reproducing kernel Hilbert spaces

- Assume k is a positive definite kernel on $\mathcal{X} \times \mathcal{X}$
- **Aronszajn theorem** (1950): there exists a Hilbert space $\mathcal{F}$ and a mapping $\Phi : \mathcal{X} \mapsto \mathcal{F}$ such that

$$\forall (x, x') \in \mathcal{X}^2, \; k(x, x') = \langle \Phi(x), \Phi(x') \rangle_{\mathcal{H}}$$

- $\mathcal{X} =$ "input space", $\mathcal{F} =$ "feature space", $\Phi =$ "feature map"
- RKHS: particular instantiation of $\mathcal{F}$ as a function space
  - $\Phi(x) = k(\cdot, x)$
  - function evaluation $f(x) = \langle f, \Phi(x) \rangle$
  - reproducing property: $k(x, y) = \langle k(\cdot, x), k(\cdot, y) \rangle$
- Notations : $f(x) = \langle f, \Phi(x) \rangle = f^\top \Phi(x)$, $\|f\|^2 = \langle f, f \rangle$

Regularization and representer theorem

- Data: $x_i \in \mathbb{R}^d$, $y_i \in \mathcal{Y}$, $i = 1, \ldots, n$, kernel k (with RKHS $\mathcal{F}$)
- Minimize with respect to f: $\boxed{\min_{f \in \mathcal{F}} \sum_{i=1}^n \ell(y_i, f^\top \Phi(x_i)) + \frac{\lambda}{2}\|f\|^2}$
- No assumptions on cost ℓ or n
- **Representer theorem** (Kimeldorf, Wahba, 1971): Optimum is reached for weights of the form
$$f = \sum_{j=1}^n \alpha_j \Phi(x_j) = \sum_{j=1}^n \alpha_j k(\cdot, x_j)$$
- $\alpha \in \mathbb{R}^n$ dual parameters, $K \in \mathbb{R}^{n \times n}$ kernel matrix:
$$K_{ij} = \Phi(x_i)^\top \Phi(x_j) = k(x_i, x_j)$$
- Equivalent problem: $\boxed{\min_{\alpha \in \mathbb{R}^n} \sum_{i=1}^n \ell(y_i, (K\alpha)_i) + \frac{\lambda}{2}\alpha^\top K \alpha}$

Kernel trick and modularity

- Kernel trick: any algorithm for finite-dimensional vectors that only uses pairwise dot-products can be applied in the feature space.
  - Replacing dot-products by kernel functions
  - Implicit use of (very) large feature spaces
  - Linear to non-linear learning methods

Kernel trick and modularity

- Kernel trick: any algorithm for finite-dimensional vectors that only uses pairwise dot-products can be applied in the feature space.
  - Replacing dot-products by kernel functions
  - Implicit use of (very) large feature spaces
  - Linear to non-linear learning methods

- Modularity of kernel methods
  1. Work on new algorithms and theoretical analysis
  2. Work on new kernels for specific data types

Representer theorem and convex duality

- The parameters $\alpha \in \mathbb{R}^n$ may also be interpreted as Lagrange multipliers
- Assumption: cost function is convex $\varphi_i(u_i) = \ell(y_i, u_i)$
- Primal problem: $\boxed{\min_{f \in \mathcal{F}} \sum_{i=1}^n \varphi_i(f^\top \Phi(x_i)) + \frac{\lambda}{2}\|f\|^2}$

	$\varphi_i(u_i)$
LS regression	$\frac{1}{2}(y_i - u_i)^2$
Logistic regression	$\log(1 + \exp(-y_i u_i))$
SVM	$(1 - y_i u_i)_+$

Representer theorem and convex duality

- Assumption: cost function is convex $\varphi_i(u_i) = \ell(y_i, u_i)$
- Primal problem: $\boxed{\min_{f \in \mathcal{F}} \sum_{i=1}^n \varphi_i(f^\top \Phi(x_i)) + \frac{\lambda}{2}\|f\|^2}$
- Dual problem: $\boxed{\max_{\alpha \in \mathbb{R}^n} -\sum_{i=1}^n \psi_i(-\lambda\alpha_i) - \frac{\lambda}{2}\alpha^\top K \alpha}$

  where $\psi_i(v_i) = \max_{u_i \in \mathbb{R}} v_i u_i - \varphi_i(u_i)$ is the Fenchel conjugate of φ_i
- Strong duality
- Relationship between primal and dual variables (at optimum):

$$f = \sum_{i=1}^n \alpha_i \Phi(x_i)$$

"Classical" kernel learning (2-norm regularization)

Primal problem $\min_{f \in \mathcal{F}} \left(\sum_i \varphi_i(f^\top \Phi(x_i)) + \frac{\lambda}{2} ||f||^2 \right)$

Dual problem $\max_{\alpha \in \mathbb{R}^n} \left(- \sum_i \psi_i(\lambda \alpha_i) - \frac{\lambda}{2} \alpha^\top K \alpha \right)$

Optimality conditions $f = - \sum_{i=1}^n \alpha_i \Phi(x_i)$

- Assumptions on loss φ_i:
  - $\varphi_i(u)$ convex
  - $\psi_i(v)$ Fenchel conjugate of $\varphi_i(u)$, i.e., $\psi_i(v) = \max_{u \in \mathbb{R}} (vu - \varphi_i(u))$

	$\varphi_i(u_i)$	$\psi_i(v)$
LS regression	$\frac{1}{2}(y_i - u_i)^2$	$\frac{1}{2}v^2 + vy_i$
Logistic regression	$\log(1 + \exp(-y_i u_i))$	$(1+vy_i)\log(1+vy_i)$ $-vy_i \log(-vy_i)$
SVM	$(1 - y_i u_i)_+$	$-vy_i \times 1_{-vy_i \in [0,1]}$

Kernel learning with convex optimization

- Kernel methods work...

  ...with the good kernel!

  $\Rightarrow$ Why not learn the kernel directly from data?

Kernel learning with convex optimization

- Kernel methods work...

  ...with the good kernel!

  $\Rightarrow$ Why not learn the kernel directly from data?

- **Proposition** (Lanckriet et al, 2004, Bach et al, 2004):

$$\begin{aligned} G(K) &= \min_{f \in \mathcal{F}} \sum_{i=1}^{n} \varphi_i(f^\top \Phi(x_i)) + \frac{\lambda}{2}\|f\|^2 \\ &= \max_{\alpha \in \mathbb{R}^n} -\sum_{i=1}^{n} \psi_i(\lambda \alpha_i) - \frac{\lambda}{2}\alpha^\top K \alpha \end{aligned}$$

  is a convex function of the Gram matrix K

- Theoretical learning bounds (Lanckriet et al, 2004)

MKL framework

- Minimize with respect to the kernel matrix K

$$G(K) = \max_{\alpha \in \mathbb{R}^n} -\sum_{i=1}^n \psi_i(\lambda \alpha_i) - \frac{\lambda}{2} \alpha^\top K \alpha$$

- Optimization domain:
  - K positive semi-definite in general
  - The set of kernel matrices is a cone $\longrightarrow$ conic representation

$$\boxed{K(\eta) = \sum_{j=1}^m \eta_j K_j, \quad \eta \geqslant 0}$$

  - Trace constraints: $\operatorname{tr} K = \sum_{j=1}^m \eta_j \operatorname{tr} K_j = 1$

- Optimization:
  - In most cases, representation in terms of SDP, QCQP or SOCP
  - Optimization by generic toolbox is costly (Lanckriet et al., 2004)

MKL - "reinterpretation" (Bach et al, 2004)

- Framework limited to $K = \sum_{j=1}^{m} \eta_j K_j$, $\eta \geqslant 0$

- Summing kernels is equivalent to concatenating feature spaces
  - m "feature maps" $\Phi_j : \mathcal{X} \mapsto \mathcal{F}_j$, $j = 1, \ldots, m$.
  - Minimization with respect to $f_1 \in \mathcal{F}_1, \ldots, f_m \in \mathcal{F}_m$
  - Predictor: $f(x) = f_1^\top \Phi_1(x) + \cdots + f_m^\top \Phi_m(x)$

$$
\begin{array}{ccccc}
 & \nearrow & \Phi_1(x)^\top & f_1 & \searrow \\
 & \vdots & \vdots & \vdots & \vdots \\
x & \longrightarrow & \Phi_j(x)^\top & f_j & \longrightarrow \; f_1^\top \Phi_1(x) + \cdots + f_m^\top \Phi_m(x) \\
 & \vdots & \vdots & \vdots & \vdots \\
 & \searrow & \Phi_m(x)^\top & f_m & \nearrow
\end{array}
$$

  - **Which regularization?**

Regularization for multiple kernels

- Summing kernels is equivalent to concatenating feature spaces
  - m "feature maps" $\Phi_j : \mathcal{X} \mapsto \mathcal{F}_j$, $j = 1, \ldots, m$.
  - Minimization with respect to $f_1 \in \mathcal{F}_1, \ldots, f_m \in \mathcal{F}_m$
  - Predictor: $f(x) = f_1^\top \Phi_1(x) + \cdots + f_m^\top \Phi_m(x)$
- Regularization by $\sum_{j=1}^m \|f_j\|^2$ is equivalent to using $K = \sum_{j=1}^m K_j$

Regularization for multiple kernels

- Summing kernels is equivalent to concatenating feature spaces
  - m "feature maps" $\Phi_j : \mathcal{X} \mapsto \mathcal{F}_j$, $j = 1, \ldots, m$.
  - Minimization with respect to $f_1 \in \mathcal{F}_1, \ldots, f_m \in \mathcal{F}_m$
  - Predictor: $f(x) = f_1^\top \Phi_1(x) + \cdots + f_m^\top \Phi_m(x)$
- Regularization by $\sum_{j=1}^m \|f_j\|^2$ is equivalent to using $K = \sum_{j=1}^m K_j$
- Regularization by $\sum_{j=1}^m \|f_j\|$ should impose sparsity at the group level
- **Main questions when regularizing by block ℓ^1-norm**:
  1. Equivalence with previous formulations
  2. Algorithms
  3. Analysis of sparsity inducing properties

MKL - duality (Bach et al, 2004)

- Primal problem:

$$\sum_{i=1}^n \varphi_i(f_1^\top \Phi_1(x_i) + \cdots + f_m^\top \Phi_m(x_i)) + \frac{\lambda}{2}\left(\|f_1\| + \cdots + \|f_m\|\right)^2$$

- **Proposition**: Dual problem (using second order cones)

$$\max_{\alpha \in \mathbb{R}^n} -\sum_{i=1}^n \psi_i(-\lambda\alpha_i) - \frac{\lambda}{2}\min_{j\in\{1,\ldots,m\}} \alpha^\top K_j \alpha$$

KKT conditions: $f_j = \eta_j \sum_{i=1}^n \alpha_i \Phi_j(x_i)$
with $\alpha \in \mathbb{R}^n$ and $\eta \geqslant 0$, $\sum_{j=1}^m \eta_j = 1$

- α is the dual solution for the clasical kernel learning problem with kernel matrix $K(\eta) = \sum_{j=1}^m \eta_j K_j$
- η corresponds to the minimum of $G(K(\eta))$

Analysis of sparsity inducing property

- Optimization problem:

$$\sum_{i=1}^{n} \varphi_i(f_1^\top \Phi_1(x_i) + \cdots + f_m^\top \Phi_m(x_i)) + \frac{\lambda}{2} \left(d_1 \|f_1\| + \cdots + d_m \|f_m\| \right)^2$$

- Sparsity inducing effect used heuristically
- Extension of the group Lasso (i.e., finite dimensional Hilbert spaces, square loss, Yuan & Lin, 2006)
- Denote $\hat{f} = (\hat{f}_1, \ldots, \hat{f}_m)$ the solution
- Consistent estimation of the sparsity pattern $J = \{j, \hat{f}_j \neq 0\}$?

Group lasso - Asymptotic analysis
Groups of finite sizes - Square loss

- Assumptions:
  1. Data (X_i, Y_i) sampled **i.i.d.**
  2. $\mathbf{w} \in \mathbb{R}^p$ denotes the (unique) minimizer of $\mathbb{E}(Y - X^\top w)^2$ (best linear predictor). Assume $\mathbb{E}((Y - \mathbf{w}^\top X)^2 | X) \geqslant \sigma_{\min}^2 > 0$ $a.s.$
  3. Finite fourth order moments: $\mathbb{E}\|X\|^4 < \infty$ and $\mathbb{E}\|Y\|^4 < \infty$.
  4. Invertible covariance: $\Sigma_{XX} = \mathbb{E}XX^\top \in \mathbb{R}^{p \times p}$ is invertible.

- Denote $\mathbf{J} = \{j, \mathbf{w}_j \neq 0\}$ the sparsity pattern of $\mathbf{w}$

- Goal: estimate consistently both $\mathbf{w}$ and $\mathbf{J}$ when n tends to infinity
  - $\forall \varepsilon > 0$, $\mathbb{P}(\|\hat{w} - \mathbf{w}\| > \varepsilon)$ tends to zero
  - $\mathbb{P}(\{j, \hat{w}_j \neq 0\} \neq \mathbf{J})$ tends to zero
  - Rates of convergence

Group lasso - Consistency conditions (Bach, 2008)

- Strict condition:

$$\max_{i \in \mathbf{J}^c} \frac{1}{d_i} \left\| \Sigma_{X_i X_\mathbf{J}} \Sigma_{X_\mathbf{J} X_\mathbf{J}}^{-1} \operatorname{Diag}(d_j / \|\mathbf{w}_j\|) \mathbf{w}_\mathbf{J} \right\| < 1$$

- Weak condition:

$$\max_{i \in \mathbf{J}^c} \frac{1}{d_i} \left\| \Sigma_{X_i X_\mathbf{J}} \Sigma_{X_\mathbf{J} X_\mathbf{J}}^{-1} \operatorname{Diag}(d_j / \|\mathbf{w}_j\|) \mathbf{w}_\mathbf{J} \right\| \leqslant 1$$

- **Theorem 1**: Strict condition is sufficient for joint regular and sparsity consistency of the group lasso ($\lambda_n / n \to 0$ and $\lambda_n n^{-1/2} \to +\infty$)

- **Theorem 2**: Weak condition is necessary for joint regular and sparsity consistency of the group lasso (for any λ_n).

Group lasso - Consistency conditions

- Condition: $\max_{i \in \mathbf{J}^c} \frac{1}{d_i} \left\| \Sigma_{X_i X_\mathbf{J}} \Sigma_{X_\mathbf{J} X_\mathbf{J}}^{-1} \operatorname{Diag}(d_j / \|\mathbf{w}_j\|) \mathbf{w}_\mathbf{J} \right\| < \text{ or } \leqslant 1$
- Extension of the Lasso consistency conditions (Zhao and Yu, 2006, Yuan and Lin, 2007, Zou, 2006, Wainwright, 2006)
- Extension to non parametric case with cross-covariance operators
- Additional questions:
  - Is strict condition necessary (as in the Lasso case)?
  - Estimate of probability of correct sparsity estimation
  - Loading independent condition
  - Other losses
  - *Negative or positive result?*

Positive or negative result?

- "Disappointing" result for Lasso/group Lasso
  - Does not always do what heuristic justification suggests!
- Can we make it always consistent?
  - Data dependent weights $\Rightarrow$ adaptive Lasso/group Lasso
- Do we care about exact sparsity consistency?
  - Recent results by Meinshausen and Yu (2007)

Importance of weights - Adaptive group lasso

- Normalization of data is subject to arbitrary choices
- Important
  - In theory: consistency condition not always satisfied

$$\max_{i \in \mathbf{J}^c} \frac{1}{d_i} \left\| \Sigma_{X_i X_\mathbf{J}} \Sigma_{X_\mathbf{J} X_\mathbf{J}}^{-1} \operatorname{Diag}(d_j / \|\mathbf{w}_j\|) \mathbf{w}_\mathbf{J} \right\| < \text{ or } \leqslant 1$$

  - In practice (Bach, Thibaux, Jordan, 2005): normalizing kernel matrices to unit trace may lead to bad predictive performance.
- Adaptive weights based on the data
  - Independent of y: depends on the rank of the kernel matrix
  - Dependent of y through simple least-square estimation: two-step adaptive group Lasso

Adaptive group lasso

- **Theorem**: Let $\hat{f}^{LS}$ be the least-square estimate with regularization parameter proportional to $n^{-1/3}$. Let $\hat{f}$ denote any minimizer of

$$\frac{1}{2n}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{m} f_j^\top \Phi_j(x_i)\right)^2 + \frac{\mu_0 n^{-1/3}}{2}\left(\sum_{j=1}^{m}\|\hat{f}_j^{LS}\|^{-\gamma}\|f_j\|\right)^2.$$

  For any $\gamma > 1$, $\hat{f}$ converges to $\mathbf{f}$ and $J(\hat{f})$ converges to $\mathbf{J}$ in probability.

- Convergence rates with more assumptions (and more work!)

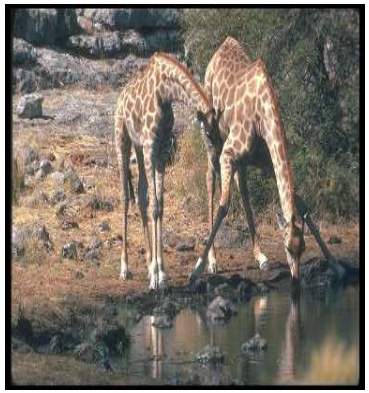
- Practical implications in applications to be determined

Applications

- Bioinformatics (Lanckriet et al., 2004)
- Image annotation (Harchaoui & Bach, 2007)

Image annotation and kernel design

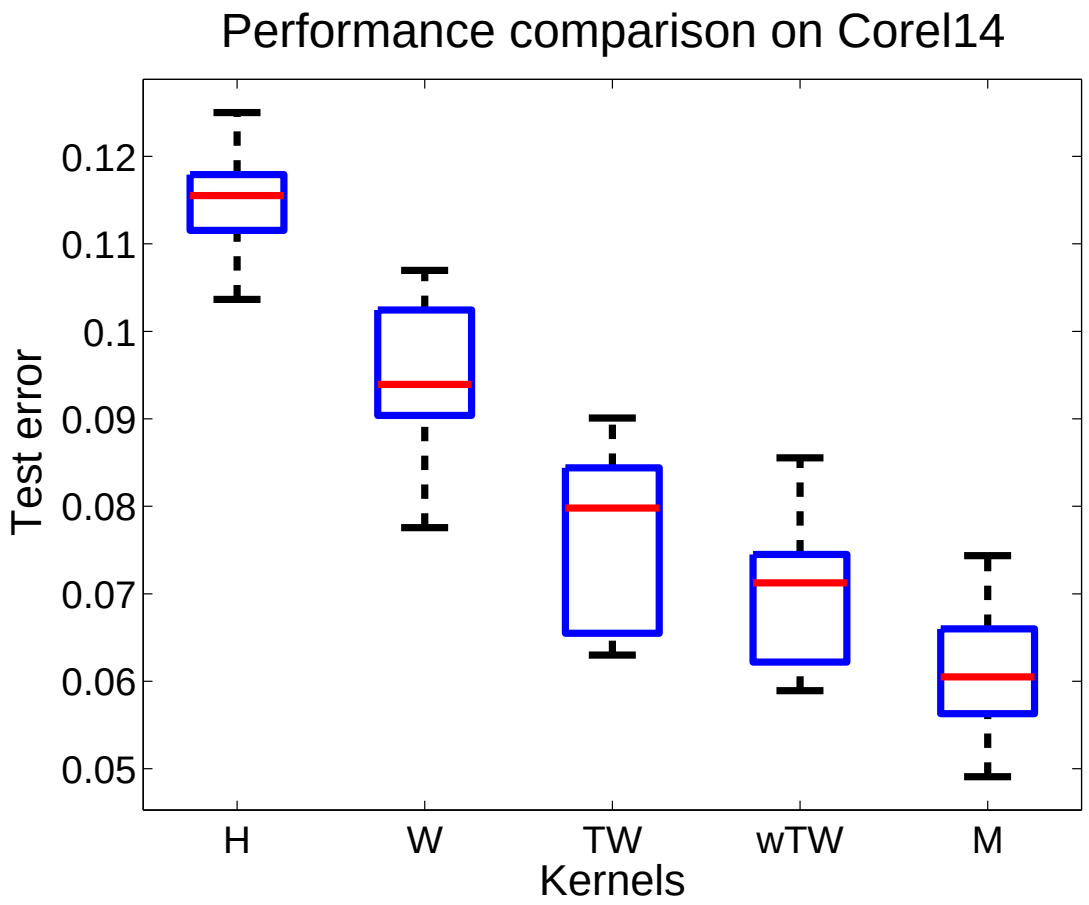
- `Corel14`: 1400 *natural images* with 14 classes






Performance on Corel14 (Harchaoui & Bach, 2007)

- Histogram kernels (**H**)
- Walk kernels (**W**)
- Tree-walk kernels (**TW**)
- Weighted tree-walks (**wTW**)
- MKL (**M**)



Conclusion

- Multiple kernel learning (MKL) for learning the kernel from data
  - Efficient algorithms
  - Theoretical analysis of kernel selection property
  - Applications to heterogeneous data fusion and hyperparameter selection

- Current research directions
  - Negative combination of kernel matrices?
  - Application to non linear variable selection
  - Analysis of other sparsity inducing norms (e.g., trace norm for matrices)