Consistency of group lasso and multiple kernel learning

Francis Bach

INRIA - Ecole Normale Supérieure

Willow project

November 2007

Summary

- Machine learning and regularization
- Group Lasso
  - Consistent estimation of groups?
- Multiple kernel learning as non parametric group Lasso
- Extension to trace norm minimization

Supervised learning and regularization

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

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

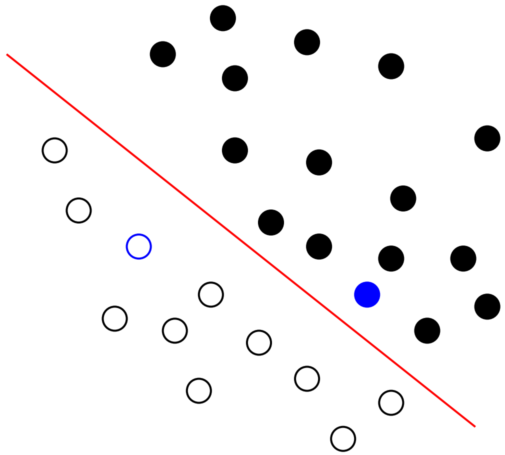
Error on data $\qquad + \qquad$ Regularization

Loss & function space ? $\qquad$ Norm ?

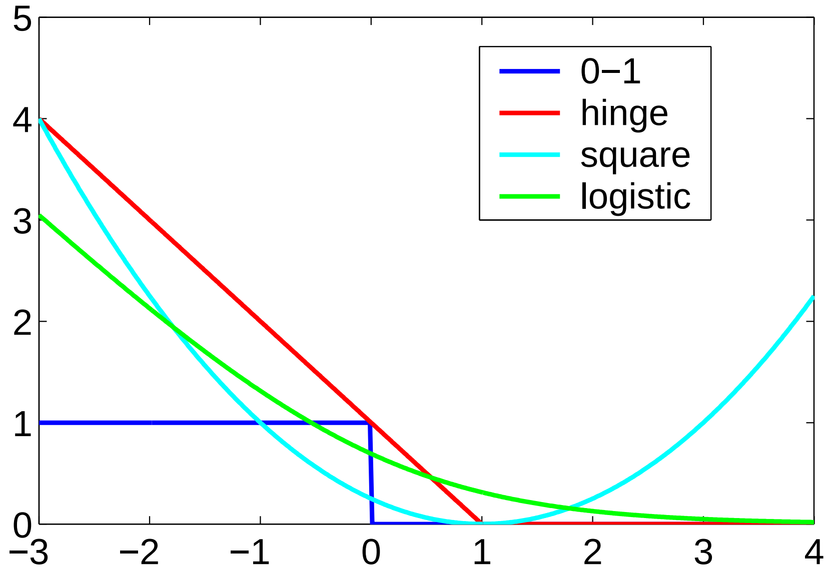
- Two issues:
  - Loss
  - Function space / norm

Usual losses

- **Regression**: $y \in \mathbb{R}$, prediction $\hat{y} = f(x)$, quadratic cost $\ell(y, f) = \frac{1}{2}(y - \hat{y})^2 = \frac{1}{2}(y - f)^2$



- **Classification** : $y \in \{-1, 1\}$ prediction $\hat{y} = \text{sign}(f(x))$
  - loss of the form $\ell(y, f) = \ell(yf)$
  - "True" cost: $\ell(yf) = 1_{yf<0}$
  - 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**

Group lasso - linear predictors

- Assume $x_i, w \in \mathbb{R}^p$ where $p = p_1 + \cdots + p_m$, i.e., m **groups**

$$x_i = (x_{i1}, \ldots, x_{im}) \qquad w = (w_1, \ldots, w_m)$$

- Goal: achieve sparsity at the levels of groups: $J(w) = \{i, \ w_i \neq 0\}$

- Main application:
  - Group selection vs. variable selection (Zhao et al., 2006)
  - Multi-task learning (Argyriou et al., 2006, Obozinsky et al., 2007)

- Regularization by block ℓ_1-norm (Yuan & Lin, 2006, Zhao et al., 2006, Bach et al., 2004):

$$\min_{w \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, w^\top x_i) + \lambda \sum d_j \|w_j\|$$

Group lasso - Main questions

$$\min_{w \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, w^\top x_i) + \lambda \sum d_j \|w_j\|$$

1. Analysis of sparsity inducing property:

   - where do $\hat{w}$ and $J(\hat{w}) = \{i,\ \hat{w}_i \neq 0\}$ converge to?
   - letting the problem grow
     - sizes of the groups $p_i, i = 1, \dots, m \Rightarrow$ **"kernelization"**
     - number of groups $m \Rightarrow$?
   - Influence of the weights d_j

2. Algorithms

   - very efficient and elegant for the Lasso (Efron et al., 2004)

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: $\Sigma_{XX} = \mathbb{E} X X^\top \in \mathbb{R}^{p \times p}$ 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

- 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 ($\lambda_n \to 0$ 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)

- 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?*

Group lasso - Strict condition necessary?

- Strict condition necessary for the Lasso (Zou, 2006, Zhao and Yu, 2006)

- Strict condition not necessary for the group Lasso
  - If weak condition is satisfied and for all $i \in \mathbf{J}^c$ such that $\frac{1}{d_i} \left\| \Sigma_{X_i X_\mathbf{J}} \Sigma_{X_\mathbf{J} X_\mathbf{J}}^{-1} \operatorname{Diag}(d_j / \|\mathbf{w}_j\|) \mathbf{w}_\mathbf{J} \right\| = 1$, we have

$$\Delta^\top \Sigma_{X_\mathbf{J} X_i} \Sigma_{X_i X_\mathbf{J}} \Sigma_{X_\mathbf{J} X_\mathbf{J}}^{-1} \operatorname{Diag}\left[d_j / \|\mathbf{w}_j\| \left(I_{p_j} - \frac{\mathbf{w}_j \mathbf{w}_j^\top}{\mathbf{w}_j^\top \mathbf{w}_j} \right) \right] \Delta > 0,$$

    with $\Delta = -\Sigma_{X_\mathbf{J} X_\mathbf{J}}^{-1} \operatorname{Diag}(d_j / \|\mathbf{w}_j\|) \mathbf{w}_\mathbf{J}$, then the group lasso estimate leads to joint regular and sparsity consistency ($\lambda_n \to 0$ and $\lambda_n n^{1/4} \to +\infty$)

Loading independent sufficient condition

- Condition on Σ and $\mathbf{J}$:

$$
\begin{aligned}
\max_{w_{\mathbf{J}}} \max_{i\in\mathbf{J}^c} \frac{1}{d_i} \left\| \Sigma_{X_i X_{\mathbf{J}}} \Sigma_{X_{\mathbf{J}} X_{\mathbf{J}}}^{-1} \operatorname{Diag}(d_j / \|\mathbf{w}_j\|) \mathbf{w}_{\mathbf{J}} \right\| &< 1 \\
\Leftrightarrow \quad \max_{i\in\mathbf{J}^c} \frac{1}{d_i} \max_{\|u_j\|=1,\ \forall j\in\mathbf{J}} \left\| \Sigma_{X_i X_{\mathbf{J}}} \Sigma_{X_{\mathbf{J}} X_{\mathbf{J}}}^{-1} \operatorname{Diag}(d_j) \mathbf{u}_{\mathbf{J}} \right\| &< 1 \\
\Rightarrow \quad \max_{i\in\mathbf{J}^c} \frac{1}{d_i} \sum_{j\in\mathbf{J}} d_j \left\| \sum_{k\in\mathbf{J}} \Sigma_{X_i X_k} \left(\Sigma_{X_{\mathbf{J}} X_{\mathbf{J}}}^{-1} \right)_{kj} \right\| &< 1
\end{aligned}
$$

- Lasso (groups of size 1): all those are equivalent

- Group lasso: stricter sufficient condition (in general)
  - NB: can obtain better one with convex relaxation (see paper)

Probability of correct selection of pattern

- Simple general result when $\lambda_n = \lambda_0 n^{-1/2}$
- Probability equal to

$$\mathbb{P}\left(\max_{i \in \mathbf{J}^c} \left\| \frac{\sigma}{n^{1/2}\lambda_n d_i} \Sigma_{X_i X_\mathbf{J}} \Sigma_{X_\mathbf{J} X_\mathbf{J}}^{-1/2} \mathbf{u} - \frac{1}{d_i} \Sigma_{X_i X_\mathbf{J}} \Sigma_{X_\mathbf{J} X_\mathbf{J}}^{-1} \operatorname{Diag}\left(\frac{d_j}{\|\mathbf{w}_j\|}\right) \mathbf{w}_\mathbf{J} \right\| \leqslant 1 \right)$$

  where **u** is normal with mean zero and identity covariance matrix.

- With additional conditions, valid when $\lambda_n n^{1/2}$ not too far from constant $\Rightarrow$ exponential rate of convergence if strict condition is satisfied
- Dependence on σ and n

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)

Relationship with multiple kernel learning (MKL) (Bach, Lanckriet, Jordan, 2004)

- Alternative equivalent formulation:

$$\min_{w \in \mathbb{R}^p} \tfrac{1}{2n}\|\bar{Y} - \bar{X}w\|^2 + \tfrac{1}{2}\mu_n \left(\textstyle\sum_{j=1}^m d_j \|w_j\| \right)^2$$

- Dual optimization problem (using conic programming):

$$\max_{\alpha \in \mathbb{R}^n} \left\{ -\tfrac{1}{2n}\|\bar{Y} - n\mu_n \alpha\|^2 - \tfrac{1}{2\mu_n} \max_{i=1,\dots,m} \frac{\alpha^\top K_i \alpha}{d_i^2} \right\}$$

Relationship with multiple kernel learning (MKL) (Bach, Lanckriet, Jordan, 2004)

- Alternative equivalent formulation:

$$\min_{w \in \mathbb{R}^p} \frac{1}{2n} \|\bar{Y} - \bar{X}w\|^2 + \frac{1}{2}\mu_n \left(\sum_{j=1}^m d_j \|w_j\| \right)^2$$

- Dual optimization problem (using conic programming):

$$\max_{\alpha \in \mathbb{R}^n} \left\{ -\frac{1}{2n} \|\bar{Y} - n\mu_n \alpha\|^2 - \frac{1}{2\mu_n} \max_{i=1,\dots,m} \frac{\alpha^\top K_i \alpha}{d_i^2} \right\}$$

$$\Leftrightarrow \max_{\alpha \in \mathbb{R}^n} \min_{\eta \geqslant 0, \sum_{j=1}^m \eta_j d_j^2 = 1} \left\{ -\frac{1}{2n} \|\bar{Y} - n\mu_n \alpha\|^2 - \frac{1}{2\mu_n} \alpha^\top \left(\sum_{j=1}^m \eta_i K_i \right) \alpha \right\}$$

Relationship with multiple kernel learning (MKL) (Bach, Lanckriet, Jordan, 2004)

- Alternative equivalent formulation:

$$\min_{w \in \mathbb{R}^p} \ \tfrac{1}{2n} \|\bar{Y} - \bar{X} w\|^2 + \tfrac{1}{2} \mu_n \left(\textstyle\sum_{j=1}^m d_j \|w_j\| \right)^2$$

- Dual optimization problem:

$$\max_{\alpha \in \mathbb{R}^n} \left\{ -\tfrac{1}{2n} \|\bar{Y} - n\mu_n \alpha\|^2 - \tfrac{1}{2\mu_n} \max_{i=1,\dots,m} \frac{\alpha^\top K_i \alpha}{d_i^2} \right\}$$

$$\Leftrightarrow \min_{\eta \geqslant 0, \sum_{j=1}^m \eta_j d_j^2 = 1} \ \max_{\alpha \in \mathbb{R}^n} \left\{ -\tfrac{1}{2n} \|\bar{Y} - n\mu_n \alpha\|^2 - \tfrac{1}{2\mu_n} \alpha^\top \left(\sum_{j=1}^m \eta_i K_i \right) \alpha \right\}$$

Relationship with multiple kernel learning (MKL)

$$\min_{\eta \geqslant 0, \sum_{j=1}^m \eta_j d_j^2 = 1} \max_{\alpha \in \mathbb{R}^n} \left\{ -\tfrac{1}{2n} \|\bar{Y} - n\mu_n \alpha\|^2 - \tfrac{1}{2\mu_n} \alpha^\top \left(\sum_{j=1}^m \eta_i K_i \right) \alpha \right\}$$

- Optimality conditions: the dual variable $\alpha \in \mathbb{R}^n$ is optimal if and only if there exists $\eta \in \mathbb{R}_+^m$ such that $\sum_{j=1}^m \eta_j d_j^2 = 1$ and α is optimal for ridge regression problem with kernel matrix $K = \sum_{j=1}^m \eta_j K_j$

- η can also be obtained as the minimizer of

$$J(\eta) = \max_{\alpha \in \mathbb{R}^n} -\frac{1}{2n} \|\bar{Y} - n\mu_n \alpha\|^2 - \frac{1}{2\mu_n} \alpha^\top \left(\sum_{j=1}^m \eta_j K_j \right) \alpha,$$

  – $J(\eta)$ is the optimal value of the objective function of the single kernel estimation problem with kernel $K = \sum_{j=1}^m \eta_j K_j$

Multiple kernel learning (MKL)

- Jointly learn optimal (sparse) combination of kernel (η) together with the estimate with this kernel (α)

- Application
  - Kernel learning
  - Heteregeneous data fusion

- Known issues
  - Algorithms
  - Influence of weights d_j (feature spaces have different sizes)
  - Consistency

Analysis of MKL as non parametric group Lasso

- Assume m Hilbert spaces $\mathcal{F}_i$, $i = 1, \dots, m$

$$\min_{f_i \in \mathcal{F}_i,\ i=1,\dots,m} \quad \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{m} f_j(x_{ji}) \right)^2 + \frac{\mu_n}{2} \left(\sum_{j=1}^{m} d_j \|f_j\| \right)^2 .$$

- Sparse generalized additive models (Hastie and Tibshirani, 1990)

- Estimate is obtained through MKL formulation

- Same question: regular and sparsity consistency when the groups are infinite-dimensional Hilbert spaces

Analysis of MKL as non parametric group Lasso (non centered) covariance operators

- Single random variable X: Σ_{XX} is a bounded linear operator from $\mathcal{F}$ to $\mathcal{F}$ such that for all $(f,g) \in \mathcal{F} \times \mathcal{F}$,

$$\langle f, \Sigma_{XX} g \rangle = \mathbb{E}(f(X)g(X))$$

  Under minor assumptions, the operator Σ_{XX} is *auto-adjoint*, *non-negative* and *Hilbert-Schmidt*

- Tool of choice for the analysis of least-squares non parametric methods (Fukumizu et al., 2005, 2006, Gretton et al., 2006, etc...)
  - Natural empirical estimate $\hat{\Sigma}_{XX} = \frac{1}{n}\sum_{i=1}^{n} k(\cdot, x_i) \otimes k(\cdot, x_i)$ converges in probability to Σ_{XX} in HS norm.

Cross-covariance operators

- Several random variables: cross-covariance operators $\Sigma_{X_i X_j}$ from $\mathcal{F}_j$ to $\mathcal{F}_i$ such that $\forall (f_i, f_j) \in \mathcal{F}_i \times \mathcal{F}_j$,

$$\langle f_i, \Sigma_{X_i X_j} f_j \rangle = \mathbb{E}(f_i(X_i) f_j(X_j))$$

- Similar convergence properties of empirical estimates

- Joint covariance operator Σ_{XX} defined by blocks

- We can define the bounded *correlation* operators through

$$\Sigma_{X_i X_j} = \Sigma_{X_i X_i}^{1/2} C_{X_i X_j} \Sigma_{X_j X_j}^{1/2}$$

- NB: the joint covariance operator is never invertible, but the correlation operator may be

Analysis of MKL as non parametric group Lasso

- Assumptions

1. $\forall j$, $\mathcal{F}_j$ is a separable RKHS associated with kernel k_j, and $\mathbb{E}k_j(X_j, X_j)^2 < \infty$.
2. **Model**: There exists functions $\mathbf{f} = (\mathbf{f}_1, \ldots, \mathbf{f}_m) \in \mathcal{F} = \mathcal{F}_1 \times \cdots \times \mathcal{F}_m$ and a function $\mathbf{h}$ of $X = (X_1, \ldots, X_m)$ such that

$$\mathbb{E}(Y|X) = \sum_{j=1}^{m} \mathbf{f}_j(X_j) + \mathbf{h}(X)$$

   with $\mathbb{E}h(X)^2 < \infty$ and $\mathbb{E}\mathbf{h}(X)f_j(X_j) = 0$ for all $j = 1, \ldots, m$ and $\mathbb{E}((Y - \sum_{j=1}^{m} \mathbf{f}_j(X_j))^2|X) \geqslant \sigma^2_{\min} > 0 \; a.s.$
3. **Compacity and invertibility** : All cross-correlation operators are compact and the joint correlation operator is invertible.
4. **Range condition**: For all j, $\exists \mathbf{g}_j \in \mathcal{F}_j$ such that $\mathbf{f}_j = \Sigma^{1/2}_{X_j X_j} \mathbf{g}_j$

Compacity and invertibility of joint correlation operator

- Sufficient condition for compacity when distributions have densities:

$$\mathbb{E}\left\{\frac{p_{X_iX_j}(x_i,x_j)}{p_{X_i}(x_i)p_{X_j}(x_j)} - 1\right\} < \infty.$$

  – Dependence between variables is not too strong

- Sufficient condition for invertibility: no exact correlation using functions in the RKHS.

  – Empty *concurvity* space assumption (Hastie and Tibshirani, 1990)

Range condition

- Technical condition: For all j, $\exists \mathbf{g}_j \in \mathcal{F}_j$ such that $\mathbf{f}_j = \Sigma_{X_j X_j}^{1/2} \mathbf{g}_j$
  - Conditions on the support of f_j with respect to the support of the data
  - Conditions on the smoothness of f_j

- Sufficient condition for translation invariant kernels

$$k(x, x') = q(x - x') \text{ in } \mathbb{R}^d:$$

  - f_j is of the form $f_j = q * g_j$ where $\int \frac{g_j^2(x_j)}{p_{X_j}(x_j)} dx_j$.

Group lasso - Consistency conditions

- Strict condition

$$\max_{i \in \mathbf{J}^c} \frac{1}{d_i} \left\| \Sigma_{X_i X_i}^{1/2} C_{X_i X_\mathbf{J}} C_{X_\mathbf{J} X_\mathbf{J}}^{-1} \operatorname{Diag}(d_j / \|\mathbf{f}_j\|) \mathbf{g}_\mathbf{J} \right\| < 1$$

- Weak condition

$$\max_{i \in \mathbf{J}^c} \frac{1}{d_i} \left\| \Sigma_{X_i X_i}^{1/2} C_{X_i X_\mathbf{J}} C_{X_\mathbf{J} X_\mathbf{J}}^{-1} \operatorname{Diag}(d_j / \|\mathbf{f}_j\|) \mathbf{g}_\mathbf{J} \right\| \leqslant 1$$

- **Theorem 1**: Strict condition is sufficient for joint regular and sparsity consistency of the lasso.

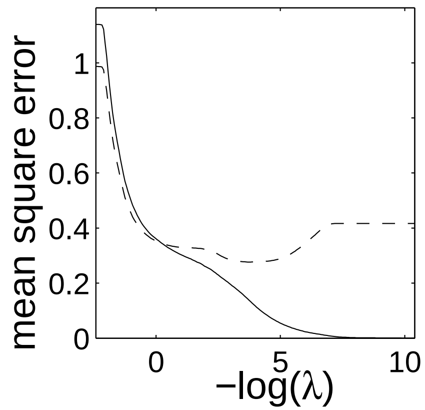
- **Theorem 2**: Weak condition is necessary for joint regular and sparsity consistency of the lasso.

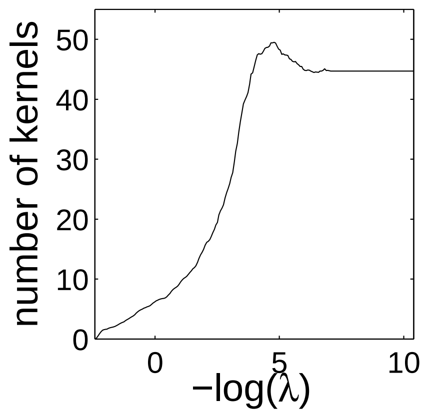
Adaptive group lasso

- Consistency condition depends on $\mathbf{w}$ or $\mathbf{f}$ and is not always satisfied!

- Empirically, the weights do matter a lot (Bach, Thibaux, Jordan, 2005)

Importance of weigts (Bach, Thibaux, Jordan, 2005)





- Canonical behavior as λ decreases
  - Training error decreases to zero
  - Testing error decreases, increases, then stabilizes

- Importance of d_j (weight of penalization $= \sum_j d_j ||w_j||$)
  - d_j should be an increasing function of the "rank" of K_j, e.g., (when matrices are normalized to unit trace):

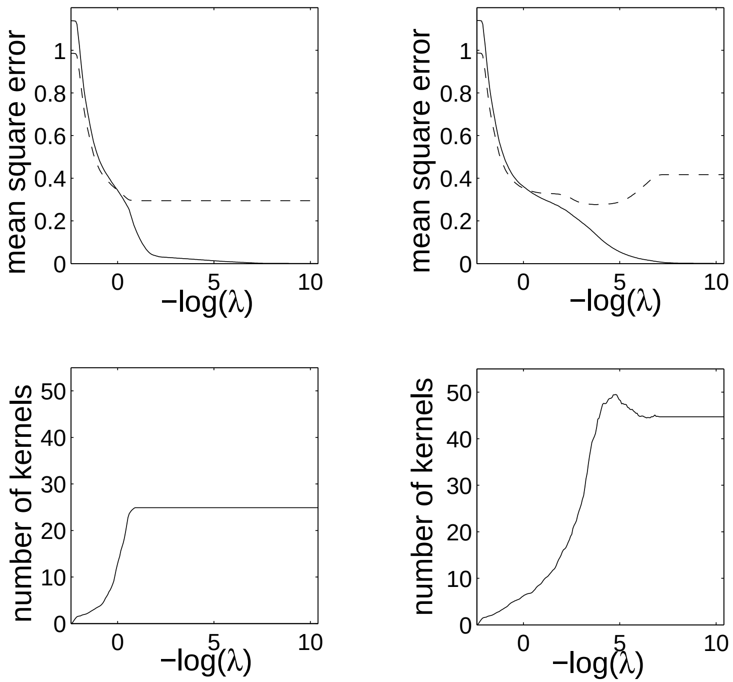
$$d_j = \left(\text{number of eigenvalue } \geqslant \frac{1}{2n} \right)^{\gamma}$$

Importance of weigts (Bach, Thibaux, Jordan, 2005)

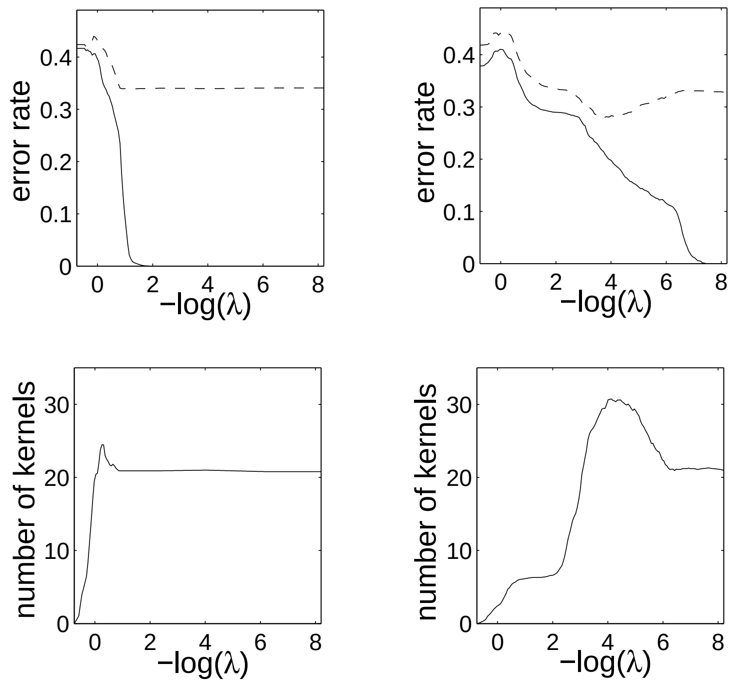
- Left: $\gamma = 0$ (unit trace, Lanckriet et al., 2004), right: $\gamma = 1$

- Top: training (bold)/testing (dashed) error
  bottom: number of kernels

Regression (Boston dataset)



Classification (Liver dataset)



Adaptive group lasso

- Consistency condition depends on $\mathbf{w}$ or $\mathbf{f}$ and is not always satisfied!
- Empirically, the weights do matter a lot (Bach, Thibaux, Jordan, 2005)
- Extension of the Lasso adaptive version (Yuan & Lin, 2006) using the regularized LS estimate $\hat{f}^{LS}_{\kappa_n}$ defined as:

$$\min_{f_i \in \mathcal{F}_i,\ i=1,\ldots,m} \quad \frac{1}{2n}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{m} f_j(x_{ji})\right)^2 + \frac{\kappa_n}{2}\sum_{j=1}^{m}\|f_j\|^2,$$

Adaptive group lasso

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

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

  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

Algorithms for group lasso and MKL

- Algorithms for general convex losses

- many different interpretations implies many different algorithms
  - Group Lasso - primal formulation w.r.t. w
  - Group Lasso - dual formulation w.r.t. α
  - Direct problem involving η

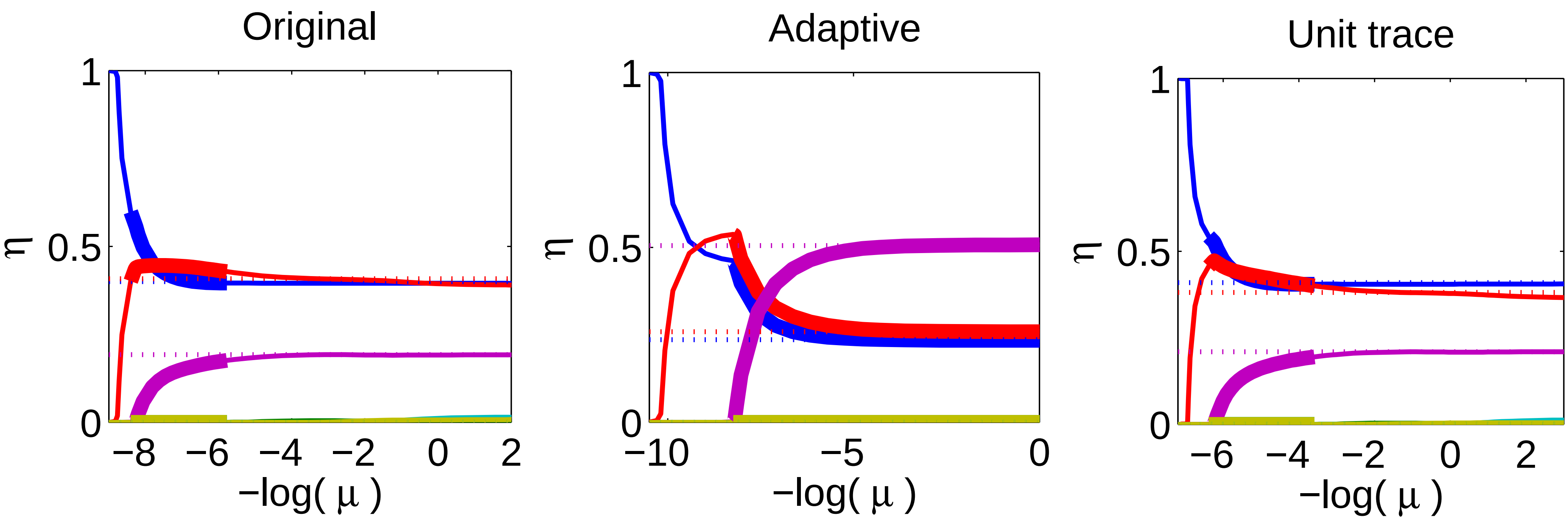
Algorithms for MKL

- (very) costly optimization with SDP, QCQP ou SOCP (Lanckriet et al., 2004)
  - $n \geqslant 1,000 - 10,000$, $m \geqslant 100$ not possible
  - "loose" required precision $\Rightarrow$ first order methods
- Shooting algorithm (Yuan & Lin, 2006)
- Dual coordinate ascent (SMO) with smoothing (Bach et al., 2004)
- Optimization of $J(\eta)$ by cutting planes (Sönnenburg et al., 2005)
- Optimization of $J(\eta)$ with steepest descent with smoothing (Rakotomamonjy et al, 2007)
- Regularization path (Bach, Thibaux & Jordan, 2005)

Illustrative toy experiments

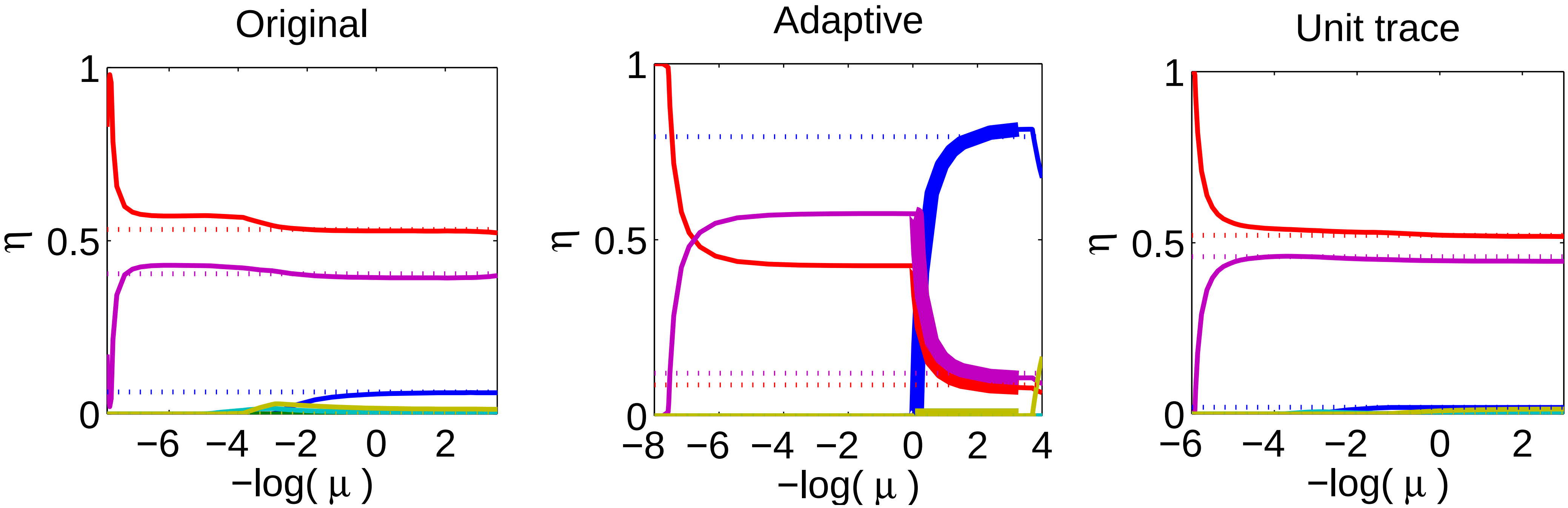
- 6 groups of size 2 - Card$(\mathbf{J}) = 3$

- Consistent condition **fullfilled**:



Illustrative toy experiments

- 6 groups of size 2 - $\mathrm{Card}(\mathbf{J}) = 3$
- Consistent condition **not fullfilled**:

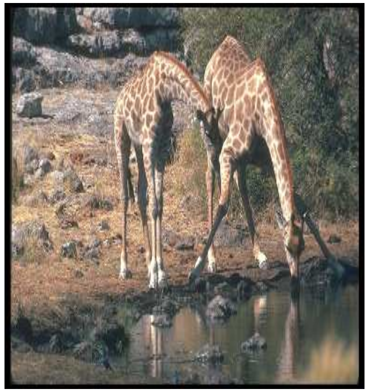


Applications

- Bioinformatics (Lanckriet et al., 2004)
  - Protein function prediction
  - ...
- Image annotation (Harchaoui & Bach, 2007)
  - Fusing information from different aspects of images

Image annotation

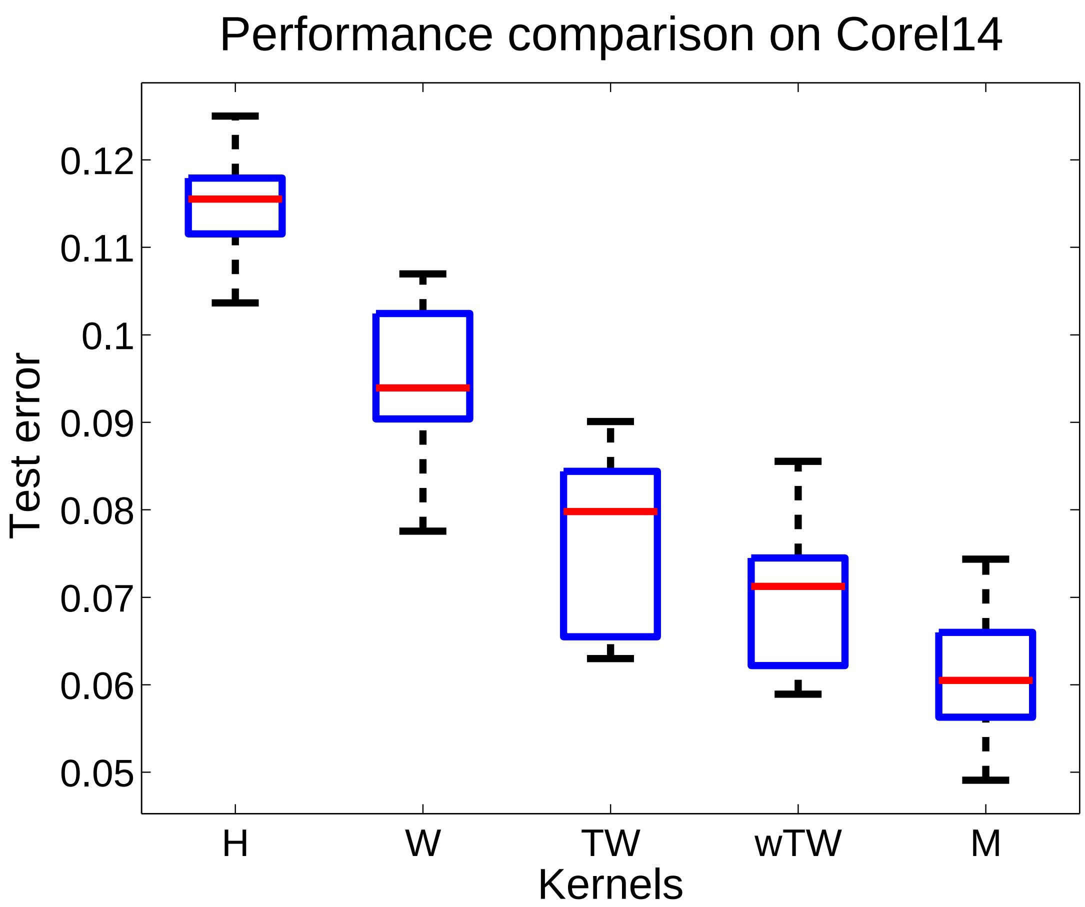
- `Corel14`: 1400 *natural images* with 14 classes






Performance on Corel14
(Harchaoui & Bach, 2007)

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



Extension to trace norm minimization

- Consider learning linear predictor where covariates X are rectangular matrices
- loading matrix W, and prediction $\operatorname{tr} W^\top X$
- Assumption of low rank loading matrix:
  - Matrix completion (Srebro et al., 2004)
  - collaborative filtering (Srebro et al., 2004, Abernethy et al., 2006)
  - Multi-task learning (Argyriou et al., 2006, Obozinsky et al., 2007)
- Equivalent of the ℓ_1 norm : **trace norm = sums of singular values**
- Do we actually get low-rank solutions?
  - Necessary and sufficient consistency conditions (Bach, 2007)
  - Extension of the group Lasso results.

Conclusion

- Analysis of sparsity behavior of the group lasso
  - infinite dimensional groups $\Rightarrow$ MKL
  - Adaptive version to define appropriate weights

- Current work:
  - Analysis for other losses
  - Consider growing number of groups
  - Analysis when consistency condition not satisfied
  - non parametric group lasso: universal consistency?
  - Infinite dimensional extensions of trace norm minimization