Sparsity by Worst-Case Quadratic Penalties

Yves Grandvalet

Heudiasyc, CNRS & Université de Technologie de Compiègne

Julien Chiquet **Christophe Ambroise**

Statistique et Génome, CNRS & Université d'Évry Val d'Essonne

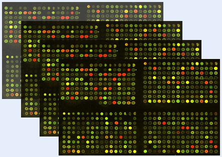
arXiv preprint

http://arxiv.org/abs/1210.2077

R-package **quadrupen**, on CRAN

Variable Selection in Bioinformatics
Microarrays

signal processing →



← pretreatment

Data Matrix $n \times p$, $n \ll p$

$n \simeq 100$, $p \simeq 10\,000$

Expression levels of p probes monitored for n patients

$$\mathbf{X} = \begin{pmatrix} x_{11} & \dots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & \dots & x_{np} \end{pmatrix}$$

⇝ Models for microarray data bet on:

- Sparsity
- Structural correlation between variables

Variable Selection in Bioinformatics

Standard Solutions

1. Univariate analysis and select effects via multiple testing
   $\rightsquigarrow$ Genomic data are often highly correlated...
2. Combine multivariate analysis and model selection techniques

$$\arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} -\mathcal{L}(\boldsymbol{\beta};\mathbf{y},\mathbf{X}) + \lambda \left\|\boldsymbol{\beta}\right\|_0$$

$\rightsquigarrow$ NP-hard in general (exact solutions only for $p < 30$)

More Recent Ideas

Use a convex relaxation of the multivariate problem

$$\arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} -\mathcal{L}(\boldsymbol{\beta};\mathbf{y},\mathbf{X}) + \lambda \left\|\boldsymbol{\beta}\right\|_1$$

...or more fancy penalties to account for structure

Contributions

1. We suggest a unifying view of sparsity-inducing penalties
   - may provide insights on these methods
     - as an interpretation: robust optimization, Bayesian framework?
     - as way to derive generic results

       ⇝ monitoring of convergence
   - results in a generic algorithm for computing solutions
2. The associated algorithm relies on solving linear systems is
   - accurate
   - efficient up to medium scale problems (thousands of variables)

     ⇝ speeds up (double) cross-validation, bootstrap/subsampling methods

     ⇝ model selection

     ⇝ stabilization

     ⇝ permutation tests

Outline

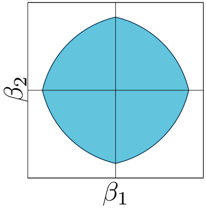
- Motivations
- Going Quadratic
  - The Variational Way
  - The Duality Way
- Benefits
  - Generality
  - Algorithm
  - Analysis
- Experiments
- Conclusion

The Variational Way

Going quadratic: solving problems amount to solve systems

Elastic-Net Example

$$\begin{cases} \min_{\boldsymbol{\beta} \in \mathbb{R}^p} \|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2 \\ \text{s. t. } \frac{1}{2}\|\boldsymbol{\beta}\|_2^2 + \eta\|\boldsymbol{\beta}\|_1 \leq s \end{cases}$$

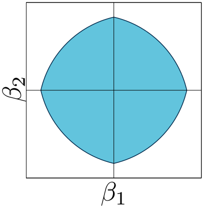


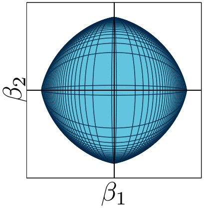
The Variational Way

Going quadratic: solving problems amount to solve systems

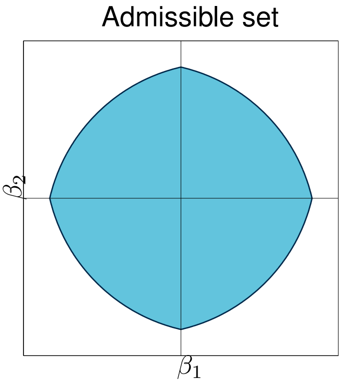
Elastic-Net Example

$$\begin{cases} \min_{\boldsymbol{\beta}\in\mathbb{R}^p} \|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2 \\ \text{s. t. } \frac{1}{2}\|\boldsymbol{\beta}\|_2^2 + \eta\|\boldsymbol{\beta}\|_1 \leq s \end{cases} \Leftrightarrow \begin{cases} \min_{\boldsymbol{\beta}\in\mathbb{R}^p} \|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2 \\ \text{s. t. } \min_{\boldsymbol{\tau}\in\mathbb{R}^p} \sum_{j=1}^{p} \left(\frac{1}{2} + \frac{\eta}{\tau_j}\right) {\beta_j}^2 \leq s \\ \qquad \|\boldsymbol{\tau}\|_1 - \|\boldsymbol{\beta}\|_1 \leq 0\,, \quad \tau_j \geq 0 \end{cases}$$



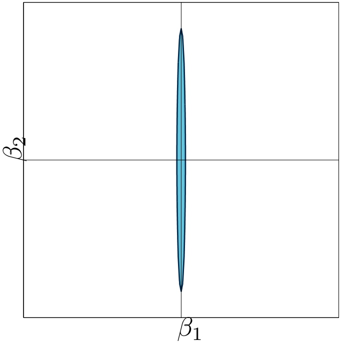


The Variational Way
Building the Admissible Set

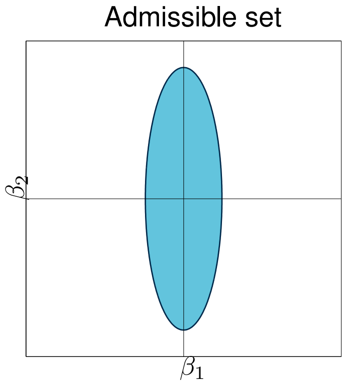


The Variational Way
Building the Admissible Set

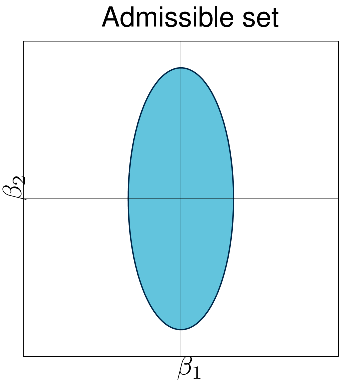
Admissible set



The Variational Way
Building the Admissible Set

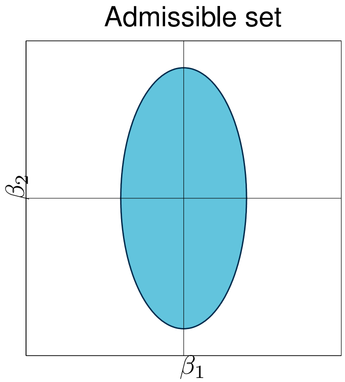


The Variational Way
Building the Admissible Set



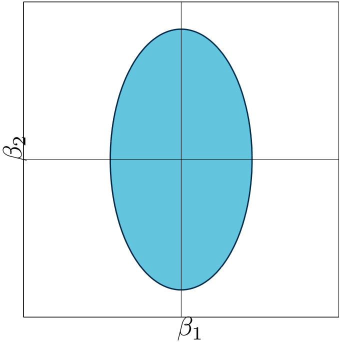
The Variational Way

Building the Admissible Set



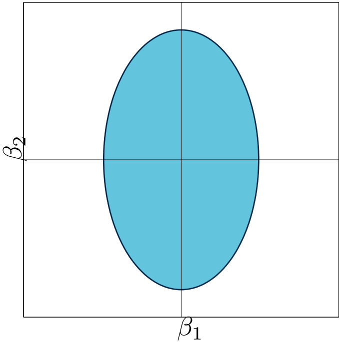
The Variational Way
Building the Admissible Set

Admissible set

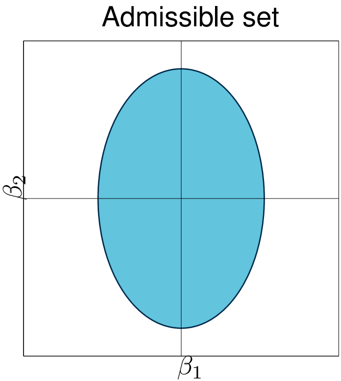


The Variational Way
Building the Admissible Set

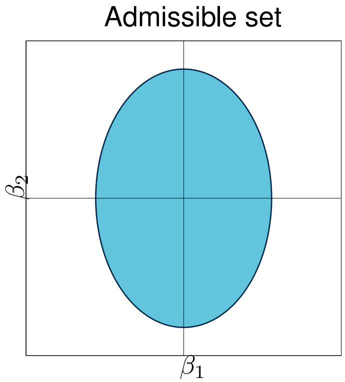
Admissible set



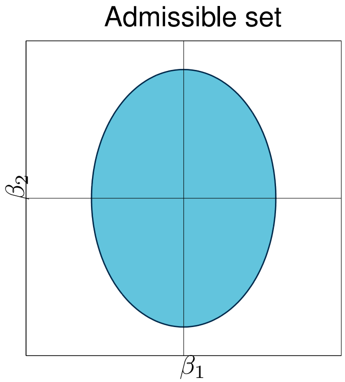
The Variational Way
Building the Admissible Set



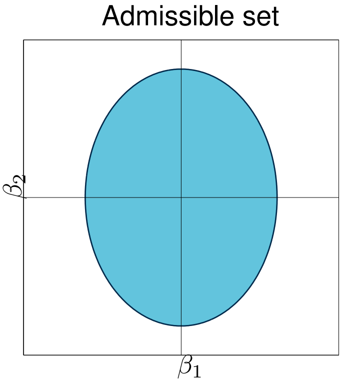
The Variational Way
Building the Admissible Set



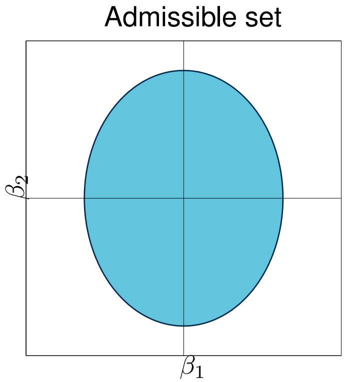
The Variational Way
Building the Admissible Set



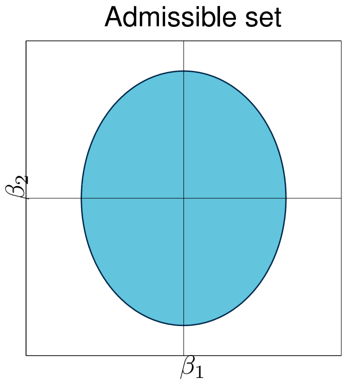
The Variational Way
Building the Admissible Set



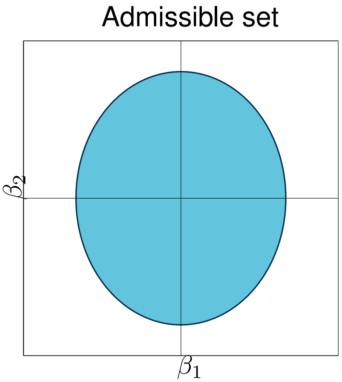
The Variational Way
Building the Admissible Set



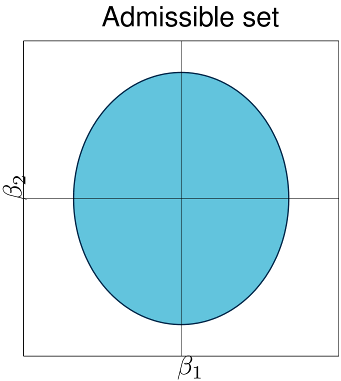
The Variational Way
Building the Admissible Set



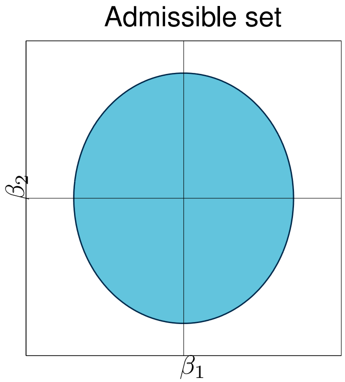
The Variational Way
Building the Admissible Set



The Variational Way
Building the Admissible Set

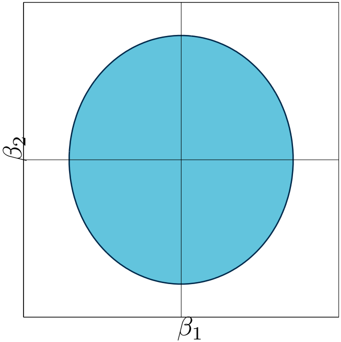


The Variational Way
Building the Admissible Set



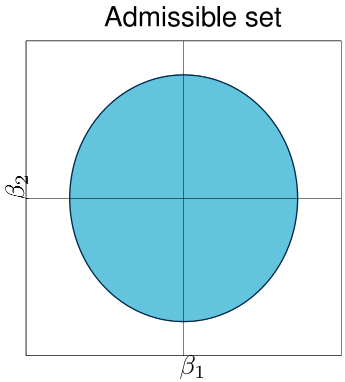
The Variational Way
Building the Admissible Set

Admissible set

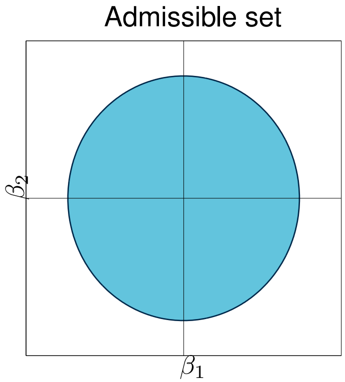


The Variational Way

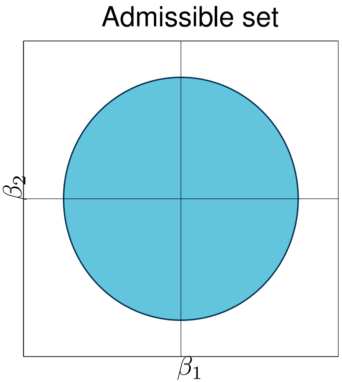
Building the Admissible Set



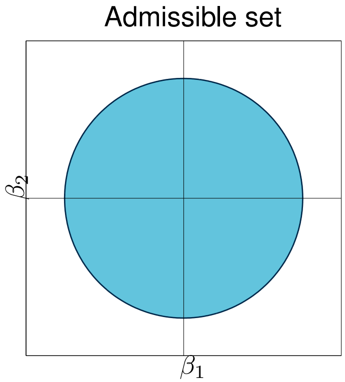
The Variational Way
Building the Admissible Set



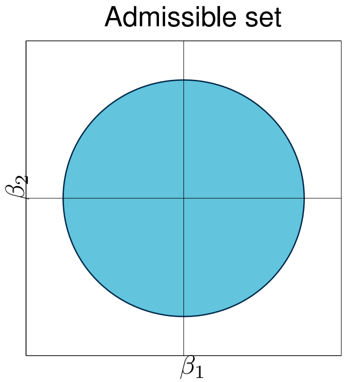
The Variational Way
Building the Admissible Set



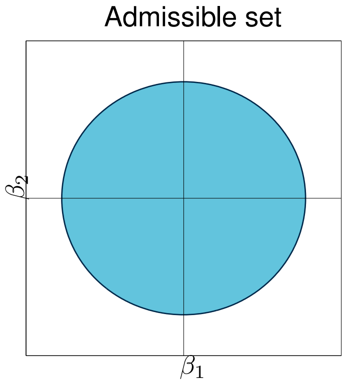
The Variational Way
Building the Admissible Set



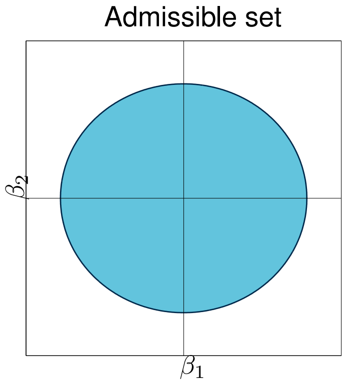
The Variational Way
Building the Admissible Set



The Variational Way
Building the Admissible Set

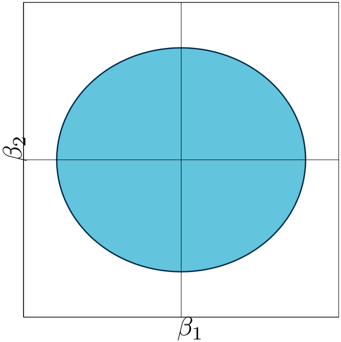


The Variational Way
Building the Admissible Set



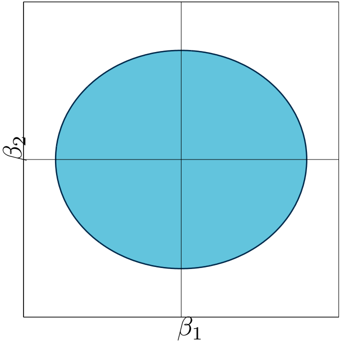
The Variational Way
Building the Admissible Set

Admissible set

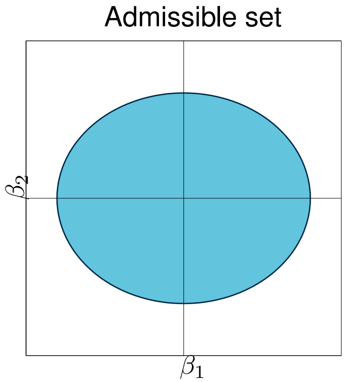


The Variational Way
Building the Admissible Set

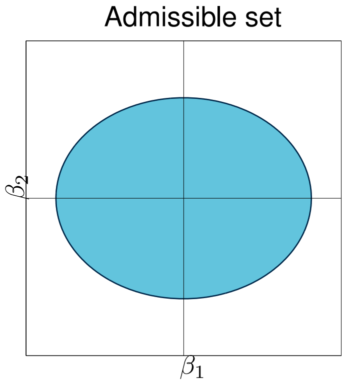
Admissible set



The Variational Way
Building the Admissible Set

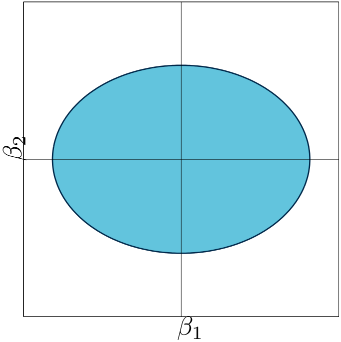


The Variational Way
Building the Admissible Set

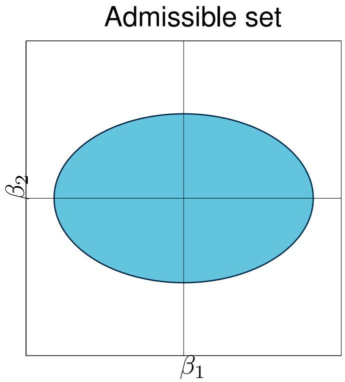


The Variational Way
Building the Admissible Set

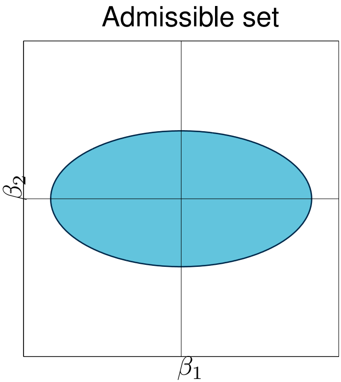
Admissible set



The Variational Way
Building the Admissible Set

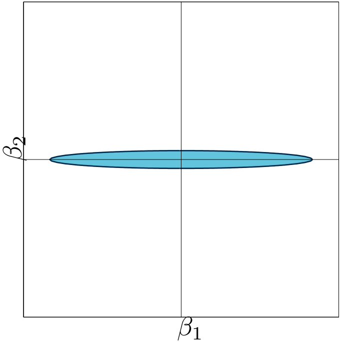


The Variational Way
Building the Admissible Set



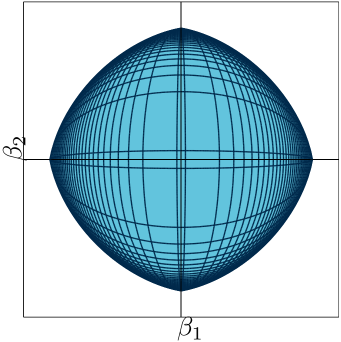
The Variational Way
Building the Admissible Set

Admissible set



The Variational Way
Building the Admissible Set

Admissible set



The admissible set is the union of ellipses

The Variational Way
Recap

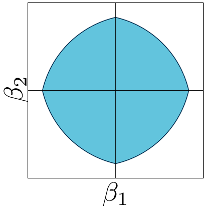
1. Provides an alternative view of sparsity-inducing penalties
   - provides insights on these methods
     - as an interpretation: in the hierarchical Bayesian framework
     - as a way to generalize them through the richness of quadratic penalties
   - allows to use some of the known results on ridge-like penalties
   - results in a generic algorithm for computing solutions
2. The associated algorithm relies on solving linear systems is
   - accurate
   - rather inefficient due to the number of systems to be solved
     - an infinite nunber of ellipses are required to cover the admissible set
     - these ellipses are degenerated at parsimonous solutions
       $\rightsquigarrow$ numerical stability issues
       $\rightsquigarrow$ alternative formulations with higher computational cost

The Duality Way

Going quadratic again: second attempt

Elastic-Net Example

$$\begin{cases} \min_{\boldsymbol{\beta} \in \mathbb{R}^p} \|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2 \\ \text{s. t. } \frac{1}{2}\|\boldsymbol{\beta}\|_2^2 + \eta\|\boldsymbol{\beta}\|_1 \leq s \end{cases}$$

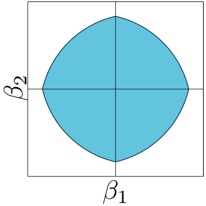


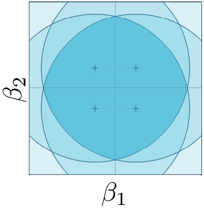
The Duality Way

Going quadratic again: second attempt

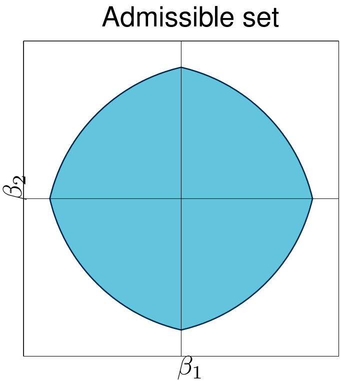
Elastic-Net Example

$$\begin{cases} \min_{\boldsymbol{\beta}\in\mathbb{R}^p} \|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2 \\ \text{s. t. } \frac{1}{2}\|\boldsymbol{\beta}\|_2^2 + \eta\|\boldsymbol{\beta}\|_1 \leq s \end{cases} \Leftrightarrow \begin{cases} \min_{\boldsymbol{\beta}\in\mathbb{R}^p} \|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2 \\ \text{s. t. } \max_{\boldsymbol{\gamma}\in\{-1,1\}^p} \frac{1}{2}\|\boldsymbol{\beta} - \eta\,\boldsymbol{\gamma}\|_2^2 \leq s + \frac{\eta^2}{2} \end{cases}$$

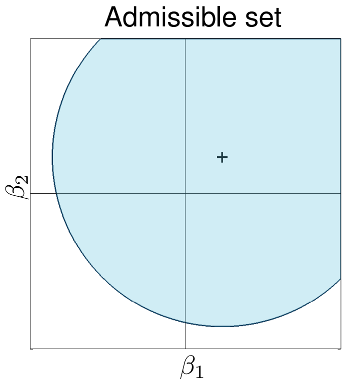




The Duality Way
Building the Admissible Set

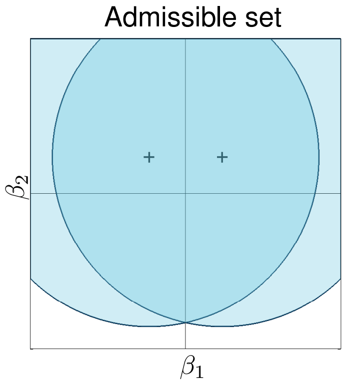


The Duality Way
Building the Admissible Set

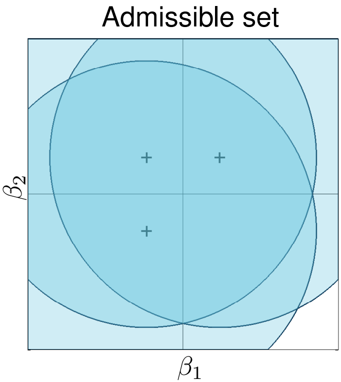


The Duality Way

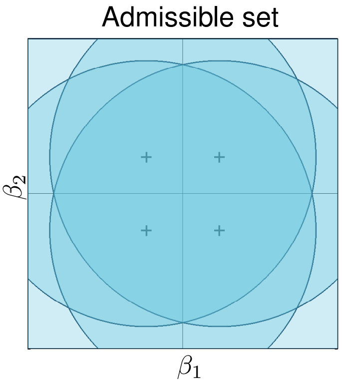
Building the Admissible Set



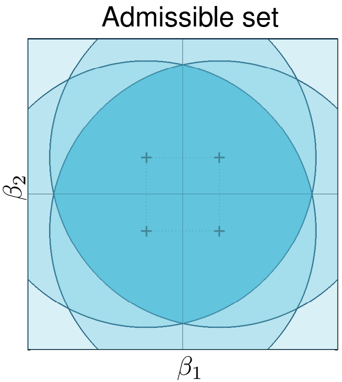
The Duality Way
Building the Admissible Set



The Duality Way
Building the Admissible Set



The Duality Way
Building the Admissible Set



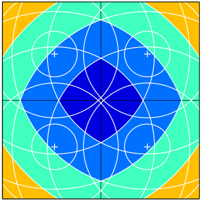
The admissible set is the intersection of ellipses

Solutions in $\boldsymbol{\beta}$ are defined by the worst-case γ

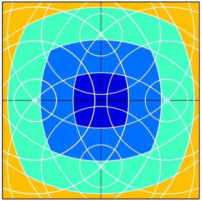
Outline

- Motivations
- Going Quadratic
  - The Variational Way
  - The Duality Way
- Benefits
  - Generality
  - Algorithm
  - Analysis
- Experiments
- Conclusion

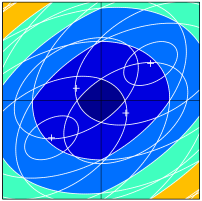
Beyond Elastic-Net



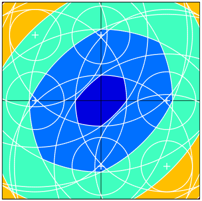
elastic-net ($\ell_1 + \ell_2$)



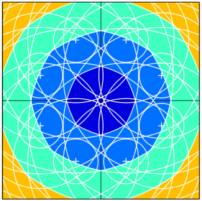
$\ell_\infty + \ell_2$



structured e.-n.



fused-lasso + ℓ_2



OSCAR + ℓ_2

Beyond Elastic-Net

General Formulation

$$\left\{\begin{array}{l}\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \|\mathbf{X}\boldsymbol{\beta}-\mathbf{y}\|_2^2 \\ \text{s. t. } \frac{1}{2}\|\boldsymbol{\beta}\|_\Omega^2+\eta\|\boldsymbol{\beta}\| \leq s\end{array}\right. \Leftrightarrow \left\{\begin{array}{l}\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \|\mathbf{X}\boldsymbol{\beta}-\mathbf{y}\|_2^2 \\ \text{s. t. } \max_{\boldsymbol{\gamma}\in\mathcal{D}_{\boldsymbol{\gamma}}} \frac{1}{2}\|\boldsymbol{\beta}\|_\Omega^2-\boldsymbol{\gamma}^t\boldsymbol{\beta} \leq s\end{array}\right.$$

where

$$\mathcal{D}_{\boldsymbol{\gamma}}=\{\boldsymbol{\gamma}\in\mathbb{R}^p : \|\boldsymbol{\gamma}\|_* \leq \eta\}$$

Simply reformulate with the dual norm to get a quadratic expression in β
γ is an adversarial prior

Generic Active Set Algorithm

S0 Initialization

$\boldsymbol{\beta} \leftarrow \boldsymbol{\beta}^0, \mathcal{A} \leftarrow \{j : \beta_j \neq 0\}$; // Start with a feasible $\boldsymbol{\beta}$

$\boldsymbol{\gamma} = \arg\max_{\mathbf{g} \in \mathcal{D}_{\boldsymbol{\gamma}}} -\mathbf{g}^t \boldsymbol{\beta}$; // Pick a worst admissible $\boldsymbol{\gamma}$

Generic Active Set Algorithm

S0 Initialization

$\boldsymbol{\beta} \leftarrow \boldsymbol{\beta}^0, \mathcal{A} \leftarrow \{j : \beta_j \neq 0\}$; // Start with a feasible $\boldsymbol{\beta}$

$\boldsymbol{\gamma} = \arg\max_{\mathbf{g} \in \mathcal{D}_{\boldsymbol{\gamma}}} -\mathbf{g}^t \boldsymbol{\beta}$; // Pick a worst admissible $\boldsymbol{\gamma}$

S1 Update active variables $\boldsymbol{\beta}_{\mathcal{A}}$

$\boldsymbol{\beta}_{\mathcal{A}} \leftarrow \left(\mathbf{X}_{\cdot\mathcal{A}}^{\intercal}\mathbf{X}_{\cdot\mathcal{A}} + \lambda \mathbf{I}_{|\mathcal{A}|}\right)^{-1} \left(\mathbf{X}_{\cdot\mathcal{A}}^{\intercal}\mathbf{y} + \lambda \boldsymbol{\gamma}_{\mathcal{A}}\right)$; // Subproblem resolution

Generic Active Set Algorithm

S0 Initialization

$\boldsymbol{\beta} \leftarrow \boldsymbol{\beta}^0, \mathcal{A} \leftarrow \{j : \beta_j \neq 0\}$; // Start with a feasible $\boldsymbol{\beta}$

$\boldsymbol{\gamma} = \arg\max_{\mathbf{g} \in \mathcal{D}_{\boldsymbol{\gamma}}} -\mathbf{g}^t \boldsymbol{\beta}$; // Pick a worst admissible $\boldsymbol{\gamma}$

S1 Update active variables $\boldsymbol{\beta}_{\mathcal{A}}$

$\boldsymbol{\beta}_{\mathcal{A}} \leftarrow \left(\mathbf{X}_{\cdot\mathcal{A}}^{\intercal} \mathbf{X}_{\cdot\mathcal{A}} + \lambda \mathbf{I}_{|\mathcal{A}|}\right)^{-1} \left(\mathbf{X}_{\cdot\mathcal{A}}^{\intercal} \mathbf{y} + \lambda \boldsymbol{\gamma}_{\mathcal{A}}\right)$; // Subproblem resolution

S2 Verify coherence of $\boldsymbol{\gamma}_{\mathcal{A}}$ with the updated $\boldsymbol{\beta}_{\mathcal{A}}$

if $-\boldsymbol{\gamma}_{\mathcal{A}}^t \boldsymbol{\beta}_{\mathcal{A}} < \max_{\mathbf{g} \in \mathcal{D}_{\boldsymbol{\gamma}}} -\mathbf{g}_{\mathcal{A}}^t \boldsymbol{\beta}_{\mathcal{A}}$ **then** // if $\boldsymbol{\gamma}_{\mathcal{A}}$ is not worst-case

$\boldsymbol{\beta}_{\mathcal{A}} \leftarrow \boldsymbol{\beta}_{\mathcal{A}}^{\text{old}} + \rho(\boldsymbol{\beta}_{\mathcal{A}} - \boldsymbol{\beta}_{\mathcal{A}}^{\text{old}})$; // Last $\boldsymbol{\gamma}_{\mathcal{A}}$-coherent solution

Generic Active Set Algorithm

S0 Initialization

$\boldsymbol{\beta} \leftarrow \boldsymbol{\beta}^0, \mathcal{A} \leftarrow \{j : \beta_j \neq 0\}$; // Start with a feasible $\boldsymbol{\beta}$

$\boldsymbol{\gamma} = \arg\max_{\mathbf{g} \in \mathcal{D}_{\boldsymbol{\gamma}}} -\mathbf{g}^t \boldsymbol{\beta}$; // Pick a worst admissible $\boldsymbol{\gamma}$

S1 Update active variables $\boldsymbol{\beta}_{\mathcal{A}}$

$\boldsymbol{\beta}_{\mathcal{A}} \leftarrow \left(\mathbf{X}_{\cdot\mathcal{A}}^{\mathsf{T}} \mathbf{X}_{\cdot\mathcal{A}} + \lambda \mathbf{I}_{|\mathcal{A}|}\right)^{-1} \left(\mathbf{X}_{\cdot\mathcal{A}}^{\mathsf{T}} \mathbf{y} + \lambda \boldsymbol{\gamma}_{\mathcal{A}}\right)$; // Subproblem resolution

S2 Verify coherence of $\boldsymbol{\gamma}_{\mathcal{A}}$ with the updated $\boldsymbol{\beta}_{\mathcal{A}}$

if $-\boldsymbol{\gamma}_{\mathcal{A}}^t \boldsymbol{\beta}_{\mathcal{A}} < \max_{\mathbf{g} \in \mathcal{D}_{\boldsymbol{\gamma}}} -\mathbf{g}_{\mathcal{A}}^t \boldsymbol{\beta}_{\mathcal{A}}$ **then** // if $\boldsymbol{\gamma}_{\mathcal{A}}$ is not worst-case

  $\boldsymbol{\beta}_{\mathcal{A}} \leftarrow \boldsymbol{\beta}_{\mathcal{A}}^{\text{old}} + \rho(\boldsymbol{\beta}_{\mathcal{A}} - \boldsymbol{\beta}_{\mathcal{A}}^{\text{old}})$; // Last $\boldsymbol{\gamma}_{\mathcal{A}}$-coherent solution

S3 Update active set $\mathcal{A}$

$g_j \leftarrow \min_{\boldsymbol{\gamma} \in \mathcal{D}_{\boldsymbol{\gamma}}} \left| \mathbf{x}_j^{\mathsf{T}} (\mathbf{X}_{\cdot\mathcal{A}} \boldsymbol{\beta}_{\mathcal{A}} - \mathbf{y}) + \lambda(\beta_j - \gamma_j) \right| \quad j = 1, \ldots, p$ // worst-case gradient

if $\exists j \in \mathcal{A} : \beta_j = 0$ *and* $g_j = 0$ **then**

  $\mathcal{A} \leftarrow \mathcal{A} \backslash \{j\}$; // Downgrade j

else

  if $\max_{j \in \mathcal{A}^c} g_j \neq 0$ **then**

    $j^\star \leftarrow \arg\max_{j \in \mathcal{A}^c} g_j$, $\mathcal{A} \leftarrow \mathcal{A} \cup \{j^\star\}$; // Upgrade $j^\star$

  else

    Stop and return $\boldsymbol{\beta}$, which is optimal

Monitoring Convergence
Optimality Gap

Proposition: Let $\mathcal{D}_{\boldsymbol{\gamma}} = \{\boldsymbol{\gamma} \in \mathbb{R}^p : \|\boldsymbol{\gamma}\|_* \leq \eta\}$. For any $\|\cdot\|_*$ and $\eta > 0$, $\forall \boldsymbol{\gamma} \in \mathbb{R}^p : \|\boldsymbol{\gamma}\|_* \geq \eta$, we have:

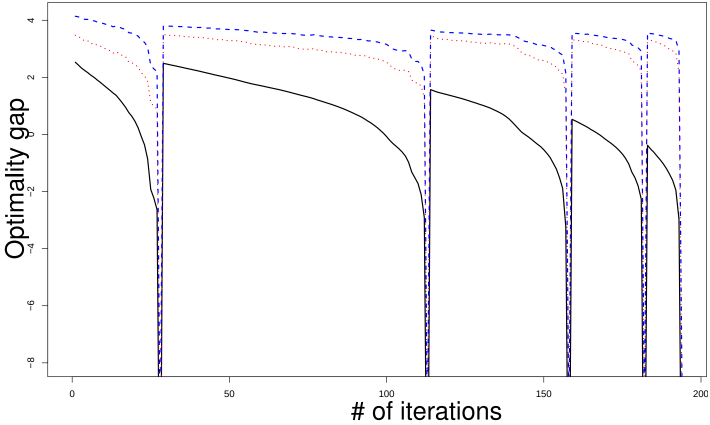
$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \max_{\boldsymbol{\gamma}' \in \mathcal{D}_{\boldsymbol{\gamma}}} J_\lambda(\boldsymbol{\beta}, \boldsymbol{\gamma}') \geq \frac{\eta}{\|\boldsymbol{\gamma}\|_*} J_\lambda\left(\boldsymbol{\beta}^\star\left(\boldsymbol{\gamma}\right), \boldsymbol{\gamma}\right) - \frac{\lambda\eta(\|\boldsymbol{\gamma}\|_* - \eta)}{\|\boldsymbol{\gamma}\|_*^2} \|\boldsymbol{\gamma}\|_2^2 \quad ,$$

where

$$J_\lambda(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2 + \lambda \|\boldsymbol{\beta} - \boldsymbol{\gamma}\|_2^2 \text{ and } \boldsymbol{\beta}^\star(\boldsymbol{\gamma}) = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} J_\lambda(\boldsymbol{\beta}, \boldsymbol{\gamma}) \quad .$$

Optimality gap: pick a γ-value such that the current worst-case gradient is null (the current $\boldsymbol{\beta}$-value then being the optimal $\boldsymbol{\beta}^\star(\boldsymbol{\gamma})$).

Monitoring Convergence
Illustration



True optimality gap along a solution path (solid black), our upper bound (dashed blue) and Fenchel's duality gap (dotted red).

Outline

- Motivations
- Going Quadratic
  - The Variational Way
  - The Duality Way
- Benefits
  - Generality
  - Algorithm
  - Analysis
- **Experiments**
- Conclusion

Comparison of Stand-Alone Implementations
Small-Medium Problem Sizes

We compare `R`-packages on Lasso problems:

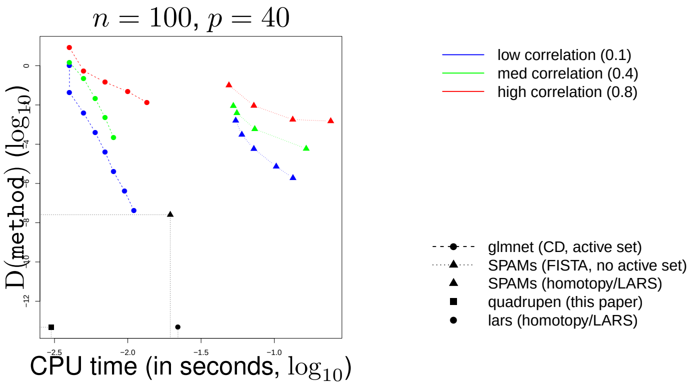
1. accelerated proximal methods – `SPAMs-FISTA` (Mairal *et al.*),
2. coordinate descent – `glmnet` (Friedman *et al.*),
3. homotopy/LARS algorithm– `lars` (Hastie and Efron) and `SPAMs-LARS`,
4. our implementation – `quadrupen`.

The distance to the optimum is averaged along the regularization path by

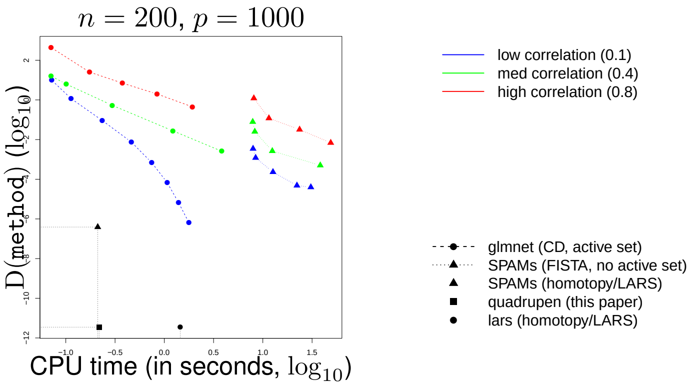
$$\mathrm{D}(\texttt{method}) = \left(\frac{1}{|\Lambda|} \sum_{\lambda \in \Lambda} \left(J_\lambda^{\text{lasso}} \left(\hat{\beta}_\lambda^{\texttt{lars}} \right) - J_\lambda^{\text{lasso}} \left(\hat{\beta}_\lambda^{\texttt{method}} \right) \right)^2 \right)^{1/2},$$

where Λ is given by the first $\min(n,p)$ steps of `lars`.
$\rightsquigarrow$ Vary $\{\rho, (p,n)\}$, fix $s = 0.25 \min(n,p)$ and average over 50 runs.

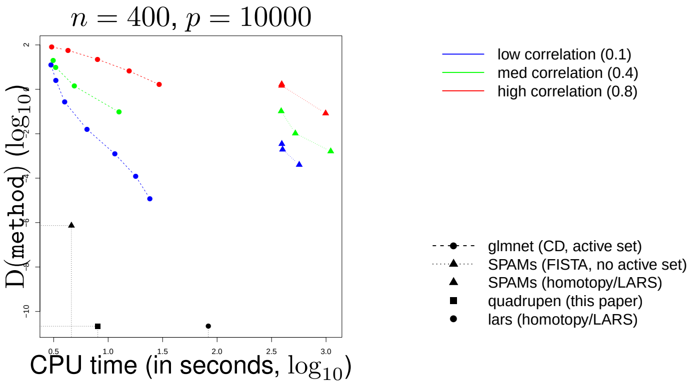
Experimental results



Experimental results



Experimental results



- Solving systems is a good strategy for this range of problem sizes
- Comparing speed is not enough: inaccuracy impacts test results

Outline

- Motivations
- Going Quadratic
  - The Variational Way
  - The Duality Way
- Benefits
  - Generality
  - Algorithm
  - Analysis
- Experiments
- Conclusion

The Duality Way
Recap

1. Provides an unifying view of sparsity-inducing penalties
   - provides insights on these methods
     - as an interpretation: robust optimization
     - as a way to build penalties $\rightsquigarrow$ which solutions should be avoided?
     - as way to derive generic results
       $\rightsquigarrow$ monitoring of convergence (limited practical use)
     - to promote efficiency
       $\rightsquigarrow$ $\mathcal{D}_\gamma$ polytope with not too many vertices
   - results in a generic algorithm for computing solutions
2. The associated algorithm relies on solving linear systems is
   - accurate
   - efficient for small to medium scale problems (thousands of variables)

Available R-package, with stability selection