Regularized Linear Models

Master's Degree in Bioinformatics and Computational Biology - Machine Learning

Carlos María Alaíz Gudín

Escuela Politécnica Superior
Universidad Autónoma de Madrid

Academic Year 2024–25

Universidad Autónoma
de Madrid

Contents

Exercise ▷ **Questionnaire**

Please, fill in the questionnaire regarding your prior knowledge about this topic.

Introduction

Need of Regularization - Exercise

Exercise

▷ **Questionnaire**

Given a 3-dimensional problem with the following data:

$x_{i,1}$	$x_{i,2}$	$x_{3,2}$	y_i
1	0	1	2
1	1	1	3

1. Define a linear model $\{b, w_1, w_2, w_3\}$ with the smaller possible Mean Squared Error (MSE). Is it possible to get a perfect training prediction?
2. Are there more than one model that can solve perfectly the problem above? Is there anyway to determine which one should be preferred?

Need of Regularization - Example

Example ("Ill-Posed" Problem)

- Regression dataset `E2006-log1p` of the LIBSVM repository.
  - 16 087 patterns for training, 3308 patterns for testing.
  - 4 272 227 features.
- Even the simplest models (linear) will have 220 free parameters per pattern.
- The complexity of the model has to be controlled.
- Probably not all the features will be relevant.
  - A model based on a subset of the features seems a sensible option.

Bias–Variance and Regularization (I)

Assumption

- $\mathbf{x}$ and y are related as $y = f^\star(\mathbf{x}) + \epsilon$.
- $f^\star(\cdot)$ is the true underlying function.
- ϵ is additive noise with zero mean and finite variance.

- The model should try to approximate the underlying function, $f \approx f^\star$.
  - The distance between f and $f^\star$ is formalised under the concept of **bias**.
  - A small bias can be achieved with highly flexible models with many parameters.

- Nevertheless, the model depends on the particular training sample, so it can be denoted by $f_{\mathcal{D}}$.
- The model should be stable, in the sense that for different datasets $\mathcal{D}$ and $\mathcal{D}'$, $f_{\mathcal{D}} \approx f_{\mathcal{D}'}$.
  - This stability is formalised under the concept of **variance**.
  - A small variance can be achieved with simple models with few parameters.

- A **trade-off** has to be found.

Bias–Variance and Regularization (II)

Bias–Variance Trade-off

- Error due to **Bias**: Difference between the expected prediction of the model and the correct value to be predicted.
- Error due to **Variance**: Variability of the model prediction for a given data point.

Definition (Regularization)

- **Regularization** usually denotes the set of techniques that attempt to improve the estimates by biasing them away from their sample-based values towards values that are deemed to be more “physically plausible”.
- The variance of the model is reduced to the expense of a potentially higher bias.

Over-Fitting and Under-Fitting (I)

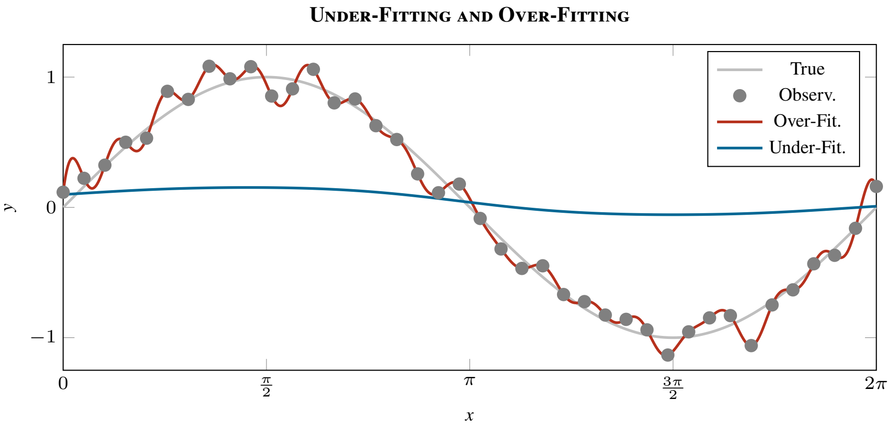
Over-Fitting

- The resultant model is overly complex to describe the data under study.
  - Limited number of training data.
  - Learning machine too complex (many free parameters).
- Large variance, small bias.

Under-Fitting

- The resultant model is overly simple to describe the data under study.
  - Learning machine too rigid.
- Large bias, small variance.

Over-Fitting and Under-Fitting (II)



Why Is Regularization Necessary?

1. There are more variables than observations ($d \gg N$).
2. The optimum estimator is not unique.
3. Numerical instabilities (e.g. if $\mathbf{X}^\intercal\mathbf{X}$ is close to singular): small changes in the data lead to large changes in the model.
4. Over-fitting avoidance: obtain more robust models that generalize well.
5. Parsimony and interpretability: simpler models can help to understand better the relation between inputs and outputs.

Notebook

The Need of Regularization

Regularized Learning

- Regularized learning consists in models trained by optimizing objective functions of the form:

$$\mathcal{S} = \mathcal{E}_{\mathcal{D}} + \gamma \mathcal{R}.$$

- The main term of the objective function is an error term $\mathcal{E}_{\mathcal{D}}$.
  - It represents how well the model fits the training data $\mathcal{D}$.
  - Examples: Mean Squared Error (regression) and minus (log)likelihood (classification).
- The additional term is a regularization term $\mathcal{R}$. It penalizes the complexity of the model, with several purposes:
  - Avoid over-fitting.
  - Introduce prior knowledge.
  - Enforce certain desirable properties.
- γ is a regularization parameter.
  - It is responsible for the balance between accuracy and complexity.

Regularization Functions

Regularization Functions

- There are different regularization functions $\mathcal{R}(\boldsymbol{\theta})$ that assign to each set of parameters $\boldsymbol{\theta}$ a measure of its complexity.
- Depending on the chosen function, the effect over $\boldsymbol{\theta}$ will change.
- The influence of the regularization functions is particularly clear on linear models.
  - Each coefficient of $\mathbf{w}$ corresponds to an input feature.
  - If $w_i = 0$, then the i-th feature is ignored.
  - If $w_i = w_j$, then the i-th feature is somehow similar to the j-th feature.

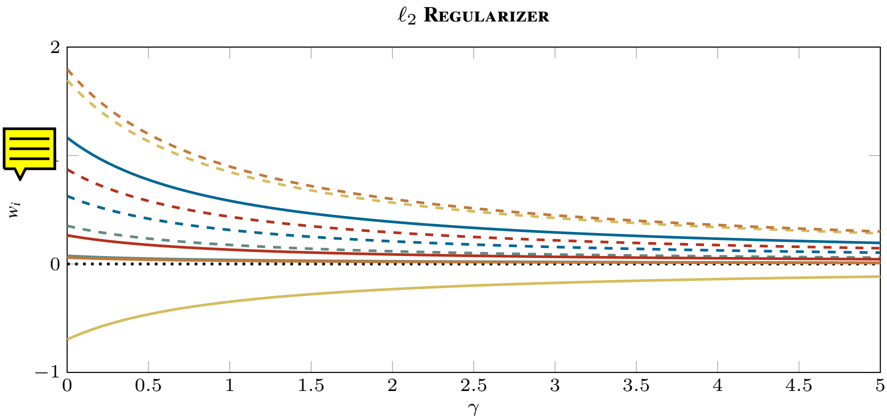
ℓ_2 Norm (I)

- Classical term, known as Tikhonov regularization, it corresponds to the sum of the squares of the entries:

$$\mathcal{R}(\mathbf{w}) = \|\mathbf{w}\|_2^2 = \sum_{i=1}^{d} w_i^2.$$

- It controls the complexity of the model.
- It is differentiable, and hence easy to optimize.
- It pushes the entries towards zero.

ℓ_2 Norm (II)



ℓ_2 Norm - Exercise

Exercise

▷ **Questionnaire**

Given the following 3-dimensional linear models, compute their squared ℓ_2 norm to check which one is simpler according to this criterion:

1. $\{w_1 = 1, w_2 = 1, w_3 = 1\}$.
2. $\{w_1 = 3, w_2 = 0, w_3 = 0\}$.
3. $\{w_1 = 2, w_2 = 2, w_3 = 0\}$.

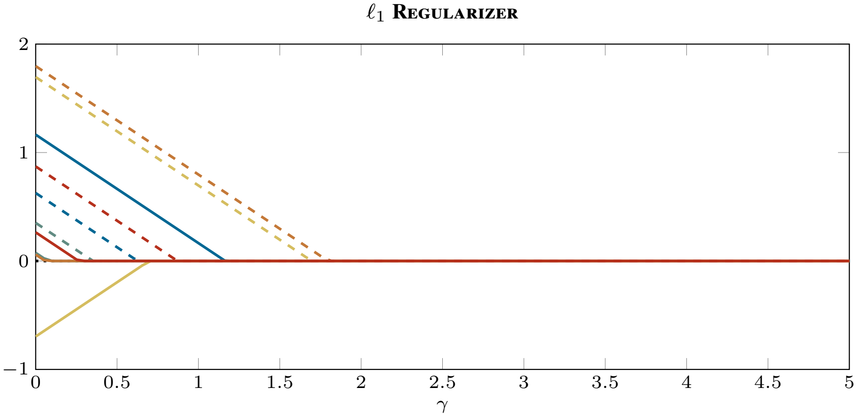
ℓ_1 Norm (I)

- It corresponds to the sum of the absolute values of the entries:

$$\mathcal{R}(\mathbf{w}) = \|\mathbf{w}\|_1 = \sum_{i=1}^{d} |w_i|.$$

- It controls the complexity of the model.
- The absolute value is non-differentiable around zero, and hence this term is more involved to optimize.
- It pushes the entries towards zero enforcing some of them to be identically zero.
  - It enforces sparsity.

ℓ_1 Norm (II)



ℓ_1 Norm - Exercise

Exercise

▷ **Questionnaire**

Given the following 3-dimensional linear models, compute their ℓ_1 norm to check which one is simpler according to this criterion:

1. $\{w_1 = 1, w_2 = 1, w_3 = 1\}$.
2. $\{w_1 = 3, w_2 = 0, w_3 = 0\}$.
3. $\{w_1 = 2, w_2 = 2, w_3 = 0\}$.

More regularizers in the appendix **Additional Regularization Functions**.

Regularized Linear Models

The Optimization Problem of a Regularized Model

- The optimization problem to train a regularized model can be formulated as:

$$\min_{\boldsymbol{\theta}} \; \{\mathcal{E}_{\mathcal{D}}(\boldsymbol{\theta}) + \gamma \mathcal{R}(\boldsymbol{\theta})\}.$$

- There exists an equivalence between this unconstrained model and the following constrained formulation:

$$\min_{\boldsymbol{\theta}} \; \{\mathcal{E}_{\mathcal{D}}(\boldsymbol{\theta})\} \; \text{ s.t. } \; \mathcal{R}(\boldsymbol{\theta}) \leq c.$$

- In the case of a regression linear model:

$$\min_{\mathbf{w} \in \mathbb{R}^d} \; \left\{\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \gamma \mathcal{R}(\mathbf{w})\right\} \equiv \min_{\mathbf{w} \in \mathbb{R}^d} \; \left\{\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2\right\} \; \text{ s.t. } \; \mathcal{R}(\mathbf{w}) \leq c.$$

- In the case of a classification linear model:

$$\min_{\mathbf{w} \in \mathbb{R}^d} \; \{\mathrm{CE}(\mathbf{w}) + \gamma \mathcal{R}(\mathbf{w})\} \equiv \min_{\mathbf{w} \in \mathbb{R}^d} \; \{\mathrm{CE}(\mathbf{w})\} \; \text{ s.t. } \; \mathcal{R}(\mathbf{w}) \leq c.$$

Ridge Regression

- This linear model uses the Tikhonov regularization:

$$\mathcal{R}(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|_2^2 = \frac{1}{2}\sum_{i=1}^{d} \mathbf{w}_i^2.$$

- The objective function is:

$$\mathcal{S}(\mathbf{w}) = \mathrm{MSE}(\mathbf{w}) + \frac{\gamma}{2}\|\mathbf{w}\|_2^2.$$

- The complexity of the model is controlled.
  - In the presence of noisy inputs:

$$\mathbf{w}^\intercal(\mathbf{x}+\boldsymbol{\epsilon}) \overset{?}{\approx} \mathbf{w}^\intercal\mathbf{x} \iff |\mathbf{w}^\intercal(\mathbf{x}+\boldsymbol{\epsilon}) - \mathbf{w}^\intercal\mathbf{x}| = |\mathbf{w}^\intercal\mathbf{x} + \mathbf{w}^\intercal\boldsymbol{\epsilon} - \mathbf{w}^\intercal\mathbf{x}| \le \|\mathbf{w}\|_2\|\boldsymbol{\epsilon}\|_2 \approx 0.$$

- No structure is imposed (the resultant model typically depends on all the variables).

- The problem is convex and differentiable.

Ridge Regression: Optimization

$$\min_{\mathbf{w}\in\mathbb{R}^d} \left\{ \frac{1}{2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \frac{\gamma}{2}\|\mathbf{w}\|_2^2 \right\}.$$

$$\begin{aligned}
\nabla_{\mathbf{w}} \mathcal{S}(\mathbf{w})\big|_{\mathbf{w}=\mathbf{w}^\star} = \mathbf{0} &\implies -\mathbf{X}^\intercal(\mathbf{y} - \mathbf{X}\mathbf{w}^\star) + \gamma\mathbf{w}^\star = \mathbf{0} \\
&\implies -\mathbf{X}^\intercal\mathbf{y} + \mathbf{X}^\intercal\mathbf{X}\mathbf{w}^\star + \gamma\mathbf{w}^\star = \mathbf{0} \\
&\implies (\mathbf{X}^\intercal\mathbf{X} + \gamma\mathbf{I})\mathbf{w}^\star = \mathbf{X}^\intercal\mathbf{y} \\
&\implies \boxed{\mathbf{w}^\star = (\mathbf{X}^\intercal\mathbf{X} + \gamma\mathbf{I})^{-1}\mathbf{X}^\intercal\mathbf{y}}.
\end{aligned}$$

Notebook

Ridge Regression

Lasso

- This linear model uses as regularizer the ℓ_1 norm:

$$\mathcal{R}(\mathbf{w}) = \|\mathbf{w}\|_1 = \sum_{i=1}^{d} |w_i|.$$

- The objective function is:

$$\mathcal{S}(\mathbf{w}) = \mathrm{MSE}(\mathbf{w}) + \gamma \|\mathbf{w}\|_1.$$

- This regularizer enforces some of the coefficients to be identically zero.
  - The model performs an implicit feature selection: the features with coefficient equal to zero can be discarded.
  - It also avoids over-fitting.

- The problem is convex but non-differentiable.

Notebook

Lasso

Elastic–Net

- This linear model combines the advantages of the ℓ_1 norm with those of the ℓ_2 norm.
- It is more stable than Lasso regarding feature selection.
- The regularizer is therefore a combination of both:

$$\mathcal{R}(\mathbf{w}) = \|\mathbf{w}\|_1 + \frac{\gamma_2'}{2}\|\mathbf{w}\|_2^2 .$$

- Thus the objective function is:

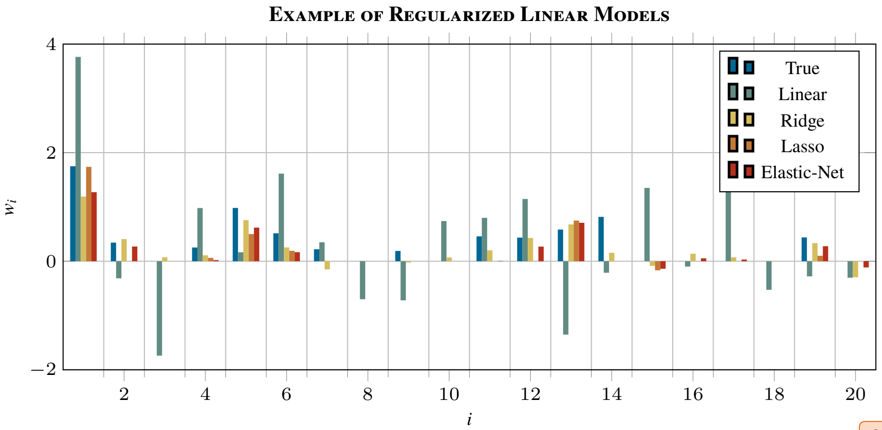
$$\mathcal{S}(\mathbf{w}) = \mathrm{MSE}(\mathbf{w}) + \gamma_1 \|\mathbf{w}\|_1 + \frac{\gamma_2}{2}\|\mathbf{w}\|_2^2 .$$

- The problem is convex but non-differentiable.

Notebook

Elastic–Net

Illustration



More models in the appendix **Additional Regularized Linear Models**.

Summary

Regularized Linear Models: Summary

- **Regularization** is often needed in real problems to control the complexity or induce structure.
- **Regularized models** are trained by minimizing both an error term and a regularization term.

- There are different choices for the regularization functions, two of the most important are:
  - The ℓ_2 norm, which controls the complexity.
  - The ℓ_1 norm, which controls the complexity and induces sparsity.

- The resultant regularized linear models are:
  - **Ridge Regression**, based on the ℓ_2 norm.
  - **Lasso**, based on the ℓ_1 norm.
  - **Elastic–Net**, based on the combination of both regularizers.

Regularized Linear Models

Carlos María Alaíz Gudín

Additional Material

$\ell_{2,1}$ Norm: Framework

- Each $\mathbf{w}$ is composed by d_g groups of $d_f = \frac{d}{d_g}$ features each group:

$$\mathbf{w} = \begin{pmatrix} w_{1,1} \\ \vdots \\ w_{1,d_f} \\ \vdots \\ w_{d_g,1} \\ \vdots \\ w_{d_g,d_f} \end{pmatrix},$$

  where $w_{g,f}$ is the f-th entry of the g-th group.
  - This framework can be easily extended to groups of different sizes.
- The variable $\mathbf{w}$ can be seen also as a matrix with d_f rows and d_g columns.
- The regularizers should respect this structure.

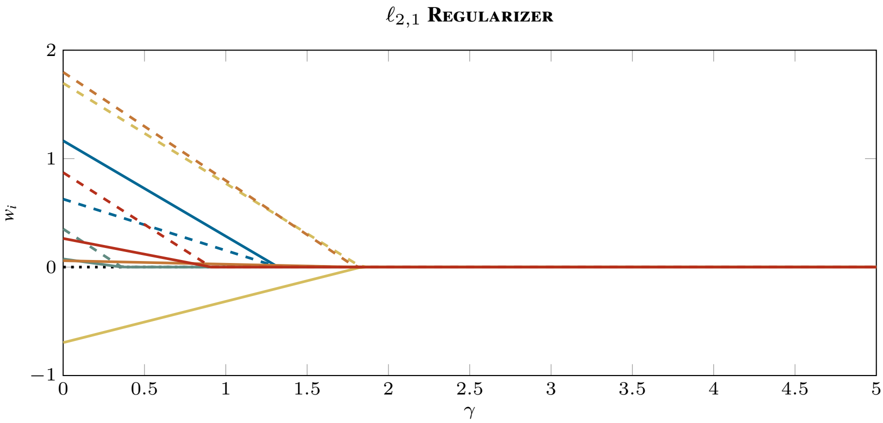
$\ell_{2,1}$ Norm (I)

- The regularizer is the $\ell_{2,1}$ norm:

$$\mathcal{R}(\mathbf{w}) = \|\mathbf{w}\|_{2,1} = \sum_{g=1}^{d_g} \sqrt{\sum_{f=1}^{d_f} w_{g,f}^2},$$

  which is just the ℓ_1 norm of the ℓ_2 norm of the different groups.
- It controls the complexity of the model.
- The ℓ_2 norm (not squared) is non-differentiable around zero, hence this term is more involved to optimize.
- It pushes the groups towards zero enforcing some of them to be identically zero.
  - It enforces sparsity at group level.

$\ell_{2,1}$ Norm (II)



Transformed Norms

- The regularization is applied over a linear transformation $\mathbf{Tw}$.
- The transformation allows for more involved structures.

Generalized ℓ_2 Norm

- The regularizer is $\mathcal{R}(\mathbf{w}) = \|\mathbf{Tw}\|_2^2$.
- It pushes the transformed vector towards zero.

Generalized Lasso

- The regularizer is $\mathcal{R}(\mathbf{w}) = \|\mathbf{Tw}\|_1$.
- It pushes the transformed vector towards zero enforcing some of the elements to be identically zero.
  - It enforces sparsity over the transformed vector.

Transformed Norms: Total Variation (I)

- The Total Variation is a family of regularizers that penalize the differences between adjacent entries.
  - It assumes some spatial location.
- It transforms the variable through a differentiating matrix:

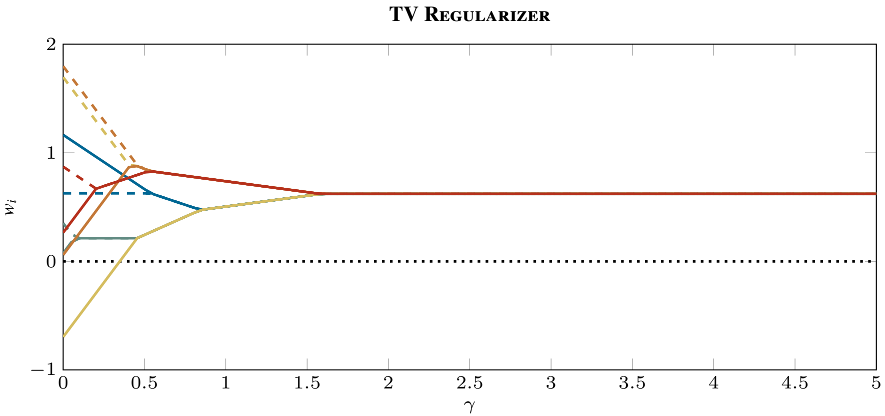
$$\mathbf{D} = \begin{pmatrix} -1 & 1 & 0 & \dots & 0 & 0 \\ 0 & -1 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & -1 & 1 \end{pmatrix}.$$

- The TV regularizer penalizes the ℓ_1 norm of the differences:

$$\mathcal{R}(\mathbf{w}) = \|\mathbf{D}\mathbf{w}\|_1 = \sum_{i=2}^{d} |w_i - w_{i-1}|.$$

  - The ℓ_1 norm enforces sparsity.
  - Some of the terms $w_i - w_{i-1}$ are zero, and hence $w_i = w_{i-1}$.
  - The vector $\mathbf{w}$ is piece-wise constant.

Transformed Norms: Total Variation (II)



Transformed Norms: Others

Graph-Based Total Variation

- An extension of the Total Variation regularizer.
- The differences between any pair of entries connected according to a graph are penalized.
- The classical Total Variation is recovered when the graph is a chain.
- When the graph is a lattice, it becomes a two-dimensional Total Variation.

Trend Filtering

- Similar idea than Total Variation but for higher degrees.
- Instead of penalizing the first differences, higher orders are penalized.

Combinations

- The previous regularizers can be combined to enforce several structures at the same time.

ℓ_1 and $\ell_{2,1}$

- Sparsity both at group level and at coefficient level.

ℓ_1 and Total Variation

- Some of the entries are identically zero.
- The remaining entries tend to be piece-wise constant.

Group Variants: Framework

- In certain circumstances, some features are grouped as corresponding to the same source.
  - E.g., different meteorological variables (wind speed, temperature) corresponding to the same geographical point.
- A grouping effect in the features is thus desirable.
  - All the features of a group should be active, or inactive, at the same time.
  - But they are different features, and they can have different coefficients.
- In this way, relevant groups can be detected.

Group Lasso and Group Elastic–Net

Group Lasso Model

- This linear model uses as regularizer the $\ell_{2,1}$ norm, $\mathcal{R}(\mathbf{w}) = \|\mathbf{w}\|_{2,1}$.
- The objective function is:

$$\mathcal{S}(\mathbf{w}) = \mathrm{MSE}(\mathbf{w}) + \gamma\|\mathbf{w}\|_{2,1}.$$

Group Elastic–Net Model

- The regularizer is a combination of the $\ell_{2,1}$ norm and the ℓ_2 norm.
- The objective function is:

$$\mathcal{S}(\mathbf{w}) = \mathrm{MSE}(\mathbf{w}) + \gamma_1\|\mathbf{w}\|_{2,1} + \frac{\gamma_2}{2}\|\mathbf{w}\|_2^2.$$

Fused Lasso

- This linear model uses as regularizer the ℓ_1 norm and the TV regularizer:
$$\mathcal{R}(\mathbf{w}) = \|\mathbf{w}\|_1 + \gamma_2' \operatorname{TV}(\mathbf{w}).$$
- It assumes that the features have some spatial location, and that they are ordered according to it.
  - A sensible model should assign similar coefficients to adjacent features.
- There are, therefore, sparse and piece-wise constant coefficients.
- The objective function is:
$$\mathcal{S}(\mathbf{w}) = \operatorname{MSE}(\mathbf{w}) + \gamma_1 \|\mathbf{w}\|_1 + \gamma_2 \operatorname{TV}(\mathbf{w}).$$

Illustration (I)

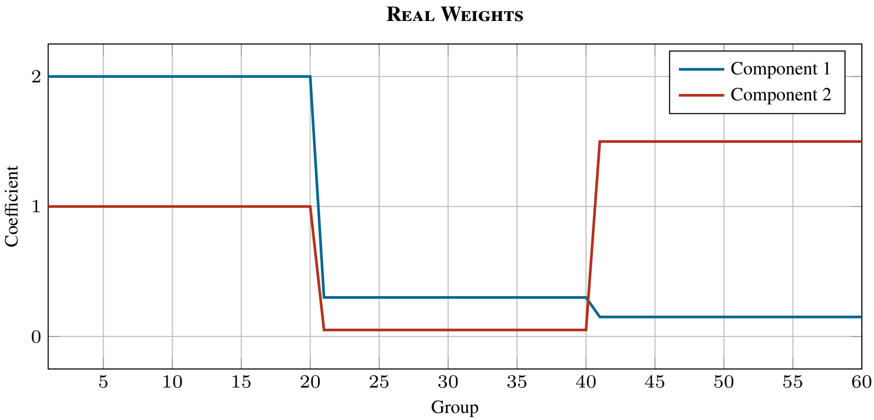


Illustration (II)

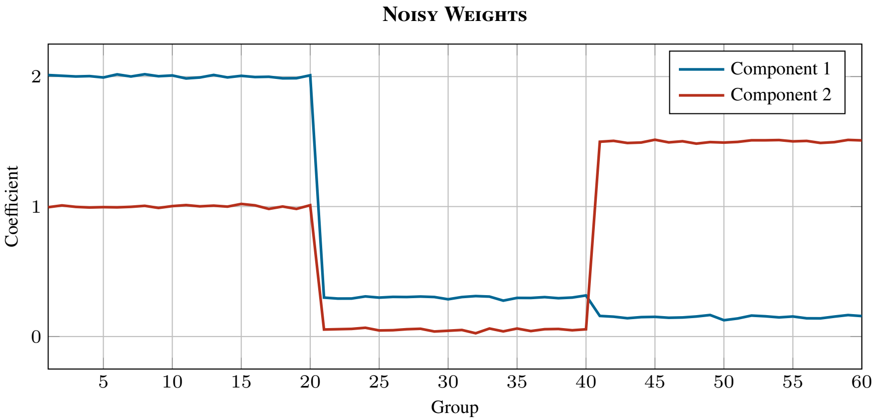


Illustration (III)

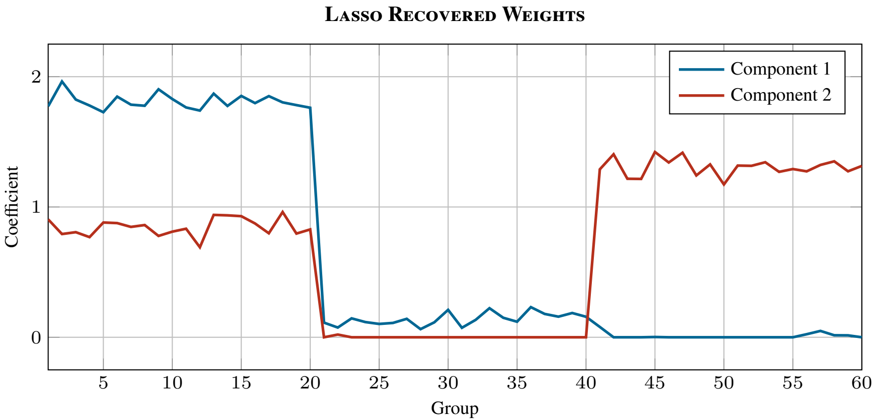


Illustration (IV)

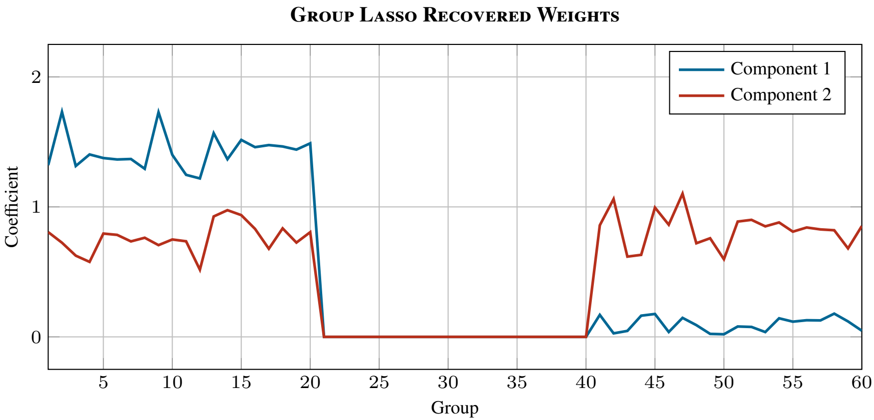


Illustration (V)