Integration of Prior Knowledge for Regression and Classification with Sparse Grids

Lukas Krenz

November 15, 2016

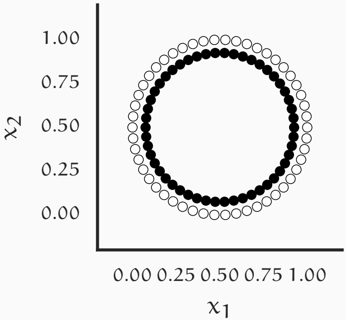


Figure 1: Original data

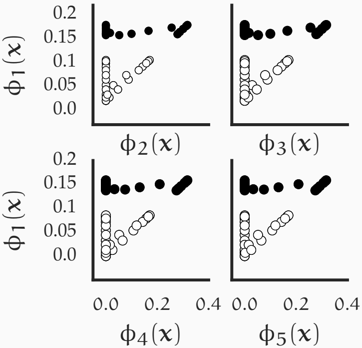


Figure 2: Transformed features

Optimization Goal

Model Matrix: p features $\rightarrow$ m grid points

$$\mathbf{\Phi}(x) = \begin{bmatrix} \phi_1(x_1) & \phi_2(x_1) & \dots & \phi_m(x_1) \\ \phi_1(x_2) & \phi_2(x_2) & \dots & \phi_m(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_1(x_n) & \phi_2(x_n) & \dots & \phi_m(x_n) \end{bmatrix} \tag{1}$$

Optimization goal:

$$\min_{\boldsymbol{\alpha}} \|\mathbf{\Phi}\boldsymbol{\alpha} - \boldsymbol{y}\|_2^2 + n\lambda\mathcal{S}(\boldsymbol{\alpha}) \tag{2}$$

Tikhonov Regularization

Impose Gaussian Prior on the Weights

$$\boldsymbol{\alpha} \sim \mathcal{N}(0, \boldsymbol{\Gamma}^{-1}) \tag{3}$$

$$\mathcal{S}(\boldsymbol{\alpha}) = \|\boldsymbol{\Gamma}\boldsymbol{\alpha}\|_2^2 \tag{4}$$

Improved Tikhonov matrix:

$$\boldsymbol{\Gamma}_{i,i} = 4^{|\boldsymbol{l}|_1 - d} \tag{5}$$

							1/64							
			1/64				1/16				1/64			
							1/64							
	1/64		1/16		1/64		1/4		1/64		1/16		1/64	
							1/64							
			1/64				1/16				1/64			
							1/64							
1/64	1/16	1/64	1/4	1/64	1/16	1/64	1	1/64	1/16	1/64	1/4	1/64	1/16	1/64
							1/64							
			1/64				1/16				1/64			
							1/64							
	1/64		1/16		1/64		1/4		1/64		1/16		1/64	
							1/64							
			1/64				1/16				1/64			
							1/64							

Figure 3: The improved prior

Tikhonov Regularization: Results Concrete

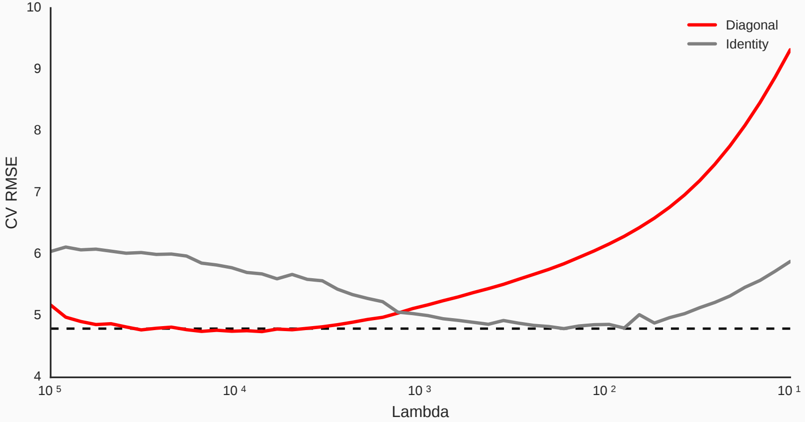


Figure 4: Results for the concrete dataset obtained with estimators with level five for two different Tikhonov matrices

Tikhonov Regularization: Results Power Plant

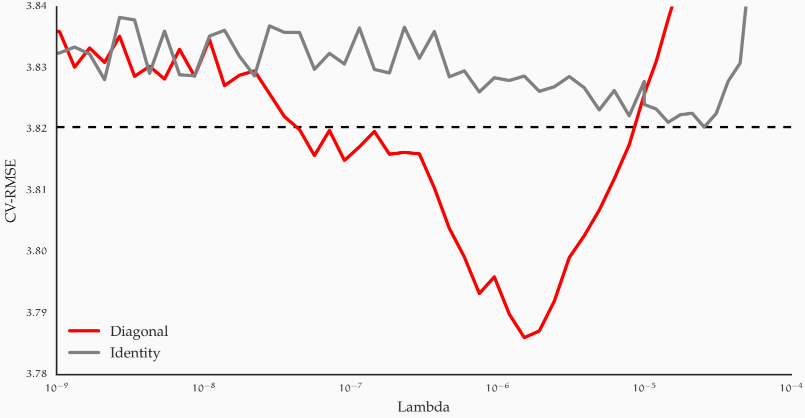
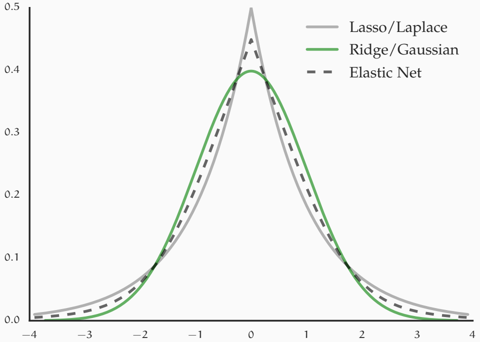


Figure 5: Results for the power plant dataset obtained with estimators with level five for two different Tikhonov matrices

Sparsity-inducing penalty.

Sparsity: Weight vector $\boldsymbol{\alpha}$, some entries are exactly zero.

$$\mathcal{S}(\boldsymbol{\alpha}) = \|\boldsymbol{\alpha}\|_1 \tag{6}$$



Sparsity-inducing Penalties: Elastic Net

Combination of Tikhonov and lasso regularization:

$$\mathcal{S}(\boldsymbol{\alpha}) = (1-\theta)\,\|\boldsymbol{\alpha}\|_2 + \theta\|\boldsymbol{\alpha}\|_1 \tag{7}$$

Improved performance for highly-correlated features.

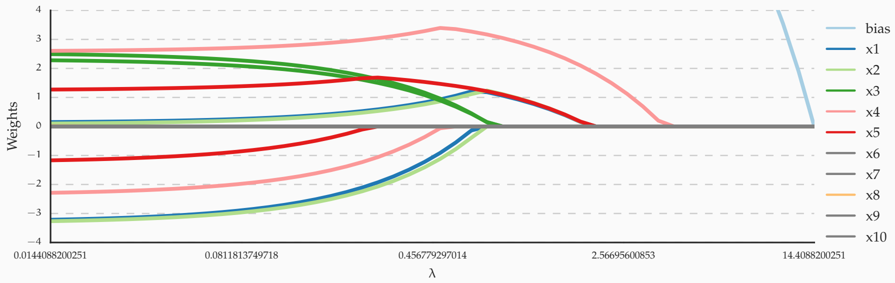
Sparsity-inducing Penalties: Group Lasso

With $\mathcal{P}$ as a partition of $\boldsymbol{\alpha}$:

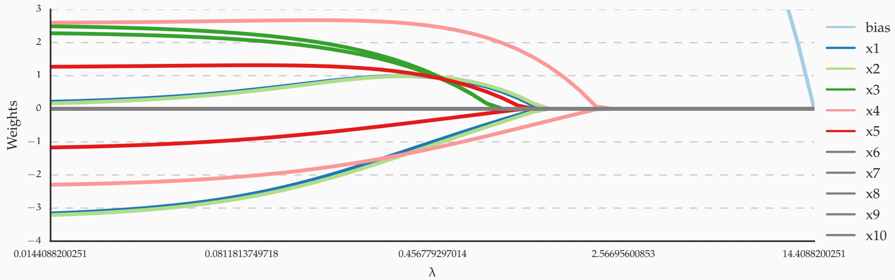
$$\mathcal{S}(\boldsymbol{\alpha}) = \sum_{\boldsymbol{p} \in \mathcal{P}} \left(\sqrt{|\boldsymbol{p}|} \right) \|\boldsymbol{p}\|_2 \tag{8}$$

Group by order:

$$\text{order}(\boldsymbol{p}) = |\{i \mid p_i \neq 0.5\}| \tag{9}$$



4
3
2
1
0
−1
−2
−3
−4
Weights
0.0144088200251
0.0811813749718
0.456779297014
2.56695600853
14.4088200251
λ
bias
x1
x2
x3
x4
x5
x6
x7
x8
x9
x10



3
2
1
0
−1
−2
−3
−4
Weights
0.0144088200251
0.0811813749718
0.456779297014
2.56695600853
14.4088200251
λ
bias
x1
x2
x3
x4
x5
x6
x7
x8
x9
x10

Sparsity-inducing Penalties: Results (1)

Reg. Method	CV-RMSE
Lasso	4.582
Elastic Net (θ = 0.95)	4.594
Group Lasso	4.650
Ridge	4.709

Table 1: Results for the concrete dataset for level five.

Akaike Information criterion

Asymptotically equivalent with leave-one-out cross-validation.

$$\mathrm{Aic}(\mathrm{df}, \mathrm{mse}) = 2\,\mathrm{df} + n \ln(\mathrm{mse}) + \text{constant} \tag{10}$$

Df. are degrees of freedom. For Lasso:

$$\mathrm{df}(\boldsymbol{\Phi}, \boldsymbol{\alpha}) = \mathrm{rank}(\boldsymbol{\Phi}_{\boldsymbol{\mathcal{A}}(\boldsymbol{\alpha})})$$

with

$$\boldsymbol{\mathcal{A}}(\boldsymbol{\alpha}) = \{a \in \boldsymbol{\alpha} \,|\, a \neq 0\}$$

For ridge:

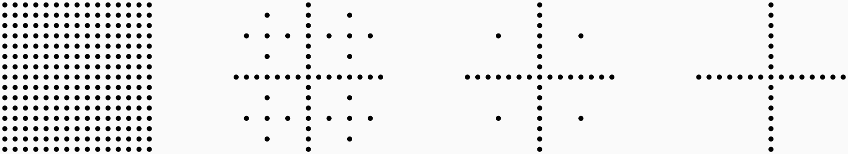
$$\mathrm{df}(\boldsymbol{\Phi}, \lambda) = \sum_i \frac{\sigma_i^2 - \lambda}{\sigma_i^2}$$

Sparsity-inducing Penalties: Results (2)

Level	Reg. Method	Gridsize	DF	AIC	Test-RMSE
5	Ridge	6650	716.7	2796.22	4.184
4	Lasso	1382	518.0	2797.09	3.850
4	Ridge	1470	558.0	2991.22	4.198
5	Lasso	6632	754.0	2997.94	3.737

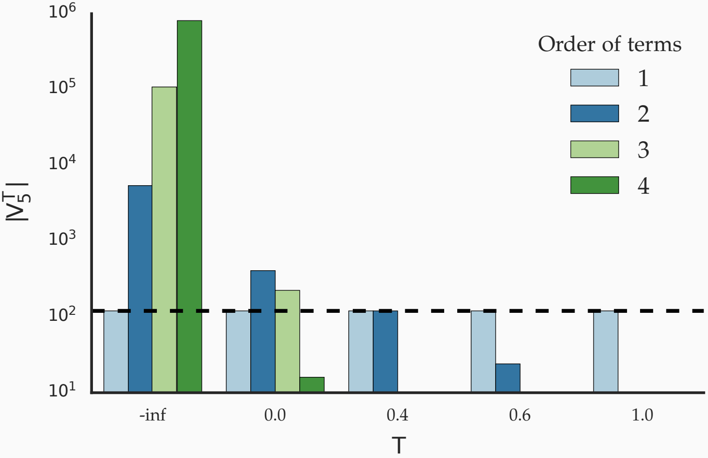
Generalized Sparse Grids

New hyper-parameter T, controls the granularity of the grid



$T = -\infty$ $T = 0$ $T = 0.5$ $T = 1$

Generalised Sparse Grids: Interaction terms



Generalized Sparse Grids: Results

T	Level	Gridsize	Root mean error
0	5	6650	4.184
0	4	1470	4.198
0.5	5	1180	3.797

Table 2: Errors of generalized grids for the concrete dataset.

Interaction-Term aware Sparse Grids

Idea: Only include some interaction terms.

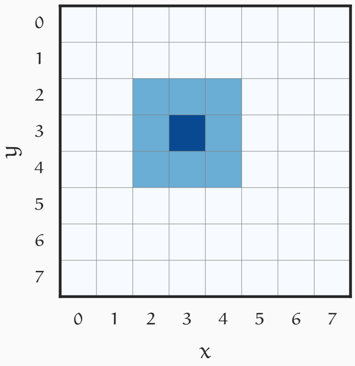


Figure 7: 3 × 3 Neighborhood

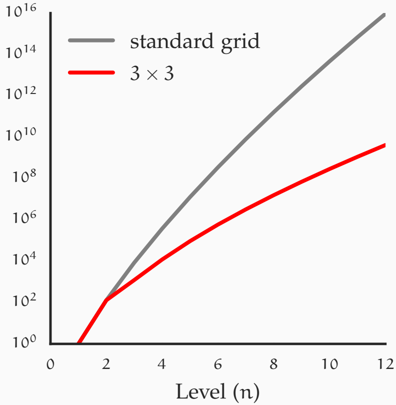


Figure 8: Effect on grid size

Optical Digits

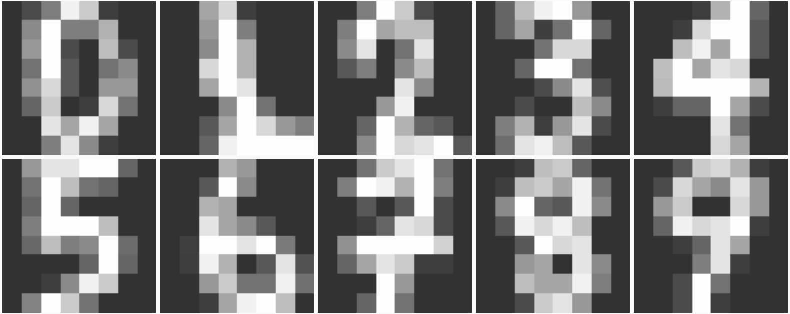


Figure 9: Digits

Optical Digits: Results

Sparse Grid Method	Level	Neighbors	Gridsize	Accuracy[%]
Standard	2	all	129	92.77
Interaction-Aware	3	3×3	1225	97.33
Adaptive	2	all	1760	97.74
Interaction-Aware	3	5×5	2569	97.83
Standard	3	all	8449	98.22

Table 3: Accuracy of sparse grids models for the optical digits dataset.

Conclusion

- Competitive results
- Prior knowledge $\rightarrow$ better & more effective solutions
- Only mild assumptions