Approximating Combined Discrete Total Variation and Correlated Sparsity With Convex Relaxations

Eugene Belilovsky
École Centrale Paris
eugene.belilovsky@ecp.fr

Andreas Argyriou
École Centrale Paris
andreas.argyriou@ecp.fr

Matthew Blaschko
Inria & École Centrale Paris
matthew.blaschko@inria.fr

Abstract

The recently introduced k-support norm has been successfully applied to sparse prediction problems with correlated features. This norm however lacks any explicit structural constraints commonly found in machine learning and image processing. We address this problem by incorporating a total variation penalty in the k-support framework. We introduce the (k, s) *support total variation norm* as the *tightest* convex relaxation of the intersection of a set of discrete sparsity and total variation penalties. We show that this norm leads to an intractable combinatorial graph optimization problem, which we prove to be NP-hard. We then introduce a tractable relaxation with approximation guarantees. We demonstrate the effectiveness of this penalty on classification in the low-sample regime, M/EEG neuroimaging analysis, and background subtracted image recovery.

1 Introduction

Regularization methods utilizing the ℓ_1 norm such as Lasso [1] have been used widely for feature selection. They have been particularly successful at learning problems in which very sparse models are required. However, in many problems a better approach is to balance sparsity against an ℓ_2 constraint. One reason is that very often features are correlated and it may be better to combine several correlated features than to select fewer of them, in order to obtain a lower variance estimator and better interpretability. This has led to the method of *elastic net* in statistics [2], which regularizes with a weighted sum of ℓ_1 and ℓ_2 penalties. More recently, it has been noted that the elastic net is not in fact the tightest convex penalty that approximates sparsity (ℓ_0) and ℓ_2 constraints at the same time [3]. The tightest convex penalty is given by the k-support norm, which is parameterized by an integer k, and can be computed efficiently. This norm has been succesfully applied to a variety of sparse vector prediction problems [4, 5, 6].

We study the problem of introducing further structural constraints to sparsity and ℓ_2. In particular, we seek to introduce a total variation smoothness prior in addition to sparsity and ℓ_2 constraints. Total variation is a popular regularizer used to enforce local smoothness in a signal [7, 8, 9]. It has succesfully been applied in image denoising and has recently become of particular interest in the neural imaging community where it can be used to reconstruct sparse but locally smooth brain activation [8, 10].

To derive a penalty incorporating these constraints we follow the approach of [3] by taking the convex hull of the intersection of our desired penalties and then recovering a norm by applying the gauge function. We then derive a formulation for the dual norm which leads us to a combinatorial optimization problem, which we prove to be NP-hard. We find an approximation to this penalty and prove a bound on the approximation error. Since the k-support norm is the tightest relaxation of sparsity and ℓ_2 constraints, we propose to use the intersection of the TV norm ball and the k-support norm ball. This leads to a convex optimization problem in which (sub)gradient computation can be achieved with a computational complexity no worse than that of the total variation. Furthermore, our approximation can be computed for variation on an arbitrary graph structure.

We demonstrate the tractability and utility of the norm through applications to M/EEG neuroimaging analysis, classification in the low-sample regime, and background subtracted image recovery.

2 Convex Relaxation of Sparsity, ℓ_2 and Total Variation

In this section we formulate the (k, s) *support total variation* norm, a tight convex relaxation of sparsity, ℓ_2, and total variation (TV) constraints.

Derivation of the Norm We start by defining the set of points corresponding to simultaneous sparsity, ℓ_2 and total variation (TV) constraints:

$$Q^2_{k,s} := \{w \in \mathbb{R}^d : \|w\|_0 \leq k, \|w\|_2 \leq 1, \|Dw\|_0 \leq s\}$$

where $k \in \{1, \dots, d\}$, $s \in \{1, \dots, m\}$ and $D \in \mathbb{R}^{m \times d}$ is a prescribed matrix. Here D will generally take the form of a discrete difference operator. It is easy to see that the set $Q^2_{k,s}$ is not convex due to the presense of the $\|\cdot\|_0$ terms. Hence using $Q^2_{k,s}$ in a regularization method is impractical. Thus we consider instead the convex hull of $Q^2_{k,s}$:

$$\begin{aligned} C^2_{k,s} :&= \operatorname{conv}(Q^2_{k,s}) \\ &= \Big\{w : w = \sum_{i=1}^r c_i z_i, \sum_{i=1}^r c_i = 1, c_i \geq 0, z_i \in \mathbb{R}^d, \|z_i\|_0 \leq k, \|z_i\|_2 \leq 1, \|Dz_i\|_0 \leq s, r \in \mathbb{N}\Big\} \end{aligned}$$

The convex set $C^2_{k,s}$ is the unit ball of a certain norm. We call this norm the (k, s) *support total variation* norm. It equals the gauge function of $C^2_{k,s}$, that is,

$$\begin{aligned} \|x\|^{sptv}_{k,s} :&= \inf\Big\{\lambda \in \mathbb{R}_+ : x = \lambda \sum_{i=1}^r c_i z_i, \sum_{i=1}^r c_i = 1, c_i \geq 0, z_i \in \mathbb{R}^d, \|z_i\|_0 \leq k, \|z_i\|_2 \leq 1, \\ &\qquad \|Dz_i\|_0 \leq s, r \in \mathbb{N}\Big\} \\ &= \inf\Big\{\sum_{i=1}^r \|v_i\|_2 : \sum_{i=1}^r v_i = x, \|v_i\|_0 \leq k, \|Dv_i\|_0 \leq s, r \in \mathbb{N}\Big\} . \end{aligned}$$

The special case $s = m$ is simply the k-support norm [3], which trades off between the ℓ_1 norm ($k = 1, s = m$) and the ℓ_2 norm ($k = d, s = m$). This norm is combinatorial in nature and it can be shown that it leads to an NP-hard problem.

Approximating the Norm Although special cases where s equals m or 1 are tractable, it can be shown that the general case for arbitrary values of s leads to an NP-hard problem. We thus approximate the solution by taking instead the intersection of the k-support norm ball and the convex relaxation of total variation. This leads to the following penalty

$$\Omega_{sptv}(w) = \max\{\|w\|^{sp}_k, \tfrac{1}{\sqrt{s}\|D\|}\|Dw\|_1\}$$

where $\|\cdot\|$ denotes the spectral norm. It can be shown that, given appropriate parameter selection, the solution to a regularized risk function with this penalty will be equivalent to regularization with a penalty of the form

$$\Omega_{k+tv}(w) = \lambda_1\|w\|^{sp}_k + \lambda_2\|Dw\|_1$$

for some regularization parameters $\lambda_1, \lambda_2 > 0$.[1] These objectives can be optimized using standard first-order convex optimization techniques such as subgradient descent or Nesterov smoothing methods [11, 12]. We can bound the approximation error as follows,

Proposition 1. *For every $w \in \mathbb{R}^d$, it holds that*

$$\Omega_{sptv}(w) \leq \|w\|_{k,s}^{sptv} .$$

Moreover, suppose that $\operatorname{range}(D^\top) = \mathbb{R}^d$ *and that for every $I \in G_k$ the submatrix D_{*I} has at least $m - s$ zero rows. Then it holds that*

$$\|w\|_{k,s}^{sptv} \leq \sqrt{1 + \frac{s\|D\|^2 \|(D^\top)^+\|_\infty^2}{k}}\, \Omega_{sptv}(w)$$

where $\|\cdot\|_\infty$ is the norm on $\mathbb{R}^{m \times d}$ induced by ℓ_∞, that is, $\|A\|_\infty = \max\limits_{i=1}^{m} \sum\limits_{j=1}^{d} |A_{ij}|$.

The hypothesis on D required for the upper bound on $\|\cdot\|_{k,s}^{sptv}$ has an intuitive interpretation when D is the transpose of an incidence matrix of a graph. It means that any group of k vertices in the graph involves at most s edges. This is true in many cases of interest, such as the grid if s is proportional to k. Thus the hypothesis is satisfied when k is not too large and the degrees of the vertices are relatively small. Finally, we note that frequently there will be some dependence on the dimensionality d, implicit in the term $\|(D^\top)^+\|_\infty$.

3 Experimental Results

We evaluate the effectiveness of the introduced penalty on several regression and classification problems. We consider a regression task with synthetic data set with spatial coherence and sparsity, a small training sample classification task using MNIST, a sparse image recovery problem, and an M/EEG prediction task.

We compare our regularizer against several basic regularizers (ℓ_1 and ℓ_2) and popular structured regularizers for problems with similar structure. In recent work TV+ℓ_1, which adds the TV and ℓ_1 constraints, has been heavily utilized for data with similar spatial assumptions and is thus one of our main benchmarks.

Synthetic 2-D Spatially Correlated Data For our synthetic problem we consider the estimation of an ideal weight vector with both spatial correlation and sparsity. We construct a 25x25 image with 84 % of coefficients set to zero. The non-sparse portion of the image correpond to Gaussian blobs. This image will serve as a set of parameters w we wish to recover. Figure 1 shows this ideal parameter vector. We construct data samples $X = Yw + \varepsilon$. Where Y is a sample from $\{-1, 1\}$ and $\varepsilon \sim \mathcal{N}$ is Gaussian noise.

We take 250 training samples, 100 validation samples, and 500 test samples. We evaluate a regularized risk function using only ℓ_1, ℓ_2, or k-support regularizers, TV+ℓ_1 regularizer, and the our k-support total variation regularizer. For each of these scenarios we perform model selection through the range of model parameters and select the model with the highest accuracy on the validation set. For this task we use hinge loss and perform optimization using stochastic subgradient descent. For the graph structure we use a grid graph with each pixel having a neighborhood consisting of the 4 adjacent pixels. We repeat this experiment with a new set of training, validation, and test samples 5 times so that we may obtain statistical significance results. The test set accuracy results for each method are shown in Table 1. For each competing method we perform a Wilcoxon signed-rank test against the k-support total variation results. In all listed cases the test rejects the null hypothesis (at a significance level of $p < 0.05$) that the samples come from the same distribution.

In some applications with spatially contigous data (e.g. the classification of registered images) the properties of the visualization afforded by the derived model parameters can be critical in providing

[1] The proof of this statement follows from the fact that optimization subject to the intersection of two constraints has a Lagrangian that is exactly a regularized risk minimization with the two corresponding penalties each with their own Lagrange multiplier.

Method	Acc. (p-value)	Stability
ℓ_2	68.12% (0.042)	96.7%
ℓ_1	70.84%(0.042)	54.8%
k-support	71.24%(0.043)	58.7%
TV+ℓ_1	79.80%(0.043)	89.5%
k-sup/TV	82.8	97.4%

Table 1: Average test accuracy results for 5 trials of synthetic data along with p-value for Wilcoxon hypothesis performed for each method against the k-support/TV result, below 0.05 for all cases. Pixel selection stability refers to the frequency of ground truth non-zero pixels remaining non-zero across all 5 trials

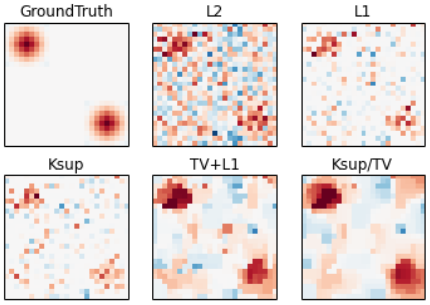

Figure 1: (*top left to bottom right*) ideal weight vector, weight vector obtained with ℓ_1,ℓ_2, k-support norm, TV+ℓ_1, and k-support/TV regularizer, and weight vector with combined total variation and k-support norm regularizer

users insight into the model. We can qualitatively evaluate this property of our sparse regularizer for data fulfilling the structure and sparsity characteristics which are hypothesised by the (k, s) support total variation norm: sparse with smooth contiguous regions.

In Figure 1 we can visualize the weight vector produced by the various regulazation methods, we can see that in Figure 1 the k-support norm alone does a poor job at reconstructing a model with any of these local correlations in place. The TV+ℓ_1-regularizer does a substantially better job at indicating the areas of interest for this task but the k-support/TV regularizer produces smoother regions (which are more easily interpertable) and substantially better classification accuracy.

We can see an additional advantage of the k-support total variation regularizer over the TV+ℓ_1 regularizer in terms of stability of the results. In applications where the support of the model parameters is used to interpret the model, methods which produce stable results in terms of selecting the same support on various folds of the data are preferable [6, 13]. Table 1 demonstrates the substantially improved selection stability for k-support total variation. Here we determine the average fraction of times a voxel is selected within the true active regions (based on the ground truth). We note that ℓ_2 constraints improve stability of the solution since they prevent correlated features from being discarded when searching for sparse solutions.

Low Sample Complexity MNIST Classification We consider a simple classification problem using the MNIST data set [14]. We select a very small subset of data to train with in order to demonstrate the effectiveness of our regularizer. We train a one versus all classifier for each digit. In the case of each digit we take 9 negative training samples, one from each other digit, and 9 positive training samples of the digit. We use a large validation set consisting of 8000 examples to perform parameter selection. We use a regularized risk function consisting of Ω_{k+tv} and logistic loss. We test on the entire MNIST test set of 10000 images. We optimize a logistic loss function combined with our k-support total variation norm and compare to results from ℓ_1, ℓ_2, k-support norm, and TV+ℓ_1 penalties combined with logistic loss. We perform optimization using FISTA on the k-support norm [3, 11] and a smoothing applied to the total variation. For the graph structure, specified by D, we use a grid graph with each pixel having a neighborhood consisting of the 4 adjacent pixels. We obtain surprisingly high classification accuracy using just 18 training examples. The results in Table 2 show classification accuracy for each one versus all classifier and the the average of the classifiers. In all but two case the k-support TV norm outperforms the other regularizers.

M/EEG Prediction We apply k-support total variation regularization to an M/EEG prediction problem from [15]. The data is preprocessed in the same way as in [15]. This results in data samples with 60 channels, each consisting of a time-series presumed to be independent across channels. Following [15] we use results only for subject 8 from this dataset. For the total variation

Class.	ℓ_1	ℓ_2	KS	ℓ_1+TV	KS+TV
D0	$93.62 \pm .01$	$93.49 \pm .01$	$93.68 \pm .02$	$96.22 \pm .01$	$\mathbf{96.27 \pm .01}$
D1	$90.1 \pm .02$	$89.56 \pm .02$	$90.08 \pm .02$	$90.57 \pm .02$	$\mathbf{92.18 \pm .02}$
D2	$78.28 \pm .03$	$77.28 \pm .03$	$78.25 \pm .03$	$\mathbf{81.47 \pm .02}$	$81.39 \pm .03$
D3	$68.58 \pm .02$	$68.05 \pm .02$	$68.60 \pm .02$	$71.63 \pm .02$	$\mathbf{73.25 \pm .02}$
D4	$83.81 \pm .01$	$82.55 \pm .01$	$83.76 \pm .01$	$84.69 \pm .01$	$\mathbf{84.79 \pm .01}$
D5	$73.7 \pm .03$	$73.2 \pm .02$	$73.69 \pm .03$	$74.52 \pm .02$	$\mathbf{74.95 \pm .02}$
D6	$93.48 \pm .01$	$93.37 \pm .01$	$93.51 \pm .01$	$93.71 \pm .01$	$\mathbf{94.08 \pm .01}$
D7	$88.88 \pm .02$	$87.21 \pm .02$	$88.85 \pm .02$	$91.67 \pm .01$	$\mathbf{92.59 \pm .01}$
D8	$70.79 \pm .02$	$72.07 \pm .03$	$72.75 \pm .02$	$73.23 \pm .02$	$\mathbf{73.10 \pm .02}$
D9	$85.48 \pm .02$	$\mathbf{85.61 \pm .02}$	$85.49 \pm .02$	$85.5 \pm .03$	$85.60 \pm .03$

Table 2: Accuracy for One versus All classifiers on MNIST using only 18 training examples and standard error computed on the test set

Classifier	Mean Acc.	Acc std.
SVM [15]	65.44%	2.29%
ksp-TV SVM	66.84%	3.42%
TV-L1 SVM	60.70%	4.66%

Table 3: Mean and standard deviation for SVM classifier, k-support total variation regularized SVM, and TV-L1 regularized SVM computed over 5 folds

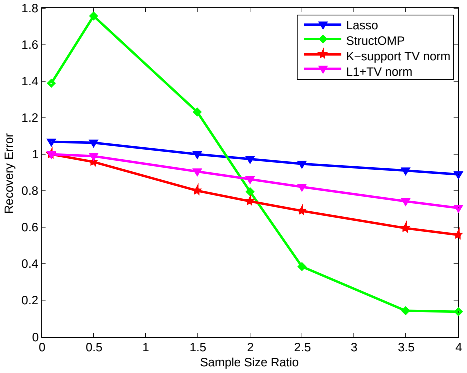

Figure 2: Average model error for background subtracted image reconstruction for various sample sizes

graph structure, we impose constraints for adjacent samples within each channel, while values from different channels are not connected within the graph. In the original work a latent variable SVM with delay parameter h is used to improve alignment of the samples. We consider only the case for $h = 0$, which reduces to the standard SVM. To directly compare our results we utilize hinge loss with a constant C of 2×10^4, the same regularization value used in [15]. Thus we optimize the following objective

$$R(w) = \frac{C}{N}\sum_{i=1}^{N} \max\{0, 1 - y_i\langle w, x_i\rangle\} + (1-\lambda)(\|w\|_k^{sp})^2 + \lambda\|Dw\|_1$$

Where λ allows us to easily tradeoff between k-support and total variation norms, while maintaining a fixed weight for our regularizer comparable to [15]. We use a k value of 2500 (approximately 80%) of the maximum and a λ parameter of 0.1. Table 3 shows the mean and standard deviation for the classification accuracy. We use the same partitioning of the data as described in [15], and obtain an improvement over the original results. We note that TV-L1 regularization has relatively poor performance. We hypothesize this is because the data used is very noisey and not extremely sparse.

Background Subtracted Image Recovery We apply k-support total variation regularization to a background subtracted image reconstruction problem frequently used in the structured sparsity literature [16, 17]. We use a similar setup to [16]. Here we apply m random projections to a background-subtracted image along with Gaussian noise, and reconstruct the image using the projections and projection matrices. Our evaluation metric for the recovery is the mean squared pixel error. For this experiment we utilize a FISTA technique with squared loss [3, 5] and a smoothed total variation term. To speed up convergence we make the smoothing parameter adaptive.

We selected 50 images from the background segmented dataset and converted them to grayscale. We use squared loss and k-support total variation to reconstruct the original images. We compute normalized recovery error for different number of samples m and compare our regularizer to LASSO, TV+ℓ_1, and StructOMP. The latter is a structured regularizer which performs best on this problem in [17]. The average normalized recovery error is shown for different sample sizes in Figure 2.

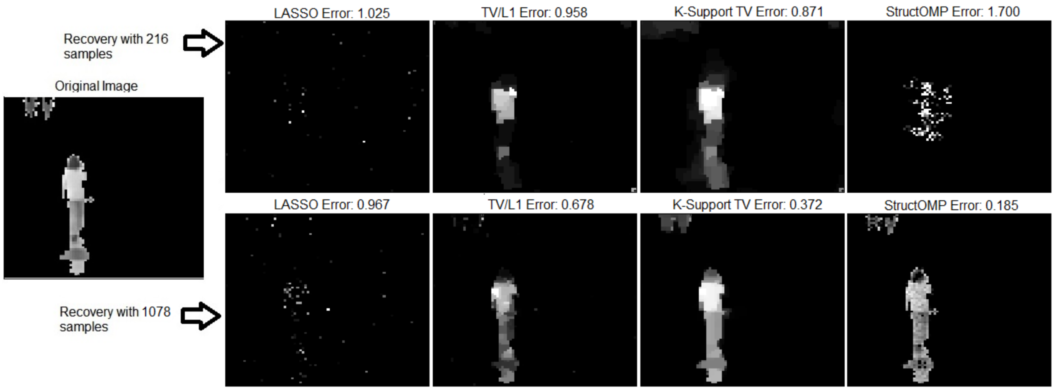

Figure 3: Background subtracted image recovery example for different methods and sample sizes

In terms of recovery error we note that k-support total variation substantially outperforms LASSO and TV+ℓ_1, and outperforms StructOMP for low sample complexity. Further examination of the images can reveal other advantages of the k-support total variation regularizer. An example for one image recovery scenario is shown at 2 different sample sizes in Figure 3. Here we can see that at low complexity StructOMP and LASSO can completely fail in terms of creating a visually coherent reconstruction of the image. TV+ℓ_1 recovery at the low sample size improves upon the latter methods, producing smooth regions, but still not resembling the human shape pictured in the original image. k-support total variation has better visual quality at this low sample complexity, due to its ability to retain multiple groups of correlated variables in addition to the smoothness prior. For the case of a larger number of samples, illustrated by the bottom row of Figure 3, we note that although the recovery performance of StructOMP is better (lower error), the visual quality of the k-support total variation regualrizer produces smoother and more coherent image segments.

4 Conclusions

We have introduced a novel norm that incorporates spatial smoothness and correlated sparsity. This norm, called the (k, s) *support total variation norm*, extends both the total variation penalty which is a standard in image processing and the recently proposed k-support norm from machine learning. The (k, s) support TV norm is the *tightest convex penalty* that combines *sparsity*, ℓ_2 and *total variation* constraints jointly. We have derived a variational form for this norm for arbitrary graph structures. We have also expressed the dual norm as a combinatorial optimization problem on the graph. This graph problem is shown to be NP-hard motivating the use of a relaxation, which is shown to be equivalent to the weighted combination of a k-support norm and a total variation penalty. We have shown that this norm approximates the (k, s) support TV norm within a factor that depends on properties of the graph as well as on the parameters k and s. Moreover, we have demonstrated that joint k support and TV regularization can be applied on a diverse variety of learning problems, such as classification with small samples, neural imaging and image recovery. These experiments have illustrated the utility of penalties combining k-support and total variation structure on problems where spatial structure, feature selection and correlations among features are all relevant.

Acknowledgements This work is partially funded by ERC Grant 259112, FP7-MC-CIG 334380, and DIGITEO 2013-0788D - SOPRANO. We would also like to thank Tianren Liu for his help with showing that computation of the (k, s) *support total variation norm* is an NP-hard problem.

References

[1] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society (Series B)*, vol. 58, pp. 267–288, 1996.

[2] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[3] A. Argyriou, R. Foygel, and N. Srebro, "Sparse prediction with the k-support norm.," in *NIPS* (P. L. Bartlett, F. C. N. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, eds.), pp. 1466–1474.

[4] K. Gkirtzou, J. Honorio, D. Samaras, R. Z. Goldstein, and M. B. Blaschko, "Fmri analysis of cocaine addiction using k-support sparsity," in *ISBI*, pp. 1078–1081, 2013.

[5] M. B. Blaschko, "A note on k-support norm regularized risk minimization," *CoRR*, vol. abs/1303.6390, 2013.

[6] M. Misyrlis, A. Konova, M. Blaschko, J. Honorio, N. Alia-Klein, R. Goldstein, D. Samaras, *et al.*, "Predicting cross-task behavioral variables from fmri data using the k-support norm," in *Sparsity Techniques in Medical Imaging (STMI)*, 2014.

[7] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Phys. D*, vol. 60, pp. 259–268, Nov. 1992.

[8] V. Michel, A. Gramfort, G. Varoquaux, E. Eger, and B. Thirion, "Total variation regularization for fmri-based prediction of behavior," *IEEE Trans. Med. Imaging*, vol. 30, no. 7, pp. 1328–1340, 2011.

[9] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight, "Sparsity and smoothness via the fused lasso," *Journal of the Royal Statistical Society Series B*, pp. 91–108, 2005.

[10] L. Baldassarre, J. Mourao-Miranda, and M. Pontil, "Structured sparsity models for brain decoding from fmri data," in *Pattern Recognition in NeuroImaging (PRNI), 2012 International Workshop on*, pp. 5–8, IEEE, 2012.

[11] Y. Nesterov, *Introductory lectures on convex optimization: A basic course*. Springer, 2004.

[12] Y. Nesterov, "Smooth minimization of non-smooth functions," *Mathematical Programming*, vol. 103, no. 1, pp. 127–152, 2005.

[13] S. Yan, X. Yang, C. Wu, Z. Zheng, and Y. Guo, "Balancing the stability and predictive performance for multivariate voxel selection in fmri study," in *Brain Informatics and Health - International Conference, BIH 2014, Warsaw, Poland, August 11-14, 2014, Proceedings*, pp. 90–99, 2014.

[14] Y. LeCun and C. Cortes, "MNIST handwritten digit database." http://yann.lecun.com/exdb/mnist/, 2010.

[15] W. Zaremba, M. P. Kumar, A. Gramfort, and M. B. Blaschko, "Learning from m/eeg data with variable brain activation delays," in *IPMI*, pp. 414–425, 2013.

[16] L. Baldassarre, J. Morales, A. Argyriou, and M. Pontil, "A general framework for structured sparsity via proximal optimization," in *AISTATS*, pp. 82–90, 2012.

[17] J. Huang, T. Zhang, and D. Metaxas, "Learning with structured sparsity," in *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, (New York, NY, USA), pp. 417–424, ACM, 2009.