Grouping Total Variation and Sparsity: Statistical Learning with Segmenting Penalties

Michael Eickenberg, Elvis Dohmatob, Bertrand Thirion, and Gaël Varoquaux

Inria Parietal, Neurospin, CEA Saclay, 91191 Gif-sur-Yvette
michael.eickenberg@nsup.org

Abstract. Prediction from medical images is a valuable aid to diagnosis. For instance, anatomical MR images can reveal certain disease conditions, while their functional counterparts can predict neuropsychiatric phenotypes. However, a physician will not rely on predictions by black-box models: understanding the anatomical or functional features that underpin decision is critical. Generally, the weight vectors of classifiers are not easily amenable to such an examination: Often there is no apparent structure. Indeed, this is not only a prediction task, but also an inverse problem that calls for adequate regularization. We address this challenge by introducing a convex region-selecting penalty. Our penalty combines total-variation regularization, enforcing spatial contiguity, and ℓ_1 regularization, enforcing sparsity, into one group: Voxels are either active with non-zero spatial derivative or zero with inactive spatial derivative. This leads to segmenting contiguous spatial regions (inside which the signal can vary freely) against a background of zeros. Such segmentation of medical images in a target-informed manner is an important analysis tool. On several prediction problems from brain MRI, the penalty shows good segmentation. Given the size of medical images, computational efficiency is key. Keeping this in mind, we contribute an efficient optimization scheme that brings significant computational gains.

1 Introduction

For certain pathologies, medical images carry weak indicators of external phenotype. For instance, in Magnetic Resonance images, a pattern of brain atrophy centered on the thalamus predicts the evolution in Alzheimer's disease [19]. Functional Magnetic Resonance Imaging (fMRI) can be used to infer subjects' behavioral state from their brain activity [11]. Machine learning methods can identify these biomarkers. With linear predictors, the weight vectors form spatial maps in the image domain. However, minimizing a prediction error gives little control on the corresponding maps. Indeed, the prediction problem is often an ill-posed inverse problem in the sense that there are less samples than features available: many different weight maps can generate exactly the same predictions. A choice among these candidates is implicitly taken by the estimator employed. In the empirical risk minimization framework, this choice is imposed via a penalty which favors maps according to certain criteria, interpretable as a "prior". Sparsity for

N. Navab et al. (Eds.): MICCAI 2015, Part I, LNCS 9349, pp. 685–693, 2015.
DOI: 10.1007/978-3-319-24553-9_84

instance, imposable in convex optimization via the ℓ_1 norm, is very useful as it selects a small number of voxels for the prediction. It has been widely used in medical imaging, from fMRI [21] to regularizing diffeomorphic registration [8].

However, imposing sparsity can often lead to less stable weight maps. Indeed, for images with high spatial correlations, adjacent voxels contain similar information and only one of them is needed for prediction. To counter this behavior, several estimators incorporate the notion of spatial contiguity in weight maps. For instance *GraphNet* [15,10,12] uses an ℓ_2 penalty on image gradients, to force adjacent voxels to have similar weights. An improvement upon this method is to impose sparsity on the spatial derivative [14], or to combine sparsity of the derivative with sparsity of the weights [1,9]. These penalties come with the mathematical property of positive homogeneity, which makes model selection easier. A drawback for these methods is that they favor flat or staircased weight maps, while one would tend to expect smooth variation within an active region.

Our goal is to detect spatially-contiguous patches in statistically estimated images and to inform their estimation of the image with these detections. Thus, our work bridges two fields: sparsity and segmentation.

Specifically, we are interested in a foreground segmentation problem: recovering small, non-zero predictive regions from a noisy background. However, in many applications, such as CT or medical imaging, the measurement process leads to strong correlations in columns of the design matrix corresponding to neighboring pixels, rendering recovery theorems non-applicable and sparse support estimation highly unstable.

The other body of literature that we draw from is that of segmentation, specifically convex variational approaches, as they can be expressed as penalties in a risk minimizer. A central aspect is the Chan-Vese functional [5] for segmentation that computes piecewise constant approximations. This variational formulation is not convex, but [16] have shown that good solutions can be achieved with a similar but convex functional, based on total variation (TV), *i.e.* the ℓ_1 norm of the image gradient. For our purposes, this approach is appealing, as the use TV as a regularizing penalty shows good properties for image denoising [17] or estimation in a linear model [4]. However, all these segmentation approaches model an object as a homogeneous constant-valued domain, thus washing out internal structure. Here, for foreground-background segmentation, we want to impose a constant structure on the background, but not in the selected image domain.

Our contribution is twofold: 1) We introduce a new penalty, *Sparse Variation*, which forces zeros on coordinates and spatial derivative jointly and smooth variations in spatially-contiguous active zones. 2) We present *FAASTA*, a novel optimization scheme for fast estimation up to a very high precision. Importantly, control on spatial maps requires solving the optimization to a tight tolerance [7]. We empirically evaluate *Sparse Variation* in regression and classification on fMRI and structural MRI data, comparing it to TV-ℓ_1 and GraphNet.

2 Sparse Variation: A New Spatially Regularizing Penalty

2.1 Penalized Regression: Problem Formulation and Prior Art

Penalized Generalized Linear Models. Let $X \in \mathbb{R}^{n \times p}$ be the design matrix and $y \in \mathbb{R}^n$ the prediction target, where $n, p \in \mathbb{N}$ are the number of samples and features. The weight vector w and the offset c are obtained by solving the optimization problem: $\arg\min_{w,c} \ell(Xw + c, y) + \Omega(w)$. Ω is the regularizer and ℓ the loss, typically a logistic loss for classification or a squared loss for regression.

Existing Regularizers. Two regularizers successfully applied to medical volume data are the GraphNet and TV-ℓ_1. In the following, ∇ will denote a finite differences spatial gradient operator acting upon an image. Generally, for a 3D grid of size $p = p_x p_y p_z$, we have $\nabla \in \mathbb{R}^{3p \times p}$. To write a function gradient, we will indicate the variable with respect to which it is calculated in subscript, e.g. "∇_w". $\|\cdot\|_2$ is the euclidean norm. For a partition $\mathcal{G}$ of coordinates the $\ell_{2,1}$ group norm is written $\|v\|_{2,1} = \sum_{g \in \mathcal{G}} \|v_g\|_2$. For all penalties, $\lambda > 0$ regulates the strength and $\rho \in [0, 1]$ is a parameter controlling the trade-off between coordinate sparsity and spatial regularity. GraphNet consists of the sum of an ℓ_1 penalty on all coordinates and a squared ℓ_2 penalty on the spatial gradient, whereas TV-ℓ_1 is the sum of an ℓ_1 penalty and an $\ell_{2,1}$ group penalty on the spatial derivative:

$$\Omega_{\mathrm{GN}}(w) = \lambda((1-\rho)\|\nabla w\|_2^2 + \rho\|w\|_1)$$
$$\Omega_{\mathrm{TV}-\ell_1}(w) = \lambda((1-\rho)\|\nabla w\|_{2,1} + \rho\|w\|_1),$$

2.2 A New Penalty for Segmentation Purposes: Sparse Variation

We propose a new penalty, *Sparse Variation*, which enforces contiguous zones of smooth activation against a background of zeros. Indeed, in TV-ℓ_1, the penalties for sparsity of the signal and sparsity of the gradient are separable: they can be active and inactive independently. A non-zero constant block, for example, is active for the ℓ_1 penalty, but inactive for the gradient, except at the borders. This property induces step functions and blockiness where one would expect smoothness. We address this issue in *Sparse Variation* by grouping coordinate activation with spatial derivative activation: Either a coordinate is active (nonzero) and its derivative is active as well - allowing for smooth variation in active zones - or both are inactive (zero). We define the *Sparse Variation* penalty as

$$\Omega_{\mathrm{SV}}(w) = \lambda\|Kw\|_{2,1}, \qquad \text{where} \quad K = \begin{pmatrix} (1-\rho)\nabla \\ \rho\,\mathrm{Id}_p \end{pmatrix}, \tag{1}$$

with Id_p the $p \times p$ identity matrix. For 3D grids, $K \in \mathbb{R}^{4p \times p}$. The $\ell_{2,1}$ norm consists of groups containing the coordinate and all derivatives at each coordinate.

2.3 Optimization Strategy

All optimization problems mentioned in this manuscript - GraphNet, TV-ℓ_1 and *Sparse Variation*, with either the logistic loss or the squared loss - have a similar global structure: a sum of two convex functions, one being smooth, that we write F, the other nonsmooth, G. This structure can be exploited in proximal splitting algorithms [6], of which we contribute a new optimized variant. These algorithms rely on an implicit subgradient step in the non-smooth function called the *proximal operator* $\mathrm{prox}_{tG}(y) := \arg\min_x \frac{1}{2t}\|y - x\|_2^2 + G(x)$.

The simplest method is the Iterative Shrinkage-Thresholding Algorithm (ISTA) [6]. It amounts to iterations of $w_{k+1} = \mathrm{prox}_{\frac{1}{L}G}\left(w_k - \frac{1}{L}\nabla_w F(w_k)\right)$, where L is the Lipschitz constant of $\nabla_w F$. To accelerate convergence, the *fast iterative shrinkage-thresholding algorithm* (fISTA) [3] adds a momentum term: the gradient steps are applied to a combination of w_k and w_{k-1}. The acceleration brought by this method comes at the cost that there is no guarantee that each step of fISTA decreases the objective function and large rebounds are common. This non-monotone behaviour can be remedied by switching to ISTA iterations whenever an increase in cost is detected, as in *monotone fISTA* (mfISTA) [2].

There is no closed-form expression for proximal operators for TV-ℓ_1 and *Sparse-Variation*: they must be solved with a second, "inner" optimization problem. Both penalties can be written as $\lambda\|K \cdot\|$ for an appropriate norm $\|\cdot\|$. The projected-gradient algorithm used in [2] for TV denoising can then be adapted to iteratively solve the proximal operator with control of the dual gap.

Fast Adaptively Accurate Shrinkage Thresholding Algorithm. Importantly, solving $\mathrm{prox}_{G/L}$ numerically is an inexact operation, which can easily prevent convergence of the outer loop. However, proximal algorithms converge if the error on $\mathrm{prox}_{G/L}$ decreases sufficiently with the iteration number k of the outer loop [18]. Accuracy can be captured by the dual gap value. Instead of using a fixed dual gap refinement strategy, we devise an adaptive method, increasing accuracy (*dgtol*) as needed, if the energy $\mathcal{L}$ increases during an $ISTA$ step (Alg. 1).

3 Empirical Results

3.1 A Simple 1D Signal Recovery Problem

To develop intuitions, we study a 1D recovery problem with simulated data. We mimic spectroscopy settings: a signal with a spectrum on a small spatially-contiguous support is measured with additive noise. The spectrum is recovered via an inverse problem with a discrete cosine transform operator. Measurements are given by $y = X_{\mathrm{DCT}}^{-1} w + \varepsilon$, where X_{DCT} is the DCT operator, w the spectrum and ε Gaussian noise of 40% signal norm. We use w of size 200, with 80% zeros and an activated region resembling that of a chemical

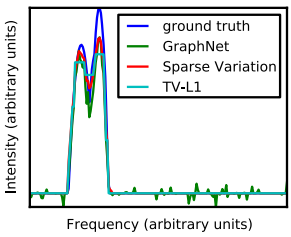

Fig. 1. Recovery for 1D spectroscopy. Note the blocky nature of the TV-ℓ_1 solution, and the noise in the GraphNet estimation.

Algorithm 1. fAASTA

Data: w_0
$ISTA \leftarrow False, \quad v_1 \leftarrow w_0, \quad k \leftarrow 0, \quad t_1 \leftarrow 1, \quad dgtol \leftarrow 0.1;$
while *not converged* **do**
  $k \leftarrow k+1, \quad w_k \leftarrow \text{prox}_{G/L}(v_k - (1/L)\nabla F(v_k), dgtol);$
  if $\mathcal{L}(w_k) > \mathcal{L}(w_{k-1})$ **then**
    $w_k \leftarrow w_{k-1}, \quad v_k \leftarrow w_{k-1};$
    if *ISTA* **then**
      $dgtol \leftarrow dgtol/2;$
      while $\mathcal{L}(\text{prox}_{G/L}(v_k - (1/L)\nabla_w F(v_k), dgtol)) > \mathcal{L}(w_{k-1})$ **do**
        $dgtol \leftarrow dgtol/2$
    $ISTA \leftarrow True;$
  else
    if *ISTA* **then**
      $v_k \leftarrow w_k, \quad ISTA \leftarrow False$
    else
      $t_k \leftarrow \frac{1+\sqrt{1+4t_{k-1}^2}}{2}, \quad v_k \leftarrow w_k + \frac{t_{k-1}-1}{t_k}(w_k - w_{k-1});$

compound signature: two overlapping smooth peaks. Fig. 1 shows the ground truth and the best recovery results: We selected the λ, ρ parameters minimizing ℓ_2 error with the ground truth. By construction, TV-ℓ_1 promotes flat signals, whereas *Sparse Variation* recovers better the smooth nature of the signal.

3.2 Segmenting Regions from MRI Data

We run experiments in both fMRI and structural MRI as well as both regression and classification settings. We compute prediction for the target variable from brain images over a full parameter grid λ, ρ. For each regularizer, weight maps of the best performing parameters in cross-validation are shown.

Classification: Intra-subject Object Recognition Study. The human ventral temporal cortex exhibits specialization to recurrent concepts such as faces, but also other object categories. We revisit the data from a seminal publication on this topic [11]: responses to visual stimuli of different categories. We test two classic contrasts, *faces versus houses* and *objects versus scramble*, with the logistic loss. Maps for optimal parameters overall detect similar regions. The top row of Fig. 2 shows the segmented right Fusiform Face Area. TV-ℓ_1 and *Sparse Variation* detect similar region size, whereas GraphNet selects a stronger sparsity. On the right, an F-statistic indicates extents of regions correlated to the stimuli. The bottom row shows the mapping of the Lateral Occipital Complex (LOC). *Sparse Variation* selects larger regions than the other two penalties. The focality of the maps is due to the single subject nature of the experiment.

Regression: Inter-subject Gain Prediction in Gambling Task. For linear regression in a multi-subject setting, we examine an fMRI experiment with gambles with varying gains [20]. Here we predict the gain of a given gamble from

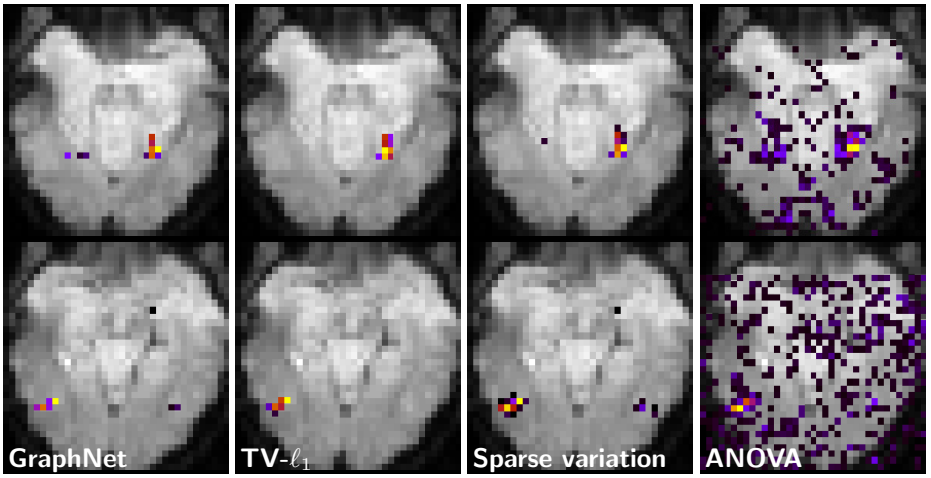

Fig. 2. Weight maps obtained from discrimination tasks between two visual concepts on data from [11]. **Top**: **FFA** (Fusiform Face Area) segmented in a face vs house discrimination. Cut at $z = -20$mm. Accuracies on held-out data: GN: 95.5%, TV-ℓ_1: 96.6%, SV: 97.7% **Bottom**: **LOC** (Lateral Occipital Complex) segmented in an object vs scramble discrimination. In this intra-subject analysis the maps are very well localized. Cut at $z = -16$mm. Accuracies: GN: 78.8%, TV-ℓ_1: 80.0%, SV: 80.0%

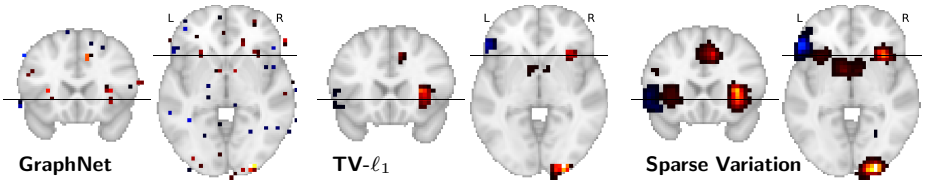

Fig. 3. Weight vectors from estimating gain on the mixed gambles task [20]. This inter-subject analysis shows broader regions of activation. Mean correlation scores on held out data: GN: 0.128, TV-ℓ_1: 0.147, SV: 0.149

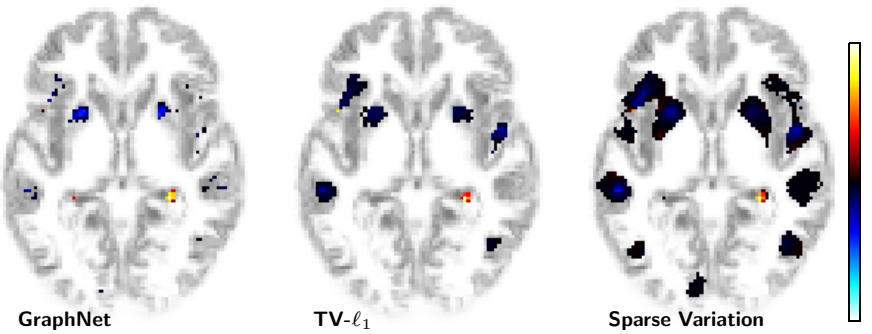

Fig. 4. Weight vectors for age prediction from VBM maps from the Oasis dataset. *Sparse Variation* selects clearly defined regions which are easily amenable to further analysis. Mean correlation scores on held-out data: GN:0.805, TV-ℓ_1:0.793, SV:0.794

the fMRI activation. At fixed $\rho = 0.5$, we evaluated the regularizers on a grid of λ. The weight maps of the best predicting parameters are shown in Fig. 3. At optimal predictive power the weight maps of TV-ℓ_1 and *Sparse Variation* show spatial contiguity and activation in expected regions, whereas GraphNet weights are scattered. The main distinction between TV-ℓ_1 and *Sparse Variation* is the "smoothness or zero" enforced by the latter in comparison to more blocky activations for the former. Larger activated regions do justice to the multi-subject setting. Note the segmentation of the Insulae, mentioned in the original study.

Regression: Estimating Age from Voxel-Based Morphometry (VBM) The Oasis database contains anatomical MRI for 400 subjects [13]. We extracted VBM images and used the different regularizers in a regression to estimate subjects' age. Fig. 4 shows the resulting weight maps. All identify the putamen, insula and para-hippocampal regions. TV-ℓ_1 selects contiguous regions where GraphNet finds sparse clouds of voxels. *Sparse Variation* segments smoother versions of regions selected by TV-ℓ_1, as well as several additional regions.

3.3 Optimization Speed of FASTAA

In data analysis optimization speed is important, practitioners may often decide to use less accurate but faster methods. We compare the adaptive refinement of the tolerance in FAASTA to others approaches: setting the dual gap tolerance to a constant, one strict (10^{-10}), one lax (0.1), and the refinement strategy of [18] (decrease dual gap as k^{-4}). We also compare to using ISTA in the outer loop in a constant dual gap (0.1) or an adaptive refinement setting.

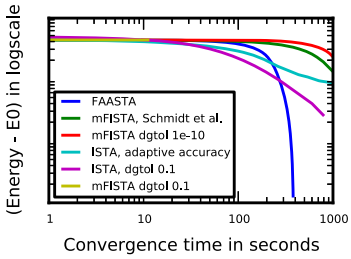

Fig. 5. Convergence on *object vs scramble*. FAASTA converges in 7mn, whereas other methods take more than 15mn

The results on Fig. 5 are striking. While the adaptive strategy always provides enough dual gap accuracy to ensure energy descent, the technique from [18] quickly becomes too strict, slowing convergence. Using a lax dual gap or the adaptive method with ISTA stalls at insufficient accuracy rates. The proposed adaptive method provides by far the fastest convergence.

Conclusion. We introduced a new region-selecting sparse convex penalty, *Sparse Variation*. It forces large regions of an image to zero, but, unlike prior art, allows smooth variation within spatially-contiguous active zones. On three brain imaging problems, this penalty shows best region segmenting properties with respect to prior art. Good convergence of the associated optimization problem is crucial to obtain reliable spatial maps. As with TV regularization, the optimization procedure necessitates an inner optimization to evaluate the proximal operator. A line-search strategy on dual gap tolerance is employed to refine the tolerance only as

much as needed for fast convergence. Compared to other schemes, our method converges fastest. In conclusion, *Sparse Variation* with *fAASTA* is the optimal choice for segmentation of medical images in a target-informed manner.

Acknowledgements. This work was funded by the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 604102 ("Human Brain Project").

References

1. Baldassarre, L., Mourao-Miranda, J., Pontil, M.: Structured sparsity models for brain decoding from fMRI data. In: PRNI, p. 5 (2012)
2. Beck, A., Teboulle, M.: Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems. IEEE Trans. Image Proc. (2009)
3. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm with application to linear inverse problems. SIAM 2, 183–202 (2009)
4. Candes, E., Romberg, J.: Signal recovery from random projections. In: Wavelet Applications in Signal and Image Processing XI, SPIE. vol. 5674, p. 76 (2005)
5. Chan, T.F., Vese, L.A.: Active contours without edges. IEEE Transactions on Image Processing 10, 266 (2001)
6. Combettes, P.L., Pesquet, J.C.: Proximal splitting methods in signal processing. In: Fixed-Point Algorithms for Inverse Problems in Science and Engineering (2011)
7. Dohmatob, E., Gramfort, A., Thirion, B., Varoquaux, G.: Benchmarking solvers for TV-l1 least-squares and logistic regression in brain imaging. PRNI (2014)
8. Durrleman, S., Prastawa, M., Gerig, G., Joshi, S.: Optimal data-driven sparse parameterization of diffeomorphisms for population analysis. In: IPMI, p. 123 (2011)
9. Gramfort, A., Thirion, B., Varoquaux, G.: Identifying predictive regions from fMRI with TV-L1 prior. In: PRNI, pp. 17–20 (2013)
10. Grosenick, L., Klingenberg, B., Katovich, K., et al.: Interpretable whole-brain prediction analysis with graphnet. NeuroImage 72, 304 (2013)
11. Haxby, J., Gobbini, I., Furey, M., et al.: Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science 293, 2425 (2001)
12. Kandel, B.M., Wolk, D.A., Gee, J.C., Avants, B.: Predicting cognitive data from medical images using sparse linear regression. In: IPMI, p. 86 (2013)
13. Marcus, D.S., Wang, T.H., Parker, J., et al.: Open access series of imaging studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. J. Cogn. Neurosci. 19, 1498 (2007)
14. Michel, V., Gramfort, A., Varoquaux, G.: other: Total variation regularization for fMRI-based prediction of behavior. IEEE Trans. Med. Im. 30, 1328 (2011)
15. Ng, B., Vahdat, A., Hamarneh, G., et al.: Generalized sparse classifiers for decoding cognitive states in fMRI. Machine Learning in Medical Imaging, p. 108 (2010)
16. Pock, T., Chambolle, A., Cremers, D., Bischof, H.: A convex relaxation approach for computing minimal partitions. In: CVPR, p. 810 (2009)

17. Rudin, L.I., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. Physica D: Nonlinear Phenomena 60, 259 (1992)
18. Schmidt, M., Roux, N.L., Bach, F.R.: Convergence rates of inexact proximal-gradient methods for convex optimization. In: NIPS, p. 1458 (2011)
19. Stonnington, C., Chu, C., Klöppel, S., et al.: Predicting clinical scores from magnetic resonance scans in alzheimer's disease. Neuroimage 51, 1405 (2010)
20. Tom, S.M., Fox, C.R., Trepel, C., Poldrack, R.A.: The neural basis of loss aversion in decision-making under risk. Science 315(5811), 515–518 (2007)
21. Yamashita, O., Sato, M.A., Yoshioka, T., et al.: Sparse estimation automatically selects voxels relevant for the decoding of fMRI activity patterns. NeuroImage 42, 1414 (2008)