HYPERPARAMETER ESTIMATION IN MAXIMUM A POSTERIORI REGRESSION USING GROUP SPARSITY WITH AN APPLICATION TO BRAIN IMAGING

Yousra Bekhti, Roland Badeau

LTCI, Télécom ParisTech,
Université Paris-Saclay
75013, Paris, France

Alexandre Gramfort

LTCI, Télécom ParisTech, Université Paris-Saclay
INRIA, Université Paris-Saclay
Paris, France

ABSTRACT

Hyperparameter estimation is a recurrent problem in the signal and statistics literature. Popular strategies are cross-validation or Bayesian inference, yet it remains an active topic of research in order to offer better or faster algorithms. The models considered here are sparse regression models with convex or non-convex group-Lasso-like penalties. Following the recent work of *Pereyra et al.* [1] we study the fixed point iteration algorithm they propose and show that, while it may be suitable for an analysis prior, it suffers from limitations when using high-dimensional sparse synthesis models. The first contribution of this paper is to show how to overcome this issue. Secondly, we demonstrate how one can extend the model to estimate a vector of regularization parameters. We illustrate this on models with group sparsity reporting improved support recovery and reduced amplitude bias on the estimated coefficients. This approach is compared with an alternative method that uses a single parameter but a non-convex penalty. Results are presented on simulations and an inverse problem relevant for neuroscience which is the localization of brain activations using magneto/electroencephalography.

1. INTRODUCTION

Hyperparameter setting is a classical statistics problem for which a number of solutions have been proposed. In signal processing, the AIC and BIC criteria are quite popular techniques historically [2]. The SURE-based techniques [3] have also been quite popular and recently explored for denoising and compressed sensing applications [4, 5]. In a standard supervised machine learning setup with independent and identically distributed (i.i.d.) observations, cross-validation (CV) is the reference approach. Also, the Bayesian approach suited for probabilistic models offers a principled way to estimate hyperparameters using hyperpriors that introduce softer constraints than solutions with fixed parameter values. This benefit yet usually comes at a price in terms of computational cost. Finally, in a number of real scenarios, humans end up setting hyperparameters, as they can have some expert knowledge that can correct model mismatch.

In statistical machine learning an hyperparameter typically aims at limiting overfitting by controlling the model complexity. In the particular case of regularized regression, classically a scalar parameter balances between the data fit and the penalty term. When using sparse regression, this parameter affects the sparsity of the solution, *i.e.*, how many covariates or regressors are used.

With CV, some independent observations are left out of the inference and the hyperparameter values that yield the best prediction performance on this data are selected. A search for the best parameter can be done with a time consuming exhaustive grid-search, smooth optimization (see [6] and references therein), sequential or even random search [7, 8]. The CV approach however needs the i.i.d. assumption to be fulfilled, which is not always the case in practice, e.g. when working with signals or arrays of sensors as in the case of our application to brain imaging.

Following [1], we consider a hierarchical Bayesian model and propose to use a maximum-a-posteriori (MAP) estimation for the hyperparameters. In this paper, we are particularly interested in the high-dimensional regression setting using group-Lasso-like structured sparsity. This formulation is particularly adapted to the ill-posed inverse problem occurring in magnetoencephalography (MEG) and electroencephalography (EEG) source localization. M/EEG are non-invasive techniques that record the electromagnetic dynamical activity produced by the brain on a few hundreds of sensors. The objective is to identify the brain sources at the origin of the signals. In the literature a number of approaches have been proposed and MAP estimates that boil down to penalized regression with smooth or non-smooth penalties are the standard approaches employed by neuroscientists [9–15].

Here we study in particular the multi-task Lasso problem also known as multiple measurement vectors (MMV) in signal processing [16]. This estimator uses a group-Lasso-like penalty with mixed ℓ_1 and ℓ_2 norms. We first study the convex case addressing the limitations of the parametrization of [1]. We then extend the model to have a vector of hyperparameters to infer. It is compared to a non-convex $\ell_{2,0.5}$ penalization. The different strategies are tested on simulations and a source reconstruction problem using public M/EEG data.

Notation. The transpose of a matrix $\mathbf{A} \in \mathbb{R}^{M\times N}$ is denoted $\mathbf{A}^\mathbf{T}$. $\mathbf{A}[i,:]$ and $\mathbf{A}[:,j]$ correspond to the i^{th} row and the j^{th} column respectively. $\|\mathbf{A}\|_F$ indicates the Frobenius norm, and $\|\mathbf{A}\|_{p,q}$ the $\ell_{p,q}$ mixed norm with $\|\mathbf{A}\|_{p,q} = \left(\sum_i \left(\sum_j |\mathbf{A}[i,j]|^p\right)^{q/p}\right)^{1/q}$. $\mathbf{I}$ denotes the identity matrix.

2. MATERIALS AND METHODS

2.1. The MMV regression model

The MMV regression model can be written as:

$$\mathbf{M} = \mathbf{G}\mathbf{X} + \mathbf{E} \tag{1}$$

where $\mathbf{M} \in \mathbb{R}^{N\times T}$ is a matrix of T measurements vectors of dimension N. To give intuitions on notations, N can be the number of sensors and T a number of time instants. Matrix $\mathbf{G} \in \mathbb{R}^{N\times S}$ is the design matrix, a known instantaneous mixing matrix also referred to as the forward matrix where $N \ll S$. This matrix relates the source to the measurements. Matrix $\mathbf{E}$ is the measurement noise, which is assumed to be additive, white, and Gaussian, $E[:,j] \sim \mathcal{N}(0,\mathbf{I})\ \forall j$. $\mathbf{X} \in \mathbb{R}^{S\times T}$ corresponds to the parameters (the sources) to be estimated.

Assuming a known regularization parameter $\lambda > 0$, the MAP estimator is given for the above model by:

$$\hat{\mathbf{X}}_\lambda = \arg\min_{\mathbf{X}} \frac{1}{2}\|\mathbf{M} - \mathbf{G}\mathbf{X}\|_F^2 + \lambda\mathcal{P}(\mathbf{X}) \ , \tag{2}$$

where $\mathcal{P}(\mathbf{X})$ is a regularization term and λ the trade-off parameter between the data fit and the penalization. In practice, the value of λ depends on the problem at hand, the noise level, and on the choice of regularization $\mathcal{P}(\mathbf{X})$. Finding a way to estimate the hyperparameter with minimal user intervention is therefore particularly important.

Recently *Pereyra et al.* [1] proposed a strategy for hyperparameter estimation in the context of MAP inference when the prior or the regularizer is a k-homogeneous function. The regularizer $\mathcal{P}$ in (2) is a k-homogeneous function if there exists $k \in \mathbb{R}^+$ such that $\mathcal{P}(\eta\mathbf{X}) = \eta^k\mathcal{P}(\mathbf{X}), \forall\mathbf{X} \in \mathbb{R}^{S\times T}$ and $\forall\eta > 0$. The k-homogeneous condition is satisfied for all $\ell_{p,q}$ mixed norms. In this paper, we focus on the estimation of the hyperparameters for hierarchical Bayesian models yielding convex $\ell_{2,1}$ ($\mathcal{P}(\mathbf{X}) = \|\mathbf{X}\|_{2,1}$) or non-convex $\ell_{2,0.5}$ penalties, which are respectively 1-homogeneous and 0.5-homogeneous. The non-convex penalization is solved using iterative reweighted convex optimization schemes, *i.e.*, each iteration is a weighted $\ell_{2,1}$-norm.

In [1], the fixed point strategy proposed is validated on an image denoising problem using an analysis prior, *i.e.* where the solution is not sparse but has a sparse representation in some transformed domain. We now illustrate and explain why the method from [1] cannot be used out-of-the-box when using a synthesis prior for an under-determined problem.

2.2. Hierarchical Bayesian modeling and reformulation

Bayesian modeling imposes hyperpriors, which are priors on the distributions of the hyperparameters. A popular choice of hyperprior is the gamma distribution:

$$p(\lambda) = \frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}\exp(-\beta\lambda)\mathbf{1}_{\mathbb{R}^+}(\lambda),\ \lambda \in \mathbb{R} \tag{3}$$

where $\mathbf{1}$ denotes the indicator function, Γ is the gamma function, and α and β are the shape parameters.

Following [1] that uses a joint MAP estimator of λ and $\mathbf{X}$, one obtains that $\hat{\lambda}$ should satisfy:

$$\hat{\lambda} = \frac{ST/k + \alpha - 1}{\mathcal{P}(\hat{\mathbf{X}}_{\hat{\lambda}}) + \beta} \ , \tag{4}$$

where $\hat{\mathbf{X}}_{\hat{\lambda}}$ is the solution of (2) for $\lambda = \hat{\lambda}$.

Looking at (4), one can observe that if ST is big, which happens for high dimensional problems, the numerator can significantly dominate the denominator, especially if the estimate $\hat{X}$ is very sparse. In practice using (4) in this scenario results rapidly in huge values of λ and empty supports. This issue is much less critical when using an analysis prior for denoising as in [1], as the size of the unknown coefficients is in this case NT, where $NT \ll ST$.

To overcome this problem, we rewrite the objective function in such a way that we obtain the same solution $\mathbf{X}$ but with a $\frac{\lambda}{ST}$. This can be written as:

$$\hat{\mathbf{X}} = \arg\min_{\mathbf{X}} \frac{ST}{2}\|\mathbf{M} - \mathbf{G}\mathbf{X}\|_F^2 + \lambda\mathcal{P}(\mathbf{X}) \tag{5}$$

Note that this is just a reparametrization of (2). In practice, this boils down to multiplying $\mathbf{M}$ and $\mathbf{G}$ by $\sqrt{ST}$. However this only solves one difficulty in the parameter's update. Another disadvantage is that none of the parameters in (4) take into account the scale of $\mathbf{G}$. In the next section, we explain how to properly calibrate the hyperprior parameters α and β given M, G and $\mathcal{P}$.

2.3. Setting hyperpriors with a single hyperparameter

As in [1], gamma hyperpriors are used to derive two iterative algorithms that simultaneously estimate a single hyperparameter λ and the entries of $\mathbf{X}$, yet the values of α and β are still to be defined. In [1], it is suggested to set α and β to 1, which turns out to be inappropriate for underdetermined inverse (deconvolution) problems as our brain imaging problem of interest.

A first observation is that α and β should default to reasonable values and be insensitive to trivial changes in matrix $\mathbf{G}$ such as scaling, i.e., multiplying $\mathbf{G}$ by a scalar. This is the problem we investigate now.

In (4), the numerator would not be affected by a rescaling of $\mathbf{G}$. However, the denominator that contains $\mathcal{P}(\mathbf{X}_{\lambda^\star})$

would. To make the estimation robust to changes of $\mathbf{G}$ such as scaling, one therefore needs to modify the numerator, hence make α a function of $\mathbf{G}$. Setting α to 1 independently of the problem, as in [1], is inadequate.

To set the value of α, we propose to take advantage of the fact that if $\mathcal{P}(\mathbf{X}) = \|\mathbf{X}\|_{2,1}$ one can analytically compute λ_{max}, which is defined as the smallest regularization parameter for which the solution is zero [17]. It is given by:

$$\lambda_{max} = \|\mathbf{G}^T\mathbf{M}\|_{2,\infty} = \max_i \|(\mathbf{G}^T\mathbf{M})[i,:]\|_2. \quad (6)$$

Parameter λ can therefore be parametrized as a fraction, or a percentage, of λ_{max}. This allows us to have a good a priori guess on the peak of the gamma distribution. We set the peak, a.k.a. the mode, to $mode = \tau \times \lambda_{max}$, with $\tau \in [0, 1]$.
Once the mode is known, it is straightforward to fix the value of α: $mode = \frac{\alpha-1}{\beta}$ for $\alpha \geq 1$. From now on we fix α as:

$$\alpha = mode \times \beta + 1 = \tau \times \lambda_{max} \times \beta + 1 \ . \quad (7)$$

Concerning the parameter β, for our specific problem of interest we fix it so that 99% of the probability density of the gamma distribution is between 20% and 70% of λ_{max}. This is motivated by the fact that in our case solutions are expected to be extremely sparse, with only a handful of active brain regions. This is of course application specific.

2.4. Estimation of a vector of hyperparameters

The penalization of the form $\mathcal{P}(\mathbf{X}) = \|\mathbf{X}\|_{2,\cdot}$ are separable in S groups of coefficients. As only a few groups are expected to be active, a natural idea is to penalize less the important groups. To do this, we propose to estimate one parameter per group of coefficients or row of $\mathbf{X}$ using the convex $\ell_{2,1}$ penalization. Rewriting (2) in the MAP framework leads to:

$$\mathbf{X}^\star = \arg\max_{\mathbf{X}} p(\mathbf{X}, \mathbf{M}|\lambda) = \arg\max_{\mathbf{X}} p(\mathbf{M}|\mathbf{X})p(\mathbf{X}|\lambda) \quad (8)$$

where $p(\mathbf{M}|\mathbf{X})$ is the likelihood function corresponding to the first term in (2) and $p(\mathbf{X}|\lambda)$ is the regularization corresponding to the second term in (2). This Bayesian formulation requires to compute the normalization factor $C(\lambda)$ in $p(\mathbf{X}|\lambda) = \exp(-\lambda\mathcal{P}(\mathbf{X}))/C(\lambda)$. Computing this constant $C(\lambda)$ in general is intractable as it involves an integration. Yet [1] showed that it admits an exact closed-form when the penalization is k-homogeneous as $C(\lambda) = D\lambda^{-ST/k}$ where $D = C(1)$ is a constant independent of λ [1].
We now propose a joint-MAP estimation with $\lambda \in \mathbb{R}^S$. We look for $(\mathbf{X}^\star, \lambda^\star) \in \mathbb{R}^{(S\times T)} \times \mathbb{R}^S$ which maximizes $p(\mathbf{X}, \lambda|\mathbf{M})$. A sufficient condition of optimality is given by:

$$(0_{(S\times T)}, 0_S) \in -\partial_{\mathbf{X},\lambda} \log p(\mathbf{X}^\star, \lambda^\star|\mathbf{M}) \quad (9)$$

$$\textit{i.e.} \quad \begin{aligned} 0_{S\times T} &\in -\partial_{\mathbf{X}} \log p(\mathbf{X}^\star, \lambda^\star|\mathbf{M}), \\ 0 &\in -\partial_{\lambda_i} \log p(\mathbf{X}^\star, \lambda^\star|\mathbf{M}) \quad \forall i, \end{aligned} \quad (10)$$

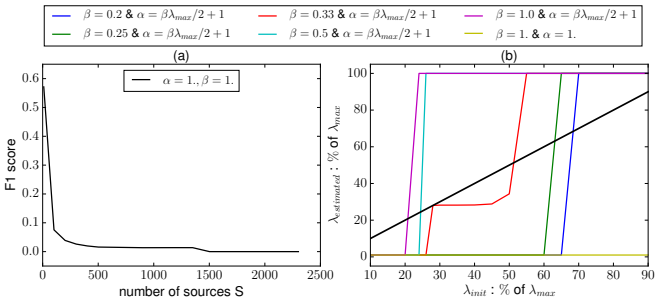

Fig. 1. (a) Source identification results for different number of sources measured with F1 score using $\alpha = 1$ and $\beta = 1$. The higher the number of regressors the worse is the performance. (b) Estimated λ as a function of λ_{init} for different values of a and b. The red curve for $\beta = 0.33$ gives the best plateau, which demonstrates that (a, b) shall be carefully adjusted.

where $\partial_{\mathbf{X},\lambda}$ is the set of subgradients (the subdifferential).

The optimization over $\mathbf{X}$ at iteration t satisfies (5):

$$\mathbf{X}^{(t)} = \arg\min_{\mathbf{X}\in\mathbb{R}^{S\times T}} \frac{ST}{2}\|\mathbf{M} - \mathbf{G}\mathbf{X}\|_F^2 + \sum_i \lambda_i^{(t-1)}\|X[i,:]\|$$

The next step is to optimize over $\lambda_i, \forall i$. Eq. (10) leads to:

$$0 \in -\partial_{\lambda_i} \log p(\mathbf{X}^{(t)}, \mathbf{M}|\lambda) - \partial_{\lambda_i} \log p(\lambda) \quad (11)$$

Using $p(\mathbf{X}^{(t)}, \mathbf{M}|\lambda) = p(\mathbf{M}|\mathbf{X}^{(t)})p(\mathbf{X}^{(t)}|\lambda)$, one has that $-\partial_{\lambda_i} \log p(\mathbf{X}^{(t)}, \mathbf{M}|\lambda) = -\partial_{\lambda_i} \log p(\mathbf{X}^{(t)}|\lambda)$. We then use the normalization factor $C(\lambda)$ which gives:
$-\partial_{\lambda_i} \log p(\mathbf{X}^{(t)}, \mathbf{M}|\lambda) = \|\mathbf{X}[i,:]\| + \partial_{\lambda_i} \log C(\lambda)$ and $\partial_{\lambda_i} \log C(\lambda) = \frac{-ST}{k\lambda_i}$. Regarding the second term in (11), (3) yields $-\partial_{\lambda_i} \log p(\lambda) = -\frac{\alpha-1}{\lambda_i} + \beta$. Completing the derivations, the equation for each λ_i, $i \in [1 \dots S]$, reads:

$$\lambda_i^\star = \frac{ST/k + \alpha - 1}{\|X_{\lambda^\star}[i,:]\| + \beta} \ . \quad (12)$$

3. APPLICATION TO M/EEG INVERSE PROBLEM

3.1. Simulation

We generated a simulation dataset with $N = 302$ sensors, $T = 190$ time samples and $S = 1500$ sources. Four sources were randomly selected to be active with realistic waveforms. The linear forward operator $\mathbf{G}$ was a random matrix, whose columns were normalized to 1. Two levels of white noise were added to the simulation. We always used $\tau = 0.5$.

In order to illustrate the issue when using a synthesis prior for large problems, we run the estimation of the hyperparameter λ as suggested in [1] using the 0.5-homogeneous non-convex prior. Fig. 1-(a) shows the F1 score of the source reconstruction (1 for good reconstruction and 0 for bad). The source estimation is failing for almost all the range of data

size. Fig. 1-(b) shows the results after reformulating the problem with different settings of α and β. One can notice that a setting as in [1] with $\alpha = 1$ and $\beta = 1$ always gives an estimated λ around 1% of λ_{max} which is not promoting the sparsity we are looking for in this kind of setting. For this aim, we varied the values of β and computed α as defined before. Fig. 1-(b) shows that for most values of β we have rather a too low estimation of $\lambda \approx 1\%$ or a too high $\lambda \geq 100\%$ resulting in zero source found active. Interestingly setting $\beta = 1/3$ gives a plateau at $\hat{\lambda}$ close to $0.3\lambda_{max}$. This is evidence of a clear fixed point for the iterative process $\lambda^{(t+1)} = f(\lambda^{(t)})$, where f is the update rule of λ in (4). We use $\beta = 1/3$ from now on and its corresponding α.

Fig. 2 represents the simulated sources with stars and the estimated ones with plain lines. Fig. 2-(a)-(b) display results with the $\ell_{2,1}$ and $\ell_{2,0.5}$ norms respectively, using one hyperparameter initialized to $\lambda = 0.5\lambda_{max}$. One can see that in Fig.2-(a), the $\ell_{2,1}$ norm recovers the four sources with an amplitude bias (the estimated amplitude is lower than the exact one), and that several sources shown in light green are almost flat around zero but still found as active sources. There is no way to reduce the support without losing one of the four simulated sources, *i.e.* the $\ell_{2,1}$ norm with one hyperparameter fails to recover the exact simulated sources. The $\ell_{2,0.5}$ norm in (b) estimates the exact four source amplitudes without amplitude bias thanks to the non-convexity [18]. On the other hand, Fig. 2-(c) shows the results for the convex penalty using one hyperparameter per source. It can be seen that it is qualitatively equivalent to the non-convex penalty. The advantage of having one hyperparameter per source is to pick up only the sources involved in the measurement $\mathbf{M}$ and drop the extra almost-zero sources visible in Fig. 2-(a) (light green). This extension produces sparser results and less amplitude bias without casting the problem as non-convex.

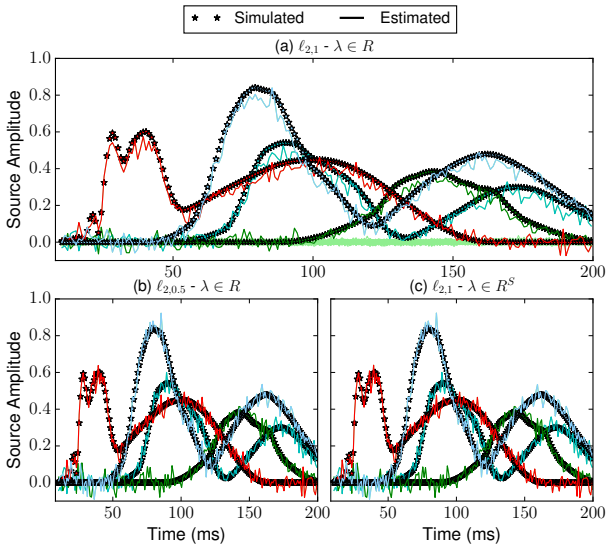

Fig. 2. Source reconstruction on simulated data. (a): Source estimates obtained using $\ell_{2,1}$ with one λ. The solution is not sparse enough (the zero-sources in light green) and there is an amplitude bias between the exact amplitudes (stars) and the estimated ones (raw lines). (b): Good reconstruction of the four sources using $\ell_{2,0.5}$ and one λ, which is equivalent to the reconstruction using the $\ell_{2,1}$ norm with $\lambda \in \mathbb{R}^S$ (c). Each of the four sources is encoded with a color.

3.2. Experimental results with MEG auditory data

We applied the estimation of a single hyperparameter and a hyperparameter per source using the convex $\ell_{2,1}$ penalty on a real open dataset (MNE sample dataset [19]). It corresponds to a dataset with $N = 305$ sensors, $T = 55$ time samples and $S = 7498$ sources. Fig. 3 shows the source amplitudes of the two auditory sources and their positions in the brain when estimating a hyperparameter per source. When using a single hyperparameter on the convex norm $\ell_{2,1}$, multiple spurious sources are found as active which replicates the simulation on Fig.2-(a). These source estimates in Fig. 3 correspond to the M100 peak (peak around 100 ms) generated in the vicinity of the bilateral auditory cortices in superior temporal gyri (the relevant auditory area).

4. DISCUSSION AND CONCLUSION

In this paper, we have explained how to address the limitations of the fixed point iteration algorithm presented in [1] when solving high-dimensional sparse synthesis problems. This required to reformulate the problem and to propose a strategy to adjust the scale parameters α and β of the Gamma prior when considering MMV problems with group-Lasso-like penalties. Finally, we extended the approach to estimate a vector of hyperparameters. The approach was applied to the M/EEG inverse problem and then compared with the estimation of a single hyperparameter using a non-convex penalty. The results on simulated data show that using a vector of hyperparameters with the convex norm is qualitatively equivalent to the non-convex norm. This can be explained by the fact that the optimization problem for the non-convex case is solved using majorization-minimization techniques, which lead to a convex problem with some reweighting. This turns out to be similar to the multi-hyperparameter approach yet using different update rules.

Concerning real data, we showed how the algorithm with a vector of hyperparameters allows us to reconstruct the two relevant sources in a MEG auditory dataset. Further investigations will focus on the extension of this hyperparameter estimation approach to the sparse group-Lasso $\ell_{2,1} + \ell_1$, which contains two different hyperparameters aiming to relax the temporal stationarity assumptions of simple group-Lasso-like penalties [18, 20, 21].

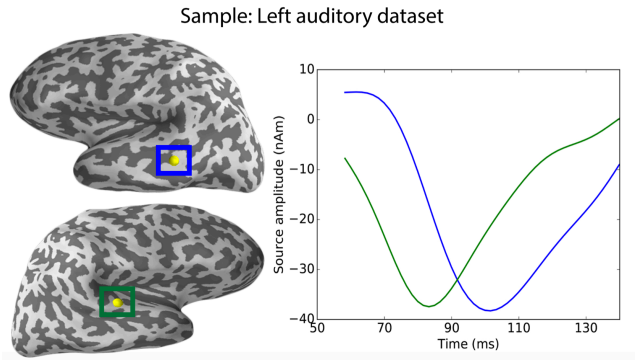

Fig. 3. Source reconstruction on MEG auditory data (sample dataset [19]). Source amplitude of two sources (blue and green) in the right and their corresponding positions in the brain on the left.

5. ACKNOWLEDGMENT

This work was supported by the French National Research Agency (ANR-14-NEUC-0002-01), the National Institutes of Health (R01 MH106174) and the European Research Council (ERC-YStG-676943).

6. REFERENCES

[1] M. Pereyra, J. M. Bioucas-Dias, and M. A. Figueiredo, "Maximum-a-posteriori estimation with unknown regularisation parameters," in *Proc. EUSIPCO*. IEEE, 2015, pp. 230–234.

[2] G. Schwarz *et al.*, "Estimating the dimension of a model," *Ann. Stat.*, vol. 6, no. 2, pp. 461–464, 1978.

[3] C. M. Stein, "Estimation of the mean of a multivariate normal distribution," *The annals of Statistics*, pp. 1135–1151, 1981.

[4] F. Luisier, T. Blu, and M. Unser, "A new SURE approach to image denoising: Interscale orthonormal wavelet thresholding," *IEEE Trans. on image processing*, vol. 16, no. 3, pp. 593–606, 2007.

[5] C. Guo and M. E. Davies, "Near optimal compressed sensing without priors: Parametric SURE approximate message passing," *IEEE Trans. on Signal Processing*, vol. 63, no. 8, pp. 2130–2141, 2015.

[6] F. Pedregosa, "Hyperparameter optimization with approximate gradient," in *Proc. ICML*, vol. 48, 2016.

[7] J. S. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, "Algorithms for hyper-parameter optimization," in *Proc. NIPS*, 2011, pp. 2546–2554.

[8] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *Journal of Machine Learning Research*, vol. 13, pp. 281–305, Feb 2012.

[9] S. Haufe, V. V. Nikulin, A. Ziehe, K.-R. Müller, and G. Nolte, "Combining sparsity and rotational invariance in EEG/MEG source reconstruction," *NeuroImage*, vol. 42, no. 2, pp. 726–738, 2008.

[10] W. Ou, M. S. Hämäläinen, and P. Golland, "A distributed spatio-temporal EEG/MEG inverse solver," *NeuroImage*, vol. 44, no. 3, pp. 932–946, 2009.

[11] A. Bolstad, B. Van Veen, and R. Nowak, "Space–time event sparse penalization for magneto-/electroencephalography," *NeuroImage*, vol. 46, no. 4, pp. 1066–1081, 2009.

[12] D. Wipf and S. Nagarajan, "A unified Bayesian framework for MEG/EEG source imaging," *NeuroImage*, vol. 44, no. 3, pp. 947–966, 2009.

[13] A. Gramfort, M. Kowalski, and M. Hämäläinen, "Mixed-norm estimates for the M/EEG inverse problem using accelerated gradient methods," *Physics in medicine and biology*, vol. 57, no. 7, p. 1937, 2012.

[14] F. Lucka, S. Pursiainen, M. Burger, and C. H. Wolters, "Hierarchical Bayesian inference for the EEG inverse problem using realistic FE head models: depth localization and source separation for focal primary currents," *Neuroimage*, vol. 61, no. 4, pp. 1364–1382, 2012.

[15] P. A. Valdés-Sosa, M. Vega-Hernández, J. M. Sánchez-Bornot, E. Martínez-Montes, and M. A. Bobes, "EEG source imaging with spatio-temporal tomographic nonnegative independent component analysis," *Human brain mapping*, vol. 30, no. 6, pp. 1898–1910, 2009.

[16] S. F. Cotter, B. D. Rao, K. Engan, and K. Kreutz-Delgado, "Sparse solutions to linear inverse problems with multiple measurement vectors," *IEEE Trans. on Signal Processing*, vol. 53, no. 7, pp. 2477–2488, 2005.

[17] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski, "Optimization with sparsity-inducing penalties," *Foundations and Trends® in Machine Learning*, vol. 4, no. 1, pp. 1–106, 2012.

[18] D. Strohmeier, Y. Bekhti, J. Haueisen, and A. Gramfort, "The iterative reweighted Mixed-Norm Estimate for spatio-temporal MEG/EEG source reconstruction," *IEEE Trans. on Medical Imaging*, 2016, to appear.

[19] A. Gramfort, M. Luessi, E. Larson, D. A. Engemann, D. Strohmeier, C. Brodbeck, L. Parkkonen, and M. S. Hämäläinen, "MNE software for processing MEG and EEG data," *Neuroimage*, vol. 86, pp. 446–460, 2014.

[20] A. Gramfort, D. Strohmeier, J. Haueisen, M. Hamalainen, and M. Kowalski, "Time-frequency mixed-norm estimates: Sparse M/EEG imaging with non-stationary source activations," *NeuroImage*, vol. 70, pp. 410 – 422, 2013.

[21] Y. Bekhti, D. Strohmeier, M. Jas, R. Badeau, and A. Gramfort, "M/EEG source localization with multiscale time-frequency dictionaries," in *6th International Workshop on Pattern Recognition in Neuroimaging (PRNI)*, 2016.