VECTOR ℓ_0 LATENT-SPACE PRINCIPAL COMPONENT ANALYSIS

Martin Luessi[⋆]*, Matti S. Hämäläinen*[⋆]*, and Victor Solo*[⋆†]

[⋆]Dept. Radiology, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA
[†]Dept. Electrical Eng. and Telecommunications, University of New South Wales, Sydney, Australia

ABSTRACT

Principal component analysis (PCA) is a widely used signal processing technique. Instead of performing PCA in the data space, we consider the problem of sparse PCA in a potentially higher dimensional latent space. To do so, we zero-out groups of variables using vector ℓ_0 regularization. The estimation is based on the maximization of the penalized log-likelihood, for which we develop an efficient coupled expectation-maximization (EM) - minorization-maximization (MM) algorithm. For the special case when the latent- and observation space are identical, our method corresponds to an existing vector ℓ_0 PCA method, which we verify using simulations. The proposed method can also be utilized for penalized linear regression and we use simulations to demonstrate superior estimation performance. As an example of a practical application, we use our method to localize cortical activity from magnetoencephalography (MEG) data.

Index Terms— principal component analysis, PCA, minorization-maximization, penalized likelihood, sparsity, l0, MEG, EEG, source localization

1. INTRODUCTION

In the noisy PCA (nPCA) problem [1, 2, 3], the signal model is given by

$$\mathbf{y}_t = \mathbf{F}\mathbf{u}_t + \boldsymbol{\epsilon}_t, \quad t = 1, \ldots, T, \tag{1}$$

where $\mathbf{y}_t \in \mathbb{R}^M$ are the observations at time t, $\mathbf{u}_t \sim \mathcal{N}(0, \mathbf{I}_k)$ is a white noise basis, and $\boldsymbol{\epsilon}_t \sim \mathcal{N}\left(0, \sigma^2 \mathbf{I}_M\right)$ is additive noise. The goal of nPCA is to estimate the $M \times k$ loading matrix $\mathbf{F}$ and the noise variance σ^2. In [2, 3] an Expectation Maximization (EM) algorithm [4] was developed to obtain maximum likelihood (ML) estimates of $\mathbf{F}$ and σ^2. However, when the number of time points T is small and M is large, the ML approach exhibits poor estimation performance. In [5] a sparse variable PCA (svnPCA) approach with a vector ℓ_0 penalty to zero-out rows in $\mathbf{F}$ was introduced. The vector ℓ_0 penalty can greatly improve the quality of the estimation in the "large M small T" setting when only a subset of variables contain the signal of interest. In this work, we consider an extension of nPCA. Namely, we assume the following signal model

$$\mathbf{y}_t = \mathbf{G}\mathbf{F}\mathbf{u}_t + \boldsymbol{\epsilon}_t, \quad t = 1, \ldots, T, \tag{2}$$

where $\mathbf{G} \in \mathbb{R}^{M \times N}$ is a known matrix and the loading matrix $\mathbf{F}$ now has dimension $N \times k$. Note that a related problem is encountered in regression where one is interested in estimating $\mathbf{X} = [\mathbf{x}_1 \cdots \mathbf{x}_T]$ from

$$\mathbf{y}_t = \mathbf{G}\mathbf{x}_t + \boldsymbol{\epsilon}_t, \quad t = 1, \ldots, T. \tag{3}$$

For the regression problem, it has been established that $\mathbf{X}$ can be estimated with high accuracy even in the underdetermined case when $M \ll N$ using ℓ_p-norm regularized regressions with $p \leq 1$ as long as each $\mathbf{x}_t$ is sparse, i.e., has a small number of non-zero entries, and $\mathbf{G}$ has certain properties [6, 7]. For the case where each $\mathbf{x}_t$ has non-zero elements at the same locations, the recovery can be improved by jointly estimating $\mathbf{X}$, such as in the Group Lasso algorithm [8] which uses an $\ell_1\ell_2$-norm penalty (ℓ_1-norm of the ℓ_2 norm of all rows in $\mathbf{X}$). While the problem is convex for $p = 1$, superior estimation quality can be achieved using vector ℓ_0 penalized regression [9]. Also Bayesian methods [10] provide high quality solutions and it can be shown that they solve an approximation to the ℓ_0-norm regularized problem [11].

In this work, we use a penalty function that is closely related to the vector ℓ_0 penalty in svnPCA0 [5]. The penalty is given by

$$\rho(\mathbf{F}) = \sum_{v \in \mathcal{V}} I(\mathbf{F}_v \neq \mathbf{0}), \tag{4}$$

where $I(\cdot)$ is the indicator function, i.e., $I(\mathbf{F}_v \neq \mathbf{0}) = 0$ only if $\mathbf{F}_v = \mathbf{0}$ and $I(\mathbf{F}_v \neq \mathbf{0}) = 1$ otherwise, $\mathcal{V}$ is a set of non-overlapping groups of row indices, and $\mathbf{F}_v$ is an $|v| \times k$ sub-matrix of $\mathbf{F}$ obtained for the indices in v. This formulation allows us to encode prior knowledge about the relationship between variables as the penalty forces *entire groups* to become exactly zero, not individual loadings as it would be the case with a scalar ℓ_0 penalty, this is why we call it a *vector ℓ_0 penalty*. The estimation procedure is based on the maximization of the penalized log-likelihood, for which we develop an efficient coupled EM-MM algorithm [12, 13]. The algorithm has several interesting properties. First, for the case where $\mathbf{G} = \mathbf{I}$ and $M = N$, it is identical to the penalized EM algorithm from [5] and in simulations we show that both algorithms indeed obtain the same solutions. Second, even when $\mathbf{G} \neq \mathbf{I}$, the maximization step in our MM algorithm has a closed form solution, which allows us to efficiently update $\mathbf{F}$ during each iteration.

Of interest is the underdetermined case when the latent space has a higher dimension than the observation space, i.e., $M \ll N$. Unlike for the regression problem, no theoretical results on recovery guarantees currently exist for latent space PCA. However, we show that our algorithm can also been utilized to solve the regression problem. Specifically, we can estimate $\mathbf{F}$ and then obtain an estimate $\hat{\mathbf{x}}_t = \mathbf{F}\hat{\mathbf{u}}_t$, where $\hat{\mathbf{u}}_t$ is the conditional expectation of $\mathbf{u}_t$ given $\mathbf{Y} = [\mathbf{y}_1 \cdots \mathbf{y}_T]$, $\mathbf{F}$ and σ^2. We use simulations to show that our method provides a better reconstruction performance than existing penalized regression methods, especially for the case where the active rows in $\mathbf{X}$ are linearly related, which is not taken into account by existing methods.

This work was supported by NIH grants P41RR14075 (NCRR), 5R01EB009048 (NIBIB), and NSF Grant 1042134, the Cognitive Rhythms Collaborative: A Discovery Network. M.L. was partially supported by the Swiss National Science Foundation Early Postdoc.Mobility fellowship 14848.

Source localization in magneto- and electroencephalography (M/EEG) is a practical application where grouping variables, as in (4), and the assumption of linear relations between rows of $\mathbf{X}$ is sensible. For this application, $\mathbf{G}$ is a gain matrix describing the relation between N fixed-orientation current dipoles distributed over the cortical mantle and the measurements at M sensors [14]. For M/EEG, row-sparsity in $\mathbf{X}$ means that only a small number of cortical sites are active, which is a valid assumption for many experimental paradigms and $\ell_1\ell_2$-norm ("mixed norm") penalized source localization methods [15, 16] can provide accurate localization of the active sources. The proposed method also lends itself to M/EEG source localization and we demonstrate its application to MEG data. A potential benefit over, e.g, $\ell_1\ell_2$-norm regularization is that the proposed method can model the activity at one cortical location using a small number of basis signals. This is the case since a single location contributes multiple rows to $\mathbf{F}$ corresponding to current dipoles at the same location but with different spatial orientations.

The method presented in this work builds on our previous work [17], where we developed a similar method with a vector ℓ_1 penalty for $\mathbf{F}$. However, in [17] we motivated the problem from a regression perspective and did not consider the application of the method to latent space PCA. For regression, the use of the vector ℓ_0 penalty in the current work improves the estimation quality when compared to our previous method, which we demonstrate using simulations.

This paper is organized as follows: In Section 2, the proposed method is introduced. In Section 3 we evaluate the performance of the proposed method using simulations. The method is applied to MEG data in Section 4 and finally, the paper is summarized and conclusions are drawn in Section 5.

2. PROPOSED METHOD

We estimate $\mathbf{F}$ and σ^2 by maximizing the penalized log-likelihood which we obtain from (2) by marginalizing over $\mathbf{u}_t$, computing the logarithm, and combining it with the regularization term (4), resulting in

$$\mathcal{L}\left(\mathbf{F}\right) = -\frac{1}{2}\mathrm{tr}\left(\mathbf{S}_\mathbf{y}\mathbf{\Omega}^{-1}\right) - \frac{1}{2}|\mathbf{\Omega}| - \frac{h}{2\sigma^2}\rho\left(\mathbf{F}\right), \tag{5}$$

where $\mathbf{\Omega} = \sigma^2\mathbf{I}_M + \mathbf{G}\mathbf{F}\mathbf{F}^T\mathbf{G}^T$, $\mathbf{S}_\mathbf{y} = {}^1\!/_T \sum_{t=1}^T \mathbf{y}_t\mathbf{y}^T$, and h is the regularization parameter. Due to the form of (5), a direct maximization with respect to $\mathbf{F}$ is difficult and we follow the same approach as in [17] and develop a coupled EM-MM algorithm to perform the maximization. First, we compute the penalized complete data log-likelihood using

$$\mathcal{L}_{\mathbf{Y},\mathbf{U}}\left(\mathbf{F}\right) = \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T}\ln \mathrm{p}\left(\mathbf{y}_t, \mathbf{u}_t|\mathbf{F}, \sigma^2\right)\right] - \frac{h}{2\sigma^2}\rho\left(\mathbf{F}\right), \tag{6}$$

where the expectation is computed with respect to

$$\mathrm{p}\left(\mathbf{u}_t|\mathbf{y}_t, \mathbf{F}, \sigma^2\right) = \mathcal{N}\left(\mathbf{W}^{-1}\mathbf{F}^T\mathbf{G}^T\mathbf{y}_t, \sigma^2\mathbf{W}^{-1}\right), \tag{7}$$

where $\mathbf{W} = \sigma^2\mathbf{I}_k + \mathbf{F}^T\mathbf{G}^T\mathbf{G}\mathbf{F}$. Which gives

$$\begin{aligned}\mathcal{L}_{\mathbf{Y},\mathbf{U}}\left(\mathbf{F}\right) = &-\frac{\mathrm{tr}\left(\mathbf{S}_\mathbf{y}\right)}{2\sigma^2} + \frac{\mathrm{tr}\left(\mathbf{F}\mathbf{\Gamma}^T\right)}{\sigma^2} \\ &- \frac{\mathrm{tr}\left(\mathbf{F}\mathbf{A}\mathbf{F}^T\mathbf{G}^T\mathbf{G}\right)}{2\sigma^2} - \frac{M}{2}\ln\sigma^2 - \frac{h}{2\sigma^2}\rho\left(\mathbf{F}\right),\end{aligned} \tag{8}$$

where

$$\mathbf{\Gamma} = \mathbf{G}^T\mathbf{S}_\mathbf{y}\mathbf{G}\mathbf{F}\mathbf{W}^{-1}, \tag{9}$$

$$\mathbf{A} = \mathbf{W}^{-1}\left(\sigma^2\mathbf{W} + \mathbf{F}^T\mathbf{G}^T\mathbf{S}_\mathbf{y}\mathbf{G}\mathbf{F}\right)\mathbf{W}^{-1}. \tag{10}$$

Notice that $\mathcal{L}_{\mathbf{Y},\mathbf{U}}\left(\mathbf{F}\right)$ minorizes $\mathcal{L}\left(\mathbf{F}\right)$; this minorization provides the basis of the well-known Expectation Maximization (EM) algorithm [4], which maximizes (5) by iteratively updating $\mathbf{W}$, $\mathbf{A}$, and $\mathbf{\Gamma}$ in the E-step and maximizing (8) with respect to $\mathbf{F}$ and σ^2 in the M-step. Surprisingly, for the svnPCA0 algorithm [5], the M-step has a closed form solution. However, due to the presence of $\mathbf{G}$ this is not the case here and the maximization of (8) with respect to $\mathbf{F}$ remains challenging. Therefore, we introduce a functional which minorizes the penalized EM functional (8) (and hence also minorizes (5)) in order to obtain a tractable optimization procedure.

First, we introduce the deviation $\tilde{\mathbf{F}} = \mathbf{F} - \mathbf{F}_0$ and rewrite (8) as follows

$$\begin{aligned}\mathcal{L}_{\mathbf{Y},\mathbf{U}}\left(\mathbf{F}\right) = &-\frac{1}{2\sigma^2}\mathrm{tr}\left(\mathbf{S}_\mathbf{y}\right) + \frac{1}{\sigma^2}\mathrm{tr}\left(\tilde{\mathbf{F}}\mathbf{\Gamma}^T\right) \\ &- \frac{1}{2\sigma^2}\mathrm{tr}\left(\tilde{\mathbf{F}}\mathbf{A}\tilde{\mathbf{F}}^T\mathbf{G}^T\mathbf{G}\right) - \frac{1}{2\sigma^2}\mathrm{tr}\left(\tilde{\mathbf{F}}\mathbf{A}\mathbf{F}_0^T\mathbf{G}^T\mathbf{G}\right) \\ &- \frac{M}{2}\ln\sigma^2 - \frac{h}{2\sigma^2}\rho\left(\mathbf{F}\right) + c,\end{aligned} \tag{11}$$

where all terms solely depending on $\mathbf{F}_0$ have been absorbed into the additive constant. We now use the maximum eigenvalue $\lambda = \lambda_{\max}\left(\mathbf{G}\mathbf{G}^T\right)$ to minorize $\mathcal{L}_{\mathbf{Y},\mathbf{U}}\left(\mathbf{F}\right)$ as follows

$$\begin{aligned}m\left(\mathbf{F}, \mathbf{F}_0\right) = &-\frac{1}{2\sigma^2}\mathrm{tr}\left(\mathbf{S}_\mathbf{y}\right) + \frac{1}{\sigma^2}\mathrm{tr}\left(\tilde{\mathbf{F}}\mathbf{\Gamma}^T\right) - \frac{\lambda}{2\sigma^2}\mathrm{tr}\left(\tilde{\mathbf{F}}\mathbf{A}\tilde{\mathbf{F}}^T\right) \\ &- \frac{1}{2\sigma^2}\mathrm{tr}\left(\tilde{\mathbf{F}}\mathbf{A}\mathbf{F}_0^T\mathbf{G}^T\mathbf{G}\right) - \frac{M}{2}\ln\sigma^2 - \frac{h}{2\sigma^2}\rho\left(\mathbf{F}\right) + c,\end{aligned} \tag{12}$$

for which we have $m\left(\mathbf{F}, \mathbf{F}_0\right) \leq \mathcal{L}_{\mathbf{Y},\mathbf{U}}\left(\mathbf{F}\right)$ where equality only holds if the deviation is zero, i.e., $\mathbf{F} = \mathbf{F}_0$ and therefore $m\left(\mathbf{F}, \mathbf{F}_0\right)$ is a minorizing functional. We now expand $m\left(\mathbf{F}, \mathbf{F}_0\right)$ and absorb all terms not depending on $\mathbf{F}$ or σ^2 into the additive constant resulting in the following MM objective function

$$\begin{aligned}\mathcal{J}\left(\mathbf{F}\right) = &-\frac{1}{\sigma^2}\mathrm{tr}\left(\mathbf{F}\mathbf{K}^T\right) + \frac{\lambda}{2\sigma^2}\mathrm{tr}\left(\mathbf{F}\mathbf{A}\mathbf{F}^T\right) \\ &+ \frac{1}{2\sigma^2}\mathrm{tr}\left(\mathbf{S}_\mathbf{y}\right) + \frac{M}{2}\ln\sigma^2 + \frac{h}{2\sigma^2}\rho\left(\mathbf{F}\right) + c\end{aligned} \tag{13}$$

where the $k \times N$ matrix $\mathbf{K}^T$ is given by

$$\mathbf{K}^T = \mathbf{\Gamma}^T + \mathbf{A}\mathbf{F}^T\left(\lambda\mathbf{I} - \mathbf{G}^T\mathbf{G}\right). \tag{14}$$

Notice that (13) we multiplied all terms with -1 and therefore the $\mathbf{F}$ and σ^2 parameters in the maximize step are obtained by minimizing (13). To do so, it is important to realize that (13) is separable in $\mathbf{F}$, i.e., we can rewrite it as follows

$$\mathcal{J}\left(\mathbf{F}\right) = \sum_{v\in\mathcal{V}} j_v\left(\mathbf{F}_v\right) + \frac{1}{2\sigma^2}\mathrm{tr}\left(\mathbf{S}_\mathbf{y}\right) + \frac{M}{2}\ln\sigma^2 + c, \tag{15}$$

where

$$\begin{aligned}j_v\left(\mathbf{F}_v\right) = &-\frac{1}{\sigma^2}\mathrm{tr}\left(\mathbf{F}_v\mathbf{K}_v^T\right) + \frac{\lambda}{2\sigma^2}\mathrm{tr}\left(\mathbf{F}_v\mathbf{A}\mathbf{F}_v^T\right) \\ &+ \frac{h}{2\sigma^2}I\left(\mathbf{F}_v \neq \mathbf{0}\right),\end{aligned} \tag{16}$$

where $\mathbf{K}_v^T$ is an $k \times |v|$ matrix obtained from $\mathbf{K}^T$ by concatenating the columns in the index set v. The separability allows us to minimize (13) with respect to $\mathbf{F}$ by minimizing each $j_v(\mathbf{F}_v)$ with respect to $\mathbf{F}_v$ separately. Due to the non-differentiability of $I(\cdot)$, care must be taken when performing the minimization. First, note that for $\mathbf{F}_v = \mathbf{0}$ we have $j_v(\mathbf{F}_v) = 0$ while for all other values the objective function is given by

$$j_v(\mathbf{F}_v) = -\frac{1}{\sigma^2}\mathrm{tr}\left(\mathbf{F}_v\mathbf{K}_v^T\right) + \frac{\lambda}{2\sigma^2}\mathrm{tr}\left(\mathbf{F}_v\mathbf{A}\mathbf{F}_v^T\right) + \frac{h}{2\sigma^2} \quad \text{if } \mathbf{F}_v \neq \mathbf{0}. \tag{17}$$

As $I(\cdot)$ is only non-differentiable at zero, we can minimize (17) by computing the derivative with respect to $\mathbf{F}_v$ and equating to zero. By doing so and by comparing the value of the objective function with the value when $\mathbf{F}_v = 0$, we obtain the following minimizer

$$\mathbf{F}_v^* = \mathbf{K}_v\mathbf{A}^{-1}I\left(h < \frac{1}{\lambda}\mathrm{tr}\left(\mathbf{K}_v\mathbf{A}^{-1}\mathbf{K}_v^T\right)\right). \tag{18}$$

To minimize (13) with respect to σ^2, we compute the derivate and equate to zero, resulting in

$$(\sigma_2)^* = \frac{1}{M}\left[\lambda\mathrm{tr}\left(\mathbf{F}\mathbf{A}\mathbf{F}^T\right) - 2\mathrm{tr}\left(\mathbf{F}\mathbf{K}^T\right) + a\right], \tag{19}$$

where $a = \mathrm{tr}(\mathbf{S}_\mathbf{y}) + h\rho(\mathbf{F})$ and we use the $\mathbf{F}$ obtained in the previous iteration.

To summarize the algorithm: In the minorize-step, we compute $\mathbf{W}$, $\mathbf{\Gamma}$, $\mathbf{A}$ and $\mathbf{K}$ and in the maximize-step we update $\mathbf{F}$ and σ^2 using (18) and (19), respectively. When the algorithm is used for regression, we use the conditional expectation of $\mathbf{u}_t$ from (7) to obtain

$$\tilde{\mathbf{x}}_t = \mathbf{F}\mathbf{W}^{-1}\mathbf{F}^T\mathbf{G}^T\mathbf{y}_t. \tag{20}$$

Finally, we point out an interesting connection to the svnPCA0 method [5]. Namely, for the case $\mathbf{G} = \mathbf{I}$, we have $\gamma = 1$ and $\mathcal{J}_{\mathrm{MM}}(\mathbf{F}) = \mathcal{L}_{\mathbf{Y},\mathbf{U}}(\mathbf{F})$, i.e., our EM-MM algorithm is equivalent to the EM algorithm from [5], which can therefore be considered a special case of the EM-MM algorithm developed here.

3. SIMULATION RESULTS

Sparse Variable PCA: In a first simulation, we consider the nPCA case where $\mathbf{G} = \mathbf{I}$. In this case, our method is equivalent to svnPCA0 [5]. To verify this fact, we compare our method to an implementation of svnPCA0 and perform an experiment similar to the ones in [5]. Specifically, we generate data with $M = 1024$, $T = 100$, $k = 10$, where the true loading matrix has 400 non-zero rows and is given by $\mathbf{F} = \left[\mathbf{V}\mathbf{S}\mathbf{U}^T, \mathbf{0}_{10\times 624}\right]^T$, where $\mathbf{V}$, and $\mathbf{U}$ are random orthonormal matrices with size 400×10 and 10×10, respectively. The matrix $\mathbf{S}$ is diagonal with entries $50^2, 45^2, \ldots, 5^2$. This loading matrix is then used with (1) to generate observations with $\sigma^2 = 500^2$. We then test the detection performance of svnPCA0 and the proposed method (ℓ_0-LSPCA) by using 50 logarithmically spaced regularization parameter settings in the range 10^3 to 10^6 and for each setting, we perform 20 simulations to compute the true positive rate (TPR) and false positive rate (FPR) from the estimated support of each method. As in [5] the proposed method and svnPCA0 are both initialized using the ML solution. As a reference, we also include the variance sorting procedure described in [5], for which the support is estimated by simply computing the variance $\sigma_i^2 = \sum_{t=1}^T (\mathbf{y}_t)_i^2$ and applying a threshold. By doing so, we compute a receiver operating characteristic (ROC) curve for each algorithm, shown in Fig. 1. As expected, the proposed method and svnPCA0 give virtually identical results. Small differences are due to slightly different implementations and numerical accuracy. It should be pointed out that while the methods give the same results, svnPCA0 is computationally more efficient in this case as the proposed method takes into account $\mathbf{G}$ which leads to unnecessary multiplications.

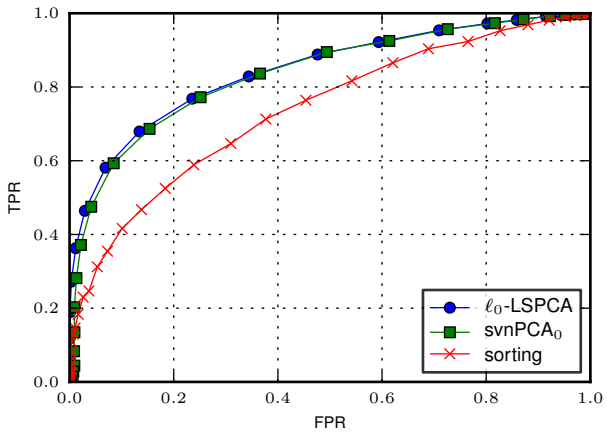

Fig. 1: ROC curves for the proposed method (ℓ_0-LSPCA), svnPCA0, and the sorting procedure.

Sparse Regression: In this experiment, we apply the proposed method to the regression problem and compare its performance to penalized regression methods which estimate $\mathbf{X}$ directly. The forward operator $\mathbf{G}$ is obtained from a uniform spherical ensemble, i.e., the columns of $\mathbf{G}$ are drawn from a uniform distribution on the M-sphere with radius 1. The true signal $\mathbf{X}$ of size 200×100 is generated using $\mathbf{X} = \mathbf{U}\mathbf{V}$, where $\mathbf{U}$ is a 200×5 matrix with 10 randomly selected non-zero rows for which we draw the entries from a zero-mean normal distribution with unit variance, and $\mathbf{V}$ is a 5×100 matrix with elements draw from the same distribution. We generate observations using (3) with a noise variance such that the signal-to-noise ratio (SNR) is 10dB ($\mathrm{SNR} = 10\log\|\mathbf{G}\mathbf{X}\|_\mathcal{F}^2/(TM\sigma^2)$). We include the following methods in the comparison: The proposed method (ℓ_0-LSPCA) for which we use $h = 1.0$ and supply the algorithm with the true noise variance σ^2 instead of trying to estimate it from the data, which is known to work poorly for sparse regression problems [10]. Due to the non-convexity of our method, a good starting point is important. We find that the method performs well when initialized using our previously developed ℓ_1 based method [17], which we also include in the comparison (ℓ_1-LSPCA) also with $h = 1.0$ and the true noise variance supplied to the method. For both the proposed method and ℓ_1-LSPCA, we use 3 settings for the parameter k, namely 5, 10, and 25. In addition, we also include $\ell_1\ell_2$-norm penalized regression in the comparison, for which we selected the regularization parameter by running the algorithm for a large number of values and retaining the one yielding the best over-all performance, which was $\alpha = 0.06 * \alpha_{\max}$, where $\alpha_{\max}$ is the $\ell_\infty\ell_2$-norm of $\mathbf{G}^T\mathbf{Y}$ [18]. To reduce the bias introduced by the ℓ_1 norm, it is common to use the $\ell_1\ell_2$-norm penalized method to detect the support and then compute a least-squares solution for the support. We also include this method ($\ell_1\ell_2$ support) with $\alpha = 0.22 * \alpha_{\max}$. Finally, we include the M-SBL method [10], which is a Bayesian method for estimating $\mathbf{X}$. The M-SBL method is supplied with the true noise

variance, which results in a high reconstruction performance for the scenarios considered here.

The reconstruction error $\|\hat{\mathbf{X}}-\mathbf{X}\|_{\mathcal{F}}^2/\|\mathbf{X}\|_{\mathcal{F}}^2$ when the number of observations M is varied is shown in Fig. 2. The proposed method clearly provides superior results. It is interesting to note that ℓ_1-

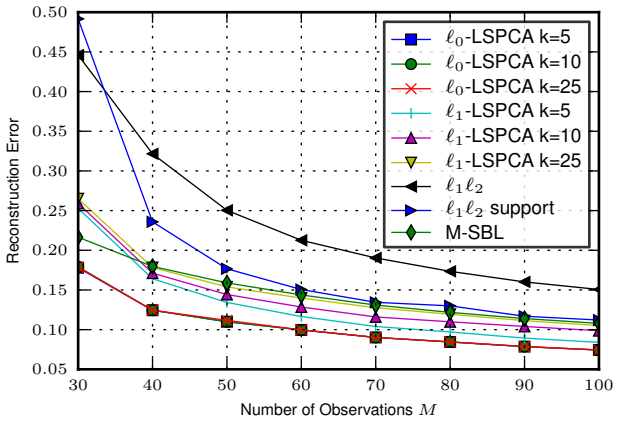

Fig. 2: Reconstruction error versus the number of observations M, $\mathbf{X}$ contains 10 non-zero rows and $\mathrm{rank}\,(\mathbf{X}) = 5$. Results are averages over 100 simulations.

LSPCA performs best when $k = 5 = \mathrm{rank}\,(\mathbf{X})$ but the reconstruction quality decreases when k is increased, while the proposed method exhibits the same high estimation quality regardless of the setting for k. This difference is likely due to the bias introduced by the ℓ_1-norm in ℓ_1-LSPCA.

4. APPLICATION TO MEG DATA

Sparse regression methods have proven their utility in localizing cortical sources from magnetoencephalography (MEG) data [15, 16]. In this section, we apply the proposed method to this problem using a publicly available dataset, which is part of the MNE software [19]. The data is from an experiment where a pure tone was applied to the left ear of the subject and was acquired using a 306-channel Elekta Neuromag Vectorview MEG system using a sampling rate of 600Hz. The MNE software was used to preprocess the data as follows: omitting one channel due to artifacts, 40Hz low-pass filtering, computation of an evoked response by averaging 63 trials from 0ms to 300ms after stimulus onset ($T = 181$). The gain matrix $\mathbf{G}$ of size 305×22494, corresponding to 7498 evenly distributed locations on the neocortex with three orthogonal current dipoles at each location, was computed using a subject specific boundary element model (BEM). Finally, a noise covariance matrix was estimated from 200ms data segments prior to the stimulus onset, which was used to whiten the evoked response and the gain matrix. The gain matrix was scaled by a factor of 10^{-8}, as the dipoles typically have small magnitudes in the order of a few nAm.

We use our ℓ_1-norm based method [17] with $h = 200$, $\sigma^2 = 1$, $k = 5$, to obtain an initial solution for the proposed method. The regularization parameter value was chosen smaller than in [17] in order to obtain a larger number of active dipoles in the initial solution (423 total.). The proposed method was then applied using $h = 1000$, $\sigma^2 = 1$, $k = 5$, resulting in a solution with 18 active dipoles corresponding to 6 locations located in the contralateral and ipsilateral auditory cortices. Note that for both methods, the three

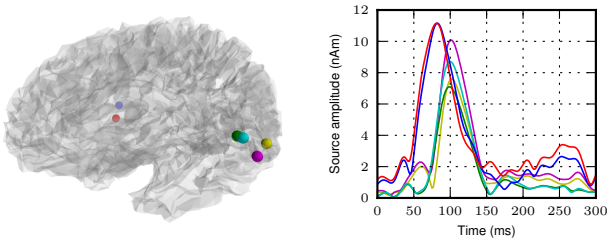

Fig. 3: MEG results: Dipole locations (left) and dipole magnitudes obtained by combining the 3 spatial components (right). The corresponding dipole location and magnitude are shown in the same color.

dipoles belonging to a single spatial location were grouped together, i.e., we used $|\mathcal{V}| = 7498$, $|v| = 3\,\forall\, v \in \mathcal{V}$, which means the dipoles at the same location will be jointly active or inactive.

Results are shown in Fig. 3. When compared to the results shown in [17] where we used our previous ℓ_1-norm penalized method, the proposed method was able to remove spurious dipoles in the right hemisphere. Furthermore, the dipole magnitudes estimated by the proposed method are significantly larger and more equal between the left and right hemispheres. Importantly, the activity in the right hemisphere, contralateral to the stimulus, is stronger and earlier than that in the ipsilateral left hemisphere.

5. CONCLUSIONS

In this work we introduce a method for sparse variable principal component analysis (PCA) in latent space. The method uses a vector ℓ_0 penalty to zero-out groups of variables, which enables it to work even for the case when the latent space is higher dimensional than the data space. We base our method on the maximization of the penalized log-likelihood, for which we derive a coupled EM-MM algorithm. Interestingly, the maximization step has a closed-form solution, which allows for efficient computation. Another interesting property is that for the special case when the latent- and data space are identical, our method is equivalent to an existing vector ℓ_0 penalized PCA method that operates in data space [5] and we use simulations to demonstrate the equivalence.

A problem related to the one considered here is underdetermined regression, as encountered in, e.g., compressive sensing [7]. We used simulations to demonstrate that our method can also be applied to this problem and provides a superior estimation quality when compared to existing methods. As a practical example, we applied our method to the problem of localizing cortical sources from magnetoencephalography (MEG) data, where it correctly localizes the sources and provides a solution with reduced amplitude bias when compared to previous results [17].

It is important to note that while we empirically demonstrate that our method performs well even in the underdetermined case, no theoretical results currently exist that establish performance guarantees, as it is the case for regression [7]. Furthermore, the regularization parameter h and the parameter k were heuristically selected in this work. For the case when $\mathbf{G} = \mathbf{I}$, the parameters can be chosen using the Bayesian information criterion (BIC) [5]. Parameter selection and a theoretical analysis will be addressed in future work.

6. REFERENCES

[1] D. N. Lawley, "A modified method of estimation in factor analysis and some large sample results," in *Proc. Uppsala Symp. Psycholog. Factor Analysis*, Sweden, 1953, vol. 17, pp. 35–42.

[2] D. B. Rubin and D. T. Thayer, "Em algorithms for ml factor analysis," *Psychometrika*, vol. 47, no. 1, pp. 69–76, 1982.

[3] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 61, no. 3, pp. 611–622, 1999.

[4] A. P. Dempster, N. M. Laird, D. B. Rubin, et al., "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977.

[5] M. O. Ulfarsson and V. Solo, "Vector l0 Sparse Variable PCA," *IEEE Trans. Signal Process.*, vol. 59, no. 5, pp. 1949–1958, 2011.

[6] D. L. Donoho, M. Elad, and V. N. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," *IEEE Trans. Inf. Theory*, vol. 52, no. 1, pp. 6–18, 2005.

[7] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inf. Theory*, vol. 52, no. 2, pp. 489–509, 2006.

[8] Ming Yuan and Yi Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, no. 1, pp. 49–67, 2006.

[9] Akila J Seneviratne and Victor Solo, "On vector l0 penalized multivariate regression," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 3613–3616.

[10] D. P. Wipf and B. D. Rao, "An Empirical Bayesian Strategy for Solving the Simultaneous Sparse Approximation Problem," *IEEE Trans. Signal Process.*, vol. 55, no. 7, pp. 3704–3716, 2007.

[11] D. P. Wipf, B. D. Rao, and S. Nagarajan, "Latent variable Bayesian models for promoting sparsity," *IEEE Trans. Inf. Theory*, vol. 57, no. 9, pp. 6236–6255, 2011.

[12] J. M. Ortega and W. C. Rheinboldt, "Iterative solution of nonlinear equations in several variables," *Academic Press, New York*, vol. 19702, pp. 504, 1970.

[13] J. De Leeuw and W. J. Heiser, "Convergence of correction matrix algorithms for multidimensional scaling," *Geometric representations of relational data*, pp. 735–752, 1977.

[14] M. Hämäläinen, R. Hari, R. J. Ilmoniemi, J. Knuutila, and O. V. Lounasmaa, "Magnetoencephalography-theory, instrumentation, and applications to noninvasive studies of the working human brain," *Rev. Mod. Phys.*, vol. 65, no. 2, pp. 413–497, Apr 1993.

[15] 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.

[16] 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*, , no. 7, pp. 1937–1961, Mar. 2012.

[17] M. Luessi, M.S. Hämäläinen, and V. Solo, "Sparse component selection with application to MEG source localization," in *2013 IEEE 10th International Symposium on Biomedical Imaging (ISBI)*, 2013, pp. 556–559.

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

[19] A. Gramfort, M. Luessi, E. Larson, D. Engemann, D. Strohmeier, C. Brodbeck, L. Parkkonen, and M. Hämäläinen, "MNE software for processing MEG and EEG data," *NeuroImage*, , no. 0, pp. –, 2013.