Empirical Bayesian Models of L1/L2 Mixed-norm Constraints

Deirel Paz-Linares, Mayrim Vega-Hernández and Eduardo Martínez-Montes
Neuroinformatics Department, Cuban Neuroscience Center, Havana, Cuba

1 OBJECTIVES

Inverse problems are common in neuroscience and neurotechnology, where usually a small amount of data is available with respect to the large number of parameters needed for modelling the brain activity. Classical examples are the EEG/MEG source localization and the estimation of effective brain connectivity. Many kinds of constraints or prior information have been proposed to regularize these inverse problems. Combination of smoothness (L2 norm-based penalties) and sparseness (L1 norm-based penalties) seem to be a promising approach due to its flexibility, but the estimation of optimal weights for balancing these constraints became a critical issue (Vega-Hernández et al., 2008). Two important examples of constraints that combine L_1/L_2 norms are the Elastic Net (Vega-Hernández et al., 2008) and the Mixed-Norm L_{12} (MxN, Gramfort et al., 2012). The latter imposes the properties along different dimensions of a matrix inverse problem. In this work, we formulate an empirical Bayesian model based on an MxN prior distribution. The objective is to pursue sparse learning along the first dimension (along rows) preserving smoothness in the second dimension (along columns), by estimating both parameter and hyperparameters (regularization weights).

2 METHODS

The matrix linear Inverse Problem consists in inferring an *S*x*T* parameter matrix $\boldsymbol{J}$ in the model $\boldsymbol{V} = \boldsymbol{KJ} + \boldsymbol{\varepsilon}$, where $\boldsymbol{V}$ (data), $\boldsymbol{\varepsilon}$ (noise) are *N*x*T*, $\boldsymbol{K}$ is *N*x*S*, with *N*<<*S*, making it an ill-posed problem due to its non-uniqueness. One approach to address this problem is the Tikhonov regularization which uses a penalty function $P(\boldsymbol{J})$ to find the inverse solution through a penalized least-squares (PLS) regression $\hat{\boldsymbol{J}} = argmin\{\|\boldsymbol{V} - \boldsymbol{KJ}\|_2^2 + \alpha P(\boldsymbol{J})\}$, where α is the regularization parameter. Another approach is the Bayesian theory, where the solution maximizes the posterior probability density function (*pdf*), given by the Bayes equation: $p(\boldsymbol{J}, \beta, \alpha|\boldsymbol{V}) \propto p(\boldsymbol{V}|\boldsymbol{J}, \beta)p(\boldsymbol{J}|\alpha)$, which is largely equivalent to the PLS model if we set the likelihood of the data to $p(\boldsymbol{V}|\boldsymbol{J}, \beta) = e^{-\frac{1}{2}tr((\boldsymbol{V}-\boldsymbol{KJ})'(\beta\boldsymbol{I})^{-1}(\boldsymbol{V}-\boldsymbol{KJ}))}/(2\pi|\beta\boldsymbol{I}|)^{\frac{T}{2}}$, and the prior distribution of the parameters as an exponential function $p(\boldsymbol{J}|\alpha) = e^{-\alpha P(\boldsymbol{J})}/Z$, where Z is a normalizing constant.

The first approach has led to development of fast and efficient algorithms for a wide range of solvers $P(\boldsymbol{J})$, but α is determined heuristically using information criteria which often do not provide optimal values. On the other hand, Bayesian approach allows inference on the hyperparameters α and β but frequently involving numerical Monte Carlo calculations that makes it very slow and computationally intensive. However, recent developments of approximate models such as Variational and Empirical Bayes, allow for fast computation of complex models.

In this work, we propose to use the squared Mixed-Norm penalty for the parameters, which is defined as the L_2 norm of the vector obtained from the L_1 norms of all columns $\{\boldsymbol{J}_t\}_{t=1}^T$ of $\boldsymbol{J}$ (Gramfort et al., 2012) and can be written as $\|\boldsymbol{J}\|_{\boldsymbol{w};1,2}^2 = \sum_t \|\boldsymbol{W}\boldsymbol{J}_t\|_1^2$, where $\boldsymbol{W} = diag(\boldsymbol{w})$ is the weights (positive) diagonal matrix. The prior *pdf* for this penalty represents a Markov Random Field (MRF) where the states of the variable $\{J_{it}\}_{i=1}^S$ are not separable.

$$p(\boldsymbol{J}_t|\alpha) \propto \left\{\prod_{i=1}^{S} e^{-\alpha w_i^2 J_{it}^2}\right\}\left\{\prod_{i=1}^{S-1}\prod_{k=i+1}^{S} e^{-2\alpha w_i w_k |J_{it}||J_{kt}|}\right\} \quad (1)$$

Using Empirical Bayes, we first transform this MRF into a Bayesian network, to arrive to a hierarchical model (figure 1) by reformulating the *pdf* of each J_{it} as $p(J_{it}|J_{i^c t}, \alpha)$, where $J_{i^c t} = \{J_{kt}\}_{k \neq i}$. In this way, the information received by J_{it} from $J_{i^c t}$ is contained in an auxiliary magnitude $\delta_{it} = \sum_{k \neq i} w_k |J_{kt}|$, leading to a Normal-Laplace joint *pdf*:

$$p(\boldsymbol{J}_t, \boldsymbol{\delta}_t|\alpha) = \left\{\prod_{i=1}^{S} \frac{e^{-\alpha w_i^2 J_{it}^2}}{Z_i} La\left(\frac{J_{it}}{2\alpha w_i \delta_{it}}\right)\right\} p(\boldsymbol{\delta}_t|\alpha) \quad (2)$$

$$p(\boldsymbol{\delta}_t|\alpha) \propto e^{-\alpha\|(\mathbf{1}_S/(s-1)-\mathbf{I}_S)\boldsymbol{\delta}_t\|_1^2}/Z\,;\, \mathbf{1}_S = ones(S\text{x}S)$$

Then, using the scaled mixture of Gaussians for the Normal-Laplace *pdf* (Li and Lin 2010), a hyper-

Paz-Linares D., Vega-Hernández M. and Martínez Montes E..
Empirical Bayesian Models of L1/L2 Mixed-norm Constraints.

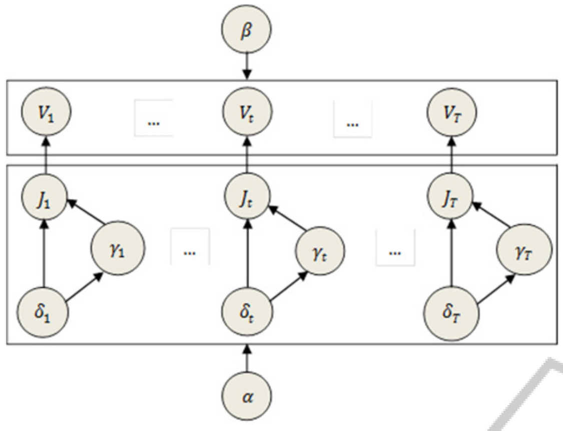

Figure 1: Hierarchical diagram of the Bayesian network obtained for the mixed-norm L_{12} model.

parameter $\boldsymbol{\gamma}_t$ is introduced to complete the joint prior $p(\boldsymbol{J}_t, \boldsymbol{\gamma}_t, \boldsymbol{\delta}_t|\alpha) = N(\boldsymbol{J}_t|\boldsymbol{0}, \boldsymbol{\Lambda}_t)p(\boldsymbol{\gamma}_t|\boldsymbol{\delta}_t, \alpha)p(\boldsymbol{\delta}_t|\alpha)$ where the variance matrix for $\boldsymbol{J}_t$ is diagonal $\boldsymbol{\Lambda}_t = diag\left[\frac{1}{2\boldsymbol{w}^2}\left(\frac{1}{\alpha} - \frac{\delta_t^2}{\gamma_t}\right)\right]$ and the *pdf* for $\boldsymbol{\gamma}_t$ is the truncated Gamma $p(\boldsymbol{\gamma}_t|\boldsymbol{\delta}_t, \alpha) = TG(1/2,1, [\alpha\boldsymbol{\delta}_t^2, \infty])$. The vector $\boldsymbol{\delta}_t$ can be updated from the $\boldsymbol{J}_t$ estimated in the previous iteration. Then we can choose gamma non-informative priors $p(\alpha)$ and $p(\beta)$ and rewrite the joint posterior *pdf* as:

$$p(\boldsymbol{J}_t, \boldsymbol{\gamma}_t, \alpha, \beta\,|\boldsymbol{V}) \propto p(\boldsymbol{V}|\boldsymbol{J}_t, \beta)p(\boldsymbol{J}_t, \boldsymbol{\gamma}_t|\alpha)p(\alpha)p(\beta) \quad (3)$$

The maximum *a posteriori* estimate for the model parameters $\boldsymbol{J}_t$ is easily derived from the Gaussian *pdf* $p\left(\boldsymbol{J}_t\middle|\boldsymbol{V}_t, \boldsymbol{\gamma}_t, \alpha, \beta\right) \propto N\left(\boldsymbol{V}_t\middle|\boldsymbol{K}\boldsymbol{J}_t, \beta\boldsymbol{I}\right)N\left(\boldsymbol{J}_t\middle|\boldsymbol{0}, \boldsymbol{\Lambda}_t\right)$ $=N(\boldsymbol{J}_t/\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$, with posterior mean and variance:

$$\boldsymbol{\mu}_t = \frac{\boldsymbol{\Sigma}_t\boldsymbol{K}'\boldsymbol{V}_t}{\beta}; \quad \boldsymbol{\Sigma}_t = (\frac{\boldsymbol{K}'\boldsymbol{K}}{\beta} + \boldsymbol{\Lambda}_t^{-1})^{-1} \quad (4)$$

The estimates for the hyperparameters are achieved by maximizing the evidence (evidence procedure), also known as the type II likelihood $\mathcal{L}$, which is obtained by integrating out the parameters $\boldsymbol{J}_t$ from the joint posterior *pdf* in (3).

$$\mathcal{L} = \sum_t \left\{\frac{1}{2}\ln|\boldsymbol{C}_t| + \boldsymbol{V}_t'\boldsymbol{C}_t^{-1}\boldsymbol{V}_t\right\} - \ln p(\boldsymbol{\gamma}_t, \alpha, \beta) \quad (5)$$

with $\boldsymbol{C}_t = \frac{\boldsymbol{I}}{\beta} + \boldsymbol{K}\boldsymbol{\Lambda}_t^{-1}\boldsymbol{K}'$. Closed estimates maximizing $\mathcal{L}$ cannot be obtained due to the nonlinear form of $\boldsymbol{C}_t$ but it can be rearranged in terms of $\boldsymbol{\mu}_t$, $\boldsymbol{\Sigma}_t$ and other differentiable expressions of $(\boldsymbol{\gamma}, \alpha, \beta)$, which after differentiation leads to the following updates for the hyperparameters:

$$\gamma_{it} = \alpha\delta_{it}^2 + \eta_{it}; \eta_{it} = \sqrt{\frac{1}{16} + \alpha^2\delta_{it}^2 w_i^2(\mu_{it}^2 + \Sigma_{ii})} - \frac{1}{4} \quad (6)$$

$$\alpha = \frac{\sum_t\sum_i\left[\frac{\gamma_{it}}{\eta_{it}} + 2\frac{\gamma_{it} - \eta_{it}}{I(z)} + 1\right]}{2\sum_t\sum_i\frac{w_i^2(\mu_{it}^2 + \Sigma_{ii})\gamma_{it}^2}{\eta_{it}^2} + 2\sum_t\|\boldsymbol{W}|\boldsymbol{\mu}_t|\|_1^2} \quad (6)$$

$$\beta = \frac{\sum_t\|\boldsymbol{V}_t - \boldsymbol{K}\boldsymbol{\mu}_t\|_2^2}{\sum_t\sum_i[\Sigma_{ii}/\Lambda_{ii}]}$$

where $I(z = (\alpha^{old}\delta_{it}^2)^{1/2})$ is related with the normalizing constant for the *TG*. Similar to the Relevance Vector Machine (Tipping 2001), the hyperparameter $\boldsymbol{\gamma}_t$ controls the sparseness of the solution, since it cannot take values equal or below $\alpha\delta_{it}^2$, allowing the J_{it} where this conditions holds to be set to zero. The iterative estimation of parameters and hyperparameters with these formulae (4 and 6) is equivalent to an EM algorithm. Here we illustrate how this model works with this algorithm (using in-house code) but future studies will focus in deriving faster and more efficient algorithms for estimating the model.

3 RESULTS

Simulations of a $\boldsymbol{J}$ matrix (800x30) with different waveforms (along columns) in well-localized rows, were performed to test the ability of the model to estimate simultaneously different levels of sparseness and smoothness in both dimensions (figure 2, left). The inverse solution was obtained (figure 2, right) from data generated using a random design matrix $\boldsymbol{K}$ (100x800, K_{ij}~N(0,4), SNR=30db), converging after 150 iterations (in about 5 min). Estimation of relevant rows is shown in figure 3.
We also considered a more realistic simulation of the EEG inverse problem. A ring of 736 cortical generators (voxels defined in MNI brain atlas), was used to simulate 3 spatio-temporal sources. The electric lead field was computed as the

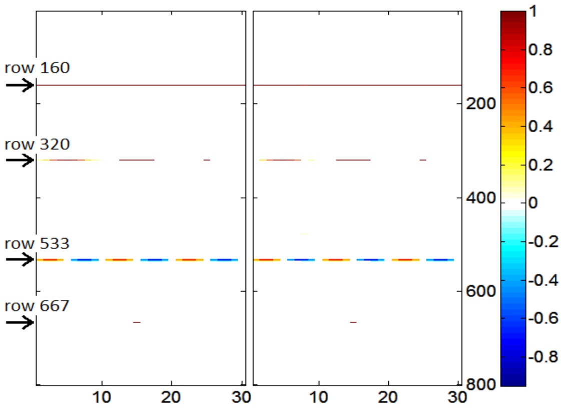

Figure 2: Left: Simulated matrix. Right: Estimated $\boldsymbol{J}$ with the Bayesian MxN model, using an EM algorithm.

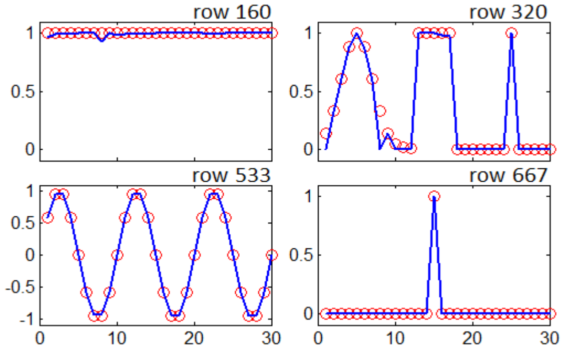

Figure 3: Simulated (true, red circles) and estimated (blue line) inverse solution for rows arrowed in figure 2.

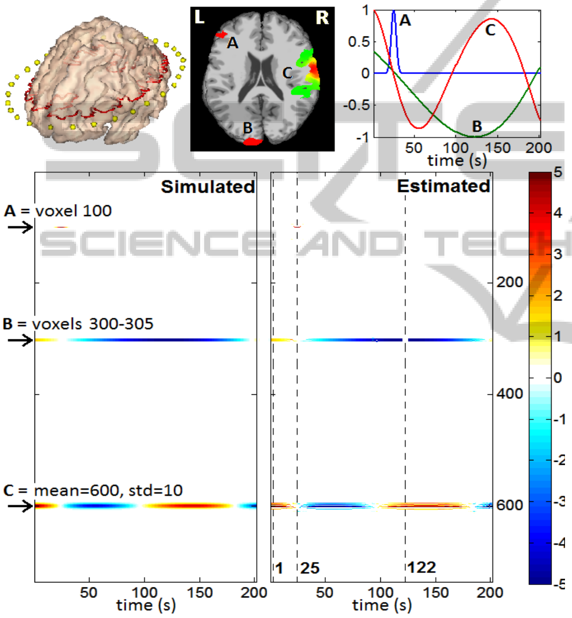

Figure 4: Realistic simulation (ring of cortical generators). Parameter matrix was formed by simulated sources A (1 voxel, temporal bell), B (5 voxels, temporal sinusoid) and C (spatial bell, temporal sinusoid). Dashed lines mark selected time points in the estimated inverse solution.

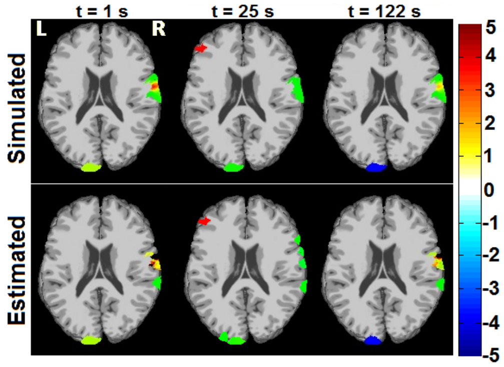

Figure 5: Simulated and estimated EEG sources in a ring of cortical generators at the maximum value for sources A (25s), B (122s) and C (1s) as marked in figure 4.

transformation matrix from the sources to 31 recording channels (figure 4, top row), and noise was added (SNR=30db). Figure 4 (bottom) shows the matrix inverse solution estimated with the Bayesian MxN model, which converged in 150 iterations (about 15 min). Figure 5 shows the spatial maps for three relevant time points.

4 DISCUSSION

The use of the Normal-Laplace distribution as the parameters' prior *pdf*, theoretically allows to flexible estimation of parameters with sparse and smooth simultaneous behaviour. Here we proposed an Empirical Bayes solution to this analytically untreatable model. Simulations showed that the method is able to reconstruct solutions that are sparse along the first dimension and smooth along the second dimension. However, it cannot accurately recover non-sparse sources in the spatial dimension. The level of sparseness is controlled by just one parameter (α) for the whole map, possibly making it difficult to estimate situations when the level of sparseness changes with time. Also, the EM algorithm showed some instability and dependence on initial values. Although further validation is needed, future efforts will also aimed at improving the model to cope with time-varying sparseness and developing more efficient and robust algorithms.

REFERENCES

Vega-Hernández M, et al. (2008): Penalized leastsquares methods for solving the *EEG inverse problem. Stat Sin*18:1535–1551.

Alexandre-Gramfort, et al. (2012): Mixed-norm estimates for the M/EEG inverse problem using accelerated gradient methods. *Physics in Medicine and Biology* 57: 1937-1961.

Qing Li and Nan Lin. (2010): *The Bayesian Elastic Net. Bayesian Analysis 5*, Number 1: 151-170.

Michael E. Tipping. 2001: Sparse Bayesian Learning and the Relevance Vector Machine. *Journal of Machine Learning Research* 1: 211-244.