J. R. Statist. Soc. B (2014)
76, *Part* 2, *pp.* 463–483

Regularized matrix regression

Hua Zhou and Lexin Li

North Carolina State University, Raleigh, USA

[Received April 2012. Revised April 2013]

Summary. Modern technologies are producing a wealth of data with complex structures. For instance, in two-dimensional digital imaging, flow cytometry and electroencephalography, matrix-type covariates frequently arise when measurements are obtained for each combination of two underlying variables. To address scientific questions arising from those data, new regression methods that take matrices as covariates are needed, and sparsity or other forms of regularization are crucial owing to the ultrahigh dimensionality and complex structure of the matrix data. The popular lasso and related regularization methods hinge on the sparsity of the true signal in terms of the number of its non-zero coefficients. However, for the matrix data, the true signal is often of, or can be well approximated by, a low rank structure. As such, the sparsity is frequently in the form of low rank of the matrix parameters, which may seriously violate the assumption of the classical lasso. We propose a class of regularized matrix regression methods based on spectral regularization. A highly efficient and scalable estimation algorithm is developed, and a degrees-of-freedom formula is derived to facilitate model selection along the regularization path. Superior performance of the method proposed is demonstrated on both synthetic and real examples.

Keywords: Electroencephalography; Multi-dimensional array; Nesterov method; Nuclear norm; Spectral regularization; Tensor regression

1. Introduction

Modern scientific applications are frequently producing data sets where the sampling unit is not in the form of a vector but instead a matrix. Examples include two-dimensional digital imaging data, which record the quantized brightness value of a colour at rows and columns of pixels, and flow cytometric data, which contain the fluorescence intensity of multiple cells at multiple channels. Our motivating example is a study of an electroencephalography data set of alcoholism. The study consists of 122 subjects with two groups, an alcoholic group and a normal control group, and each subject was exposed to a stimulus. Voltage values were measured from 64 channels of electrodes placed on the subject's scalp for 256 time points, so each sampling unit is a 256×64 matrix. It is of scientific interest to study the association between alcoholism and the pattern of voltage over times and channels. The generalized linear model (GLM) (McCullagh and Nelder, 1983) offers a useful tool for that purpose, where the response Y is the binary indicator of alcoholic or control, and the predictors include the matrix-valued electroencephalography data $\mathbf{X}$ and possible covariate vector $\mathbf{Z}$ such as age and gender. However, the classical GLM deals with a vector of covariates, and the presence of matrix-type covariates poses fresh challenges to statistical analysis. First, naively turning a matrix into a vector results in an exceedingly large dimensionality; for instance, for the electroencephalography data, the dimension is $p = 256 \times 64 = 16384$, whereas the sample size

Address for correspondence: Hua Zhou, Department of Statistics, North Carolina State University, Box 8203, Raleigh, NC 27695-8203, USA.
E-mail: hua_zhou@ncsu.edu

 1369–7412/14/76463

is only $n = 122$. Second, vectorization destroys the wealth of structural information that is inherently possessed in the matrix data; for example it is commonly expected that the voltage values of the adjacent time points and channels are highly correlated. Given the ultrahigh dimensionality and the complex structure, *regularization* becomes crucial for the analysis of such data. In this paper, we propose a novel regularization solution for regression with matrix covariates, which efficiently tackles the ultrahigh dimensionality while preserving the matrix structure.

A wide variety of regularization methods have been developed in recent years. Among them, penalization has been playing a powerful role in stabilizing the estimates, improving the risk property and increasing the generalization power in classical regressions. Popular penalization techniques include the lasso (Tibshirani, 1996; Donoho and Johnstone, 1994), smoothly clipped absolute deviation (SCAD) (Fan and Li, 2001), the fused lasso (Tibshirani *et al.*, 2005), the elastic net (Zou and Hastie, 2005) and many others. For regressions with matrix covariates, a direct approach is first to vectorize the covariates and then to apply the classical penalization techniques. Regularization helps to alleviate the problem that the dimensionality far exceeds the sample size. However, this is unsatisfactory, since it fails to incorporate the matrix structural information. More importantly, the solution is based on a fundamental assumption that the true underlying signal is sparse in terms of the l_0-norm of the regression parameters. In matrix regressions, however, often the true signal is of, or can be well approximated by, a low rank structure. As such, sparsity is in terms of the *rank* of the matrix parameters, which is intrinsically different from sparsity in the number of non-zero entries.

To see how such a difference affects signal estimation in matrix regressions, we consider the following illustrative example. We generated a normal response Y with mean $\mu = \boldsymbol{\gamma}^{\mathrm{T}}\mathbf{Z} + \langle \mathbf{B}, \mathbf{X} \rangle$, and variance 1. $\mathbf{Z} \in \mathbb{R}^5$ denotes a usual vector of covariates with standard normal entries, and $\boldsymbol{\gamma} = (1, \ldots, 1)^{\mathrm{T}}$. $\mathbf{X} \in \mathbb{R}^{64 \times 64}$ denotes the matrix covariates, of which all entries are standard normal, and $\mathbf{B}$ is the coefficient matrix of the same size. $\mathbf{B}$ is binary, with the true signal region, which is a cross shape in our example, equal to 1 and the rest 0. The inner product between two matrices is defined as $\langle \mathbf{B}, \mathbf{X} \rangle = \mathrm{tr}(\mathbf{B}^{\mathrm{T}}\mathbf{X}) = \langle \mathrm{vec}(\mathbf{B}), \mathrm{vec}(\mathbf{X}) \rangle$, where $\mathrm{vec}(\cdot)$ is the vectorization operator that stacks the columns of a matrix into a vector. We sampled $n = 500$ instances $\{(y_i, \mathbf{x}_i, \mathbf{z}_i), i = 1, \ldots, 500\}$, and the goal is to identify $\mathbf{B}$ through a regression of y_i on $(\mathbf{x}_i, \mathbf{z}_i)$. We note that our problem differs from the usual edge detection or object recognition in imaging processing (Qiu, 2005, 2007). In our set-up, all elements of the image $\mathbf{X}$ follow the same distribution. The signal region is defined through the coefficient image $\mathbf{B}$ and needs to be inferred from the association between Y and $\mathbf{X}$ after adjusting for $\mathbf{Z}$. We also note that the total number of entries in $\mathbf{B}$ is $4096 = 64^2$, and the number of non-zero entries is 240 (about 5.8%). We applied two approaches. The first is the lasso to the vectorized $\mathbf{X}$, i.e. we solve the optimization problem

$$\min_{\mathbf{B}} \frac{1}{2} \sum_{i=1}^{n} (y_i - \boldsymbol{\gamma}^{\mathrm{T}}\mathbf{z}_i - \langle \mathbf{B}, \mathbf{x}_i \rangle)^2 + \lambda \|\mathrm{vec}(\mathbf{B})\|_1,$$

where $\|\mathrm{vec}(\mathbf{B})\|_1$ is the l_1-norm of the vectorized $\mathbf{B}$, and λ is the regularization parameter. Fig. 1(e) displays the Bayesian information criterion BIC along the lasso solution path, which suggests a model with the maximum number of predictors (500) that is allowed given the sample size. The parameter estimate under this model is shown Fig. 1(c), which appears far away from the truth. The second solution that we consider is penalizing the nuclear norm of $\mathbf{B}$, i.e. we solve

$$\min_{\mathbf{B}} \frac{1}{2} \sum_{i=1}^{n} (y_i - \boldsymbol{\gamma}^{\mathrm{T}}\mathbf{z}_i - \langle \mathbf{B}, \mathbf{x}_i \rangle)^2 + \lambda \|\mathbf{B}\|_*,$$

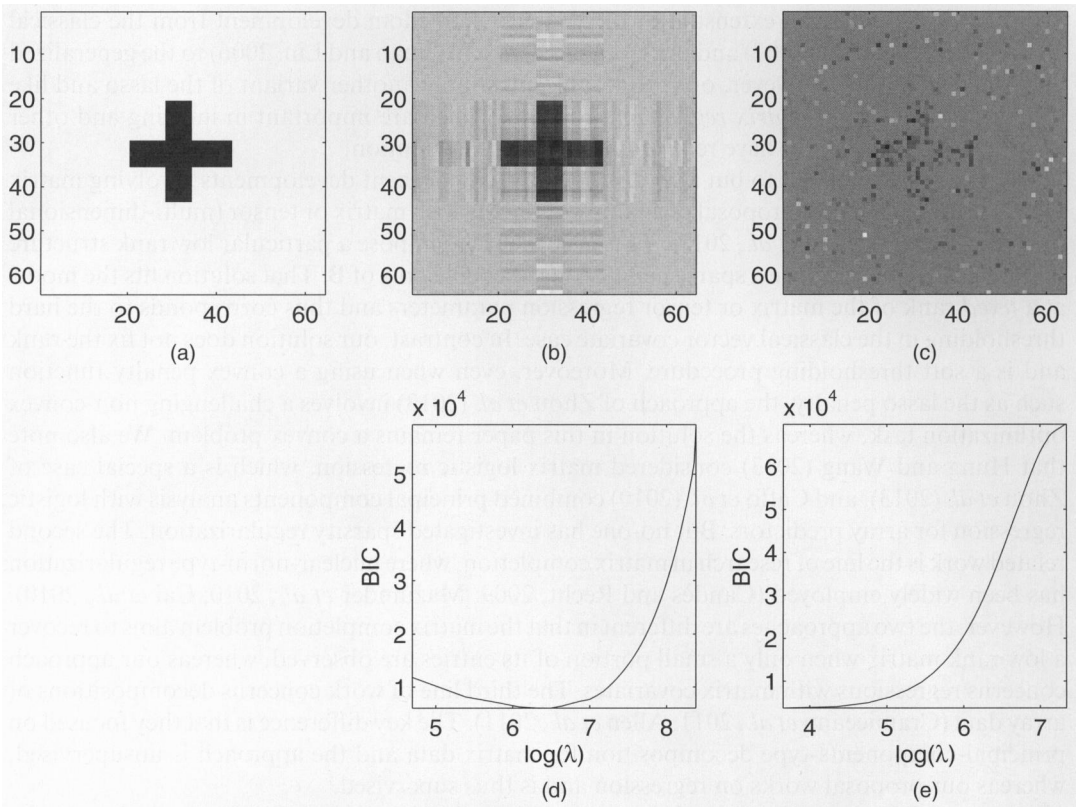

Fig. 1. Comparison of the nuclear-norm-regularized estimation with the classical lasso (the matrix covariate is 64 × 64, and the sample size is 500): (a) true signal, (b) nuclear-norm-regularized estimate with minimal BIC; (c) classical lasso estimate with minimal BIC; (d) BIC along solution paths for the nuclear norm regularization; (e) BIC along solution paths for classical lasso regularization

where the nuclear norm $\|\mathbf{B}\|_* = \Sigma_j \sigma_j(\mathbf{B})$, and $\sigma_j(\mathbf{B})$s are the singular values of the matrix $\mathbf{B}$. The nuclear norm $\|\mathbf{B}\|_*$ is a suitable measure of the 'size' of a matrix and is a convex relaxation of $\text{rank}(\mathbf{B}) = \|\sigma(\mathbf{B})\|_0$. This is analogous to the l_1-norm for a vector (Recht *et al.*, 2010). Fig. 1(d) displays BIC along the solution path of the nuclear norm penalized matrix regression, and Fig. 1(b) shows the corresponding estimate with minimal BIC. It is clearly seen that the nuclear norm estimate achieves a substantially better recovery than the lasso estimate. One might argue that the fused lasso (Tibshirani *et al.*, 2005) might give better recovery of such piecewise constant signals. However, there are numerous low rank signals, e.g. $(01\ldots01)^{\mathrm{T}}(10\ldots10)$, which are extremely non-smooth and would fail the fused lasso.

More generally, in this paper, we propose a family of regularized regression models with matrix covariates based on spectral regularization. Our contributions are multifold. First, we employ a spectral regularization formulation within the GLM framework. The resulting model works for a variety of penalization functions, including the lasso, elastic net, SCAD and many others, as well as different types of response variables, including normal, binary and count outcomes. Second, we develop a highly efficient and scalable algorithm for model estimation with explicit, non-asymptotic convergence rate. Such a highly scalable algorithm is critical for analysing large-scale and ultrahigh dimensional matrix data. Third, we derive the effective degrees of freedom of selected models, which are crucial for tuning the regularization parameter.

This can be viewed as an extension of the degrees-of-freedom development from the classical lasso model (Zou *et al.*, 2007) and the group lasso model (Yuan and Lin, 2006) to the generalized linear matrix model. However, our proposal is *not* simply another variant of the lasso and like techniques. We aim at *matrix regression* problems, which are important in imaging and other scientific applications but have received relatively little attention.

Our proposal is related to but also distinct from some recent developments involving matrix data. The first is a recent proposal of a family of GLMs with matrix or tensor (multi-dimensional array) covariates (Zhou *et al.*, 2013). The basic idea is to impose a particular low rank structure on $\mathbf{B}$ and then to introduce a sparse penalty on the coefficients of $\mathbf{B}$. That solution fits the model at a *fixed* rank of the matrix or tensor regression parameters and thus corresponds to the hard thresholding in the classical vector covariate case. In contrast, our solution does not fix the rank and is a soft thresholding procedure. Moreover, even when using a convex penalty function such as the lasso penalty, the approach of Zhou *et al.* (2013) involves a challenging non-convex optimization task, whereas the solution in this paper remains a convex problem. We also note that Hung and Wang (2013) considered matrix logistic regression, which is a special case of Zhou *et al.* (2013), and Caffo *et al.* (2010) combined principal components analysis with logistic regression for array predictors. But no-one has investigated sparsity regularization. The second related work is the line of research in matrix completion, where nuclear-norm-type regularization has been widely employed (Candès and Recht, 2009; Mazumder *et al.*, 2010; Cai *et al.*, 2010). However, the two approaches are different in that the matrix completion problem aims to recover a low rank matrix when only a small portion of its entries are observed, whereas our approach concerns regressions with matrix covariates. The third line of work concerns decompositions of array data (Crainiceanu *et al.*, 2011; Allen *et al.*, 2011). The key difference is that they focused on principal-components-type decomposition for matrix data and the approach is unsupervised, whereas our proposal works on regression and is thus supervised.

The rest of the paper is organized as follows. We formulate the spectral regularization for matrix regression in Section 2 and develop a highly scalable algorithm for the associated optimization in Section 3. We derive the degrees-of-freedom formula in Section 4 and investigate the numerical performance of the proposed method in Section 5. We conclude the paper with a discussion of potential future research in Section 6. All technical proofs are delegated to Appendix A.

2. Spectral regularization

We start with some notation. For any matrix $\mathbf{B} \in \mathbb{R}^{p_1 \times p_2}$, $\sigma(\mathbf{B}) = (\sigma_1(\mathbf{B}), \ldots, \sigma_q(\mathbf{B}))$, $q = \min\{p_1, p_2\}$, denotes the vector of singular values of $\mathbf{B}$ arranged in decreasing order, i.e. $\sigma_1(\mathbf{B}) \geqslant \sigma_2(\mathbf{B}) \geqslant \ldots \geqslant \sigma_r(\mathbf{B}) > \sigma_{r+1}(\mathbf{B}) = \ldots = \sigma_q(\mathbf{B}) = 0$, where $r = \mathrm{rank}(\mathbf{B})$. For any scalar function f, ∇f is the column gradient vector and $\mathrm{d}^2 f$ is the Hessian matrix. Let Y denote the response variable, $\mathbf{Z} \in \mathbb{R}^{p_0}$ the vector covariate, $\mathbf{X} \in \mathbb{R}^{p_1 \times p_2}$ the two-dimensional matrix covariate and $(y, \mathbf{x}, \mathbf{z})$ their sample instances.

We consider the GLM set-up, where Y belongs to an exponential family with probability mass function or density

$$p(y|\mathbf{x}, \mathbf{z}) = \exp\left\{ \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right\}.$$

Here θ is the natural parameter and $\phi > 0$ is a dispersion parameter. $a(\phi)$ usually has the form $a(\phi) = \phi w^{-1}$, where w is a prespecified weight. The first conditional moment is $E(Y|\mathbf{X}, \mathbf{Z}) = \mu = b'(\theta)$, and μ is of the form

$$q(\mu) = \gamma^{\mathrm{T}}\mathbf{Z} + \langle \mathbf{B}, \mathbf{X} \rangle, \tag{1}$$

where q is a known link function, $\gamma \in \mathbb{R}^{p_0}$, and $\mathbf{B} \in \mathbb{R}^{p_1 \times p_2}$. For simplicity, we drop the vector covariate $\mathbf{Z}$ and its associated parameter γ in subsequent development. However, the results can be readily extended to incorporate $\mathbf{Z}$ and γ. Also, we consider only GLMs with a univariate response. Extensions to more complex models such as quasi-likelihood models and multivariate responses are straightforward.

We consider the spectral regularization problem

$$\min_{\mathbf{B}} h(\mathbf{B}) = l(\mathbf{B}) + J(\mathbf{B}), \tag{2}$$

where $l(\mathbf{B})$ is a loss function; for the GLM, we use the negative log-likelihood as the loss. $J(\mathbf{B}) = f \circ \sigma(\mathbf{B})$, where $f : \mathbb{R}^q \to \mathbb{R}$ is a function of the singular values of $\mathbf{B}$. The least squares loss combined with $f(\mathbf{w}) = \lambda \Sigma_{j=1}^{q} |w_j|$ corresponds to the special case of the nuclear norm regularization problem that was considered in Section 1. In general, for sparsity of the spectrum, f takes the general form

$$f(\mathbf{w}) = \sum_{j=1}^{q} P_\eta(|w_j|, \lambda),$$

where P is a scalar penalty function, η is the parameter indexing the penalty family and λ is the tuning constant. We list some commonly used penalty functions below.

(a) Power family (Frank and Friedman, 1993):

$$P_\eta(|w|, \lambda) = \lambda |w|^\eta, \qquad \eta \in (0, 2].$$

Two important special cases of this family are the lasso penalty when $\eta = 1$ (Tibshirani, 1996; Chen *et al.*, 2001) and the ridge penalty when $\eta = 2$ (Hoerl and Kennard, 1970).

(b) The elastic net (Zou and Hastie, 2005):

$$P_\eta(|w|, \lambda) = \lambda\{(\eta - 1)w^2/2 + (2 - \eta)|w|\}, \qquad \eta \in [1, 2].$$

Varying η from 1 to 2 bridges the lasso to the ridge penalty.

(c) The log-penalty (Candès *et al.*, 2008; Armagan *et al.*, 2013)

$$P_\eta(|w|, \lambda) = \lambda \ln(\eta + |w|), \qquad \eta > 0.$$

(d) SCAD (Fan and Li, 2001), in which the penalty is defined via its partial derivative:

$$\frac{\partial}{\partial |w|} P_\eta(|w|, \lambda) = \lambda \left\{ \mathbf{1}_{\{|w| \leqslant \lambda\}} + \frac{(\eta\lambda - |w|)_+}{(\eta - 1)\lambda} \mathbf{1}_{\{|w| > \lambda\}} \right\}, \qquad \eta > 2.$$

Integration shows that SCAD is a natural quadratic spline with knots at λ and $\eta\lambda$

$$P_\eta(|w|, \lambda) = \begin{cases} \lambda |w| & |w| < \lambda, \\ \lambda^2 + \dfrac{\eta\lambda(|w| - \lambda)}{\eta - 1} - \dfrac{w^2 - \lambda^2}{2(\eta - 1)} & |w| \in [\lambda, \eta\lambda], \\ \dfrac{\lambda^2(\eta + 1)}{2} & |w| > \eta\lambda. \end{cases}$$

For small signals $|w| < \lambda$, it acts as a lasso; for large signals $|w| > \eta\lambda$, the penalty flattens and leads to the unbiasedness of the regularized estimate.

(e) The MC+-penalty (Zhang, 2010), which is similar to SCAD and is defined by the partial derivative

$$\frac{\partial}{\partial |w|} P_\eta(|w|, \lambda) = \lambda \left(1 - \frac{|w|}{\lambda\eta}\right)_+.$$

Integration shows that the penalty function

$$P_\eta(|w|,\lambda)=\left(\lambda|w|-\frac{w^2}{2\eta}\right)\mathbf{1}_{\{|w|<\lambda\eta\}}+\frac{\lambda^2\eta}{2}\mathbf{1}_{\{|w|\geqslant\lambda\eta\}},\qquad \eta>0,$$

is quadratic on $[0,\lambda\eta]$ and flattens beyond $\lambda\eta$. Varying η from 0 to ∞ bridges hard thresholding (l_0-regression) to lasso (l_1-) shrinkage.

We also comment that, besides these sparsity penalties, other forms of regularization can be useful, depending on the scientific question of interest. For instance, the choice $f(\mathbf{w})=\lambda\Sigma_{j=1}^{q-1}|w_j-w_{j+1}|=\lambda(w_1-w_r)$ produces the regularization for the 'spiked' matrix model, i.e. matrices with clustered eigenvalues or singular values (Johnstone, 2001).

Convexity eases the study of convergence properties in many optimization problems. We first state the necessary and sufficient condition for the convexity of the regularizer J. Its proof follows from the theory of the spectral function (Borwein and Lewis, 2006).

Lemma 1. The functional $J(\mathbf{B})=f\circ\sigma(\mathbf{B})$ is convex and lower semicontinuous if and only if f is convex and lower semicontinuous. Furthermore, for a convex f, the subdifferential of J at $\mathbf{B}$, which admits singular value decomposition $\mathbf{U}\,\mathrm{diag}(\mathbf{b})\mathbf{V}^{\mathrm{T}}$, is

$$\partial J(\mathbf{B})=\partial(f\circ\sigma)(\mathbf{B})=\mathbf{U}\,\mathrm{diag}\{\partial f(\mathbf{b})\}\mathbf{V}^{\mathrm{T}}.$$

Here 'diag' denotes a diagonal matrix. Lemma 1 immediately leads to the optimality condition when both loss and regularizer are convex.

Theorem 1. When both the loss l and f are convex, all local minima of the regularized programme (2) are global minima and are unique if l is strictly convex. A matrix $\mathbf{B}=\mathbf{U}\,\mathrm{diag}(\mathbf{b})\mathbf{V}^{\mathrm{T}}$ is a global minimum if and only if

$$\mathbf{0}_{p_1\times p_2}\in\nabla l(\mathbf{B})+\mathbf{U}\,\mathrm{diag}\{\partial f(\mathbf{b})\}\mathbf{V}^{\mathrm{T}}.$$

When either the loss l or f is non-convex, the regularized objective function (2) may be non-convex and lacks an easy-to-check optimality condition.

3. Estimation algorithm

We utilize the powerful Nesterov optimal gradient method (Nesterov, 1983, 2004; Beck and Teboulle, 2009a) for minimizing the non-smooth and possibly non-convex objective function (2). We first state a matrix thresholding formula for spectral regularization, which forms the building blocks of the Nesterov algorithm.

Proposition 1. For a given matrix $\mathbf{A}$ with singular value decomposition $\mathbf{A}=\mathbf{U}\,\mathrm{diag}(\mathbf{a})\mathbf{V}^{\mathrm{T}}$, the optimal solution to

$$\min_{\mathbf{B}}\tfrac{1}{2}\|\mathbf{B}-\mathbf{A}\|_{\mathrm{F}}^2+f\circ\sigma(\mathbf{B})$$

shares the same singular vectors as $\mathbf{A}$ and its ordered singular values are the solution to

$$\min_{\mathbf{b}}\tfrac{1}{2}\|\mathbf{b}-\mathbf{a}\|_2^2+f(\mathbf{b}).$$

Here $\mathbf{A}$ and $\mathbf{B}$ denote two generic matrices, with $\mathbf{a}$ and $\mathbf{b}$ as their singular values accordingly. An immediate consequence of proposition 1 is the following well-known singular value thresholding formula for nuclear norm regularization (Cai *et al.*, 2010).

Corollary 1. For a given matrix $\mathbf{A}$ with singular value decomposition $\mathbf{A} = \mathbf{U}\,\mathrm{diag}(\mathbf{a})\mathbf{V}^{\mathrm{T}}$, the optimal solution to

$$\min_{\mathbf{B}} \tfrac{1}{2}\|\mathbf{B} - \mathbf{A}\|_{\mathrm{F}}^2 + \lambda\|\mathbf{B}\|_*$$

shares the same singular vectors as $\mathbf{A}$ and its singular values are $b_i = (a_i - \lambda)_+$.

Given this matrix thresholding formula, we are ready to present the Nesterov algorithm for minimizing problem (2). The Nesterov method has attracted increasing attention in recent years owing to its efficiency in solving regularization problems (Beck and Teboulle, 2009a). It resembles the classical gradient descent algorithm in that only the first-order gradients of the objective function are utilized to produce the next algorithmic iterate from the current search point and as such is simple to implement. It differs from the gradient descent algorithm by extrapolating the previous two algorithmic iterates to generate the next search point. This extrapolation step incurs trivial computational cost but improves the convergence rate dramatically. It has been shown to be optimal among a wide class of convex smooth optimization problems (Nemirovski, 1994; Nesterov, 2004).

We summarize our Nesterov method for solving spectral regularization problem (2) in algorithm 1 (Table 1). Each iteration consists of three stages:

(a) predict a search point $\mathbf{S}$ by a linear extrapolation from the previous two iterates (step 3 of algorithm 1),
(b) perform gradient descent from the search point $\mathbf{S}$ possibly with an Armijo-type line search (steps 4–10) and
(c) force the descent property of the next iterate (steps 11–15).

In step (a), $\alpha^{(t)}$ is a scalar sequence that plays a critical role in the extrapolation. We update this sequence as in the original Nesterov method (step 16 of algorithm 1), whereas other sequences,

Table 1. Algorithm 1: Nesterov method for spectral regularized matrix regression (2)

Step	*Description*	*Comment*
1	Initialize $\mathbf{B}^{(0)} = \mathbf{B}^{(1)}$, $\alpha^{(0)} = 0$, $\alpha^{(1)} = 1$, $\delta > 0$	
2	*Repeat*	
3	$\mathbf{S}^{(t)} \leftarrow \mathbf{B}^{(t)} + \dfrac{\alpha^{(t-1)} - 1}{\alpha^{(t)}}(\mathbf{B}^{(t)} - \mathbf{B}^{(t-1)})$	Extrapolation
4	*Repeat*	Line search
5	$\mathbf{A}_{\mathrm{temp}} \leftarrow \mathbf{S}^{(t)} - \delta\nabla l(\mathbf{S}^{(t)})$	
6	Compute singular value decomposition $\mathbf{A}_{\mathrm{temp}} = \mathbf{U}\,\mathrm{diag}(\mathbf{a})\mathbf{V}^{\mathrm{T}}$	
7	$\mathbf{b} \leftarrow \arg\min_{\mathbf{x}} (2\delta)^{-1}\|\mathbf{x} - \mathbf{a}\|_2^2 + f(\mathbf{x})$	
8	$\mathbf{B}_{\mathrm{temp}} \leftarrow \mathbf{U}\,\mathrm{diag}(\mathbf{b})\mathbf{V}^{\mathrm{T}}$	
9	$\delta \leftarrow \delta/2$	
10	*Until* $h(\mathbf{B}_{\mathrm{temp}}) \leqslant g(\mathbf{B}_{\mathrm{temp}} \mid \mathbf{S}^{(t)}, \delta)$	Force descent
11	*If* $h(\mathbf{B}_{\mathrm{temp}}) \leqslant (\mathbf{B}^t)$ *then*	
12	$\mathbf{B}^{(t+1)} \leftarrow \mathbf{B}_{\mathrm{temp}}$	
13	*Otherwise*	
14	$\mathbf{B}^{(t+1)} \leftarrow \mathbf{B}^{(t)}$	
15	*End*	
16	$\alpha^{(t+1)} \leftarrow [1 + \sqrt{\{1 + (2\alpha^{(t)})^2\}}]/2$	Update α
17	*Until* objective value converges	

for instance $\alpha^{(t)} = (t-1)/(t+2)$, can also be used. In stage (b), the gradient descent is based on the first-order approximation to the loss function at the current search point $\mathbf{S}^{(t)}$,

$$g(\mathbf{B}|\mathbf{S}^{(t)}, \delta) = l(\mathbf{S}^{(t)}) + \langle \nabla l(\mathbf{S}^{(t)}), \mathbf{B} - \mathbf{S}^{(t)} \rangle + \frac{1}{2\delta} \|\mathbf{B} - \mathbf{S}^{(t)}\|_{\mathrm{F}}^2 + J(\mathbf{B})$$
$$= \frac{1}{2\delta} \|\mathbf{B} - \{\mathbf{S}^{(t)} - \delta \nabla l(\mathbf{S}^{(t)})\}\|_{\mathrm{F}}^2 + J(\mathbf{B}) + c^{(t)},$$

where the constant δ is determined during the line search and the constant $c^{(t)}$ collects terms that are irrelevant to the optimization. The 'ridge' term $(2\delta)^{-1}\|\mathbf{B} - \mathbf{S}^{(t)}\|_{\mathrm{F}}^2$ acts as a trust region and shrinks the next iterate towards $\mathbf{S}^{(t)}$. If the loss function $l \in \mathcal{C}_{1,1}$, which denotes the class of functions that are convex and continuously differentiable and the gradient satisfies $\|\nabla l(\mathbf{u}) - \nabla l(\mathbf{v})\| \leqslant \mathcal{L}(l)\|\mathbf{u} - \mathbf{v}\|$ with a known gradient Lipschitz constant $\mathcal{L}(l)$ for all $\mathbf{u}$ and $\mathbf{v}$, then δ is fixed at $\mathcal{L}(l)^{-1}$. In practice, the gradient Lipschitz constant is often unknown. Then δ is updated dynamically to capture the unknown $\mathcal{L}(l)$ by using the classical Armijo line search rule (Nocedal and Wright, 2006; Lange, 2004). Solution to the surrogate function $g(\mathbf{B}|\mathbf{S}^{(t)}, \delta)$ is given by proposition 1. Singular value decomposition is performed on the intermediate matrix $\mathbf{A}_{\text{temp}} = \mathbf{S}^{(t)} - \delta \nabla l(\mathbf{S}^{(t)})$. The next iterate $\mathbf{B}^{(t+1)}$ shares the same singular vectors as $\mathbf{A}$ and its singular values $\mathbf{b}^{(t+1)}$ are determined by minimizing $(2\delta)^{-1}\|\mathbf{b} - \mathbf{a}\|_2^2 + f(\mathbf{b})$, where $\mathbf{a} = \sigma(\mathbf{A}_{\text{temp}})$. For a nuclear norm regularization $f(\mathbf{w}) = \lambda \Sigma_j |w_i|$, the solution is given by soft thresholding the singular values $b_i^{(t+1)} = (a_i - \lambda\delta)_+$. In this special case, only the top singular values or vectors need to be retrieved. The Lanczos method (Golub and Van Loan, 1996) is extremely efficient for this purpose. For a linear regularization function $f(\mathbf{w}) = \lambda\|\mathbf{D}\mathbf{w}\|_1$ where $\mathbf{D}$ has full column rank, reparameterization $\mathbf{c} = \mathbf{D}\mathbf{b}$ turns the problem into $\min_{\mathbf{c}}(1/2\delta)\|(\mathbf{D}^{\mathrm{T}}\mathbf{D})^{-1}\mathbf{D}^{\mathrm{T}}\mathbf{c} - \mathbf{a}\|_2^2 + \lambda\|\mathbf{c}\|_1$, which is a standard lasso problem with many efficient solvers available. When $\mathbf{D}$ does not have a full column rank, we append extra rows such that the expanded matrix, which is denoted by $\tilde{\mathbf{D}}$, has full column rank and then solve the above lasso problem with $\tilde{\mathbf{D}}$ and only part of $\mathbf{c}$ penalized.

For minimization of a smooth convex function l in $\mathcal{C}_{1,1}$, it is well known that the Nesterov method is optimal with the convergence rate of order $O(t^{-2})$, where t indicates the iterate number. In contrast, the gradient descent has a slower convergence rate of $O(t^{-1})$. Our spectral regularization problem (2) is non-smooth, but a similar non-asymptotic convergence result can be established, which is summarized in theorem 2. Its proof has been omitted for brevity, and readers are referred to Beck and Teboulle (2009a).

Theorem 2. Suppose that l is continuously differentiable with a gradient Lipschitz constant $\mathcal{L}(l)$. Let $\mathbf{B}^{(t)}$ be the iterates generated by the Nesterov method described in algorithm 1. Then the objective value $h(\mathbf{B}^{(t)})$ monotonically converges. Furthermore, if J is convex, then

$$h(\mathbf{B}^{(t)}) - h(\mathbf{B}^*) \leqslant \frac{4\mathcal{L}(l)\|\mathbf{B}^{(0)} - \mathbf{B}^*\|_{\mathrm{F}}^2}{(t+1)^2} \tag{3}$$

for all $t \geqslant 0$ and any minimum point $\mathbf{B}^*$.

We make two remarks here. The first regards the monotonicity of the objective function during iterations. Because of the extrapolation step, the objective values of algorithmic iterates $h(\mathbf{B}^{(t)})$ are not guaranteed to be monotonically decreasing. When the loss $l \in \mathcal{C}_{1,1}$ and the regularizer J is convex, convergence of the objective values is guaranteed with the explicit convergence rate (3). Because of potential use of a non-convex J, we enforce monotonicity of algorithmic iterates (rows 11–15 in algorithm 1), which is essential for the convergence of at least the objective values. After each gradient descent step, if the new iterate fails to decrease the objective value,

then the current iterate is the same as the previous iterate. In other words, the next gradient descent is initiated from the previous iterate. Fortunately, the fast convergence rate (3) still holds under the assumptions $l \in C_{1,1}$ and J is convex. See Beck and Teboulle (2009b) for the argument.

The second remark is about a crude estimate of the Lipschitz constant $\mathcal{L}(l)$ for the GLM loss l. Each step halving in the line search part of algorithm 1 involves an expensive singular value decomposition. Therefore even a rough initial estimate of $\mathcal{L}$ potentially cuts the computational cost significantly. Recall that a twice differentiable function f is continuously differentiable with a Lipschitz constant L if and only if $\mathbf{v}^{\mathrm{T}} \mathrm{d}^2 f(\mathbf{u})\mathbf{v} \leqslant L\|\mathbf{v}\|_2^2$ for all $\mathbf{v}$. The Fisher information matrix of a GLM model with systematic part (1) is

$$\mathbf{I}(\mathbf{B}) = E\{\mathrm{d}^2 l(\mathbf{B})\} = \sum_{i=1}^{n} \omega_i (\mathrm{vec}(\mathbf{x}_i))(\mathrm{vec}(\mathbf{x}_i))^{\mathrm{T}},$$

where $\omega_i = \{\mu_i'(\eta_i)/\sigma_i\}^2$, η_i is the systematic part, μ_i is the mean and σ_i^2 is the variance corresponding to the ith observation. Then in light of the Cauchy–Schwartz inequality

$$\mathbf{v}^{\mathrm{T}} \mathbf{I}(\mathbf{B})\mathbf{v} = \sum_i \omega_i (\mathbf{v}^{\mathrm{T}} \mathrm{vec}(\mathbf{x}_i))(\mathbf{v}^{\mathrm{T}} \mathrm{vec}(\mathbf{x}_i))^{\mathrm{T}} \leqslant \sum_i \omega_i \|\mathrm{vec}(\mathbf{x}_i)\|_2^2 \|\mathbf{v}\|_2^2 = \|\mathbf{v}\|_2^2 \left(\sum_i \omega_i \|\mathbf{x}_i\|_{\mathrm{F}}^2 \right),$$

and thus an initial estimate of L is given by $L \approx \Sigma_i \mu_i'(\eta_i)^2/\sigma_i^2 \|\mathbf{X}_i\|_{\mathrm{F}}^2$. Note that this L provides an upper bound to the smallest Lipschitz constant. Setting initial δ one or two orders of magnitude larger than L^{-1} works well in practice.

4. Degrees of freedom

In this section, we address the problem of choosing the tuning parameter λ that yields the best model along the regularization path according to certain criteria. Cross-validation is commonly used for parameter tuning in practice. However, for large data, it may incur considerable computation burden. There are computationally attractive alternatives, such as the Akaike information criterion AIC (Akaike, 1974) and BIC (Schwarz, 1978), which often yield performance that is comparable with cross-validation in practice.

Consider a normal model under the GLM (1). For simplicity, we again drop the covariate vector $\mathbf{Z}$:

$$Y = \langle \mathbf{X}, \mathbf{B} \rangle + \varepsilon \tag{4}$$

where ε is a normal error with mean 0 and variance v^2. Let y_i denote the ith observation of Y and $\hat{y}_i(\lambda)$ denote the estimated response under a given tuning parameter λ from the minimization of problem (2). Then, for this normal model, AIC and BIC are defined by

$$\mathrm{AIC}(\lambda) = \frac{\sum_i \{y_i - \hat{y}_i(\lambda)\}^2}{v^2} + 2\,\mathrm{df}(\lambda),$$

$$\mathrm{BIC}(\lambda) = \frac{\sum_i \{y_i - \hat{y}_i(\lambda)\}^2}{v^2} + \ln(n)\,\mathrm{df}(\lambda).$$

In applications, the variance v^2 is often unknown but can be estimated from the fitted value by least squares estimation. An essential element in the above model selection criteria is the effective degrees of freedom $\mathrm{df}(\lambda)$ of the selected model. Using Stein's theory of unbiased risk estimation (Stein, 1981), Efron (2004) showed that, under a differentiability condition on $\hat{\mathbf{y}}(\lambda)$,

$$\mathrm{df}(\lambda) = E\left[\mathrm{tr}\left\{\frac{\partial \hat{\mathbf{y}}(\lambda)}{\partial \mathbf{y}}\right\}\right] = \frac{1}{\tau^2}\sum_{i=1}^{n}\mathrm{cov}\{\hat{y}_i(\lambda), y_i\}$$

with expectation taken with respect to Y, $\mathbf{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$ and $\hat{\mathbf{y}} = (\hat{y}_1(\lambda), \ldots, \hat{y}_n(\lambda))^{\mathrm{T}}$. This formulation has been productively used to derive the degrees-of-freedom estimate in least angle regression (Efron *et al.*, 2004), the lasso (Zou *et al.*, 2007), group-penalized regression (Yuan and Lin, 2006) and sign coherent group-penalized regression (Chiquet *et al.*, 2012).

We derive a degrees-of-freedom estimate for the nuclear-norm-regularized estimate under a normal model. For an orthonormal design, i.e. $\boldsymbol{\Xi}^{\mathrm{T}}\boldsymbol{\Xi} = \mathbf{I}_{p_1 p_2}$ with the matrix $\boldsymbol{\Xi}$ having rows $(\mathrm{vec}\,(\mathbf{x}_i))^{\mathrm{T}}$, the estimate derived is unbiased for the true degrees of freedom. In practice, it yields results that are comparable with cross-validation even for non-orthogonal designs.

Theorem 3. Assume that the data are generated from model (4) with $\mathrm{vec}(\mathbf{x}_i)$ orthonormal. Consider the nuclear-norm-regularized estimate

$$\hat{\mathbf{B}}_\lambda = \arg\min_{\mathbf{B}} \frac{1}{2}\sum_i (y_i - \langle \mathbf{X}_i, \mathbf{B}\rangle)^2 + \lambda\|\mathbf{B}\|_*$$

with singular values $\sigma(\hat{\mathbf{B}}_\lambda) = (b_1(\lambda), \ldots, b_q(\lambda))$ where $q = \min\{p_1, p_2\}$. Let $\hat{\mathbf{B}}_{\mathrm{LS}}$ be the usual least squares estimate and assume that it has distinct positive singular values $\sigma_1 > \ldots > \sigma_q > 0$. With the convention $\sigma_i = 0$ for $i > q$, the following expression is an unbiased estimate of the degrees of freedom of the regularized fit:

$$\widehat{\mathrm{df}}(\lambda) = \sum_{i=1}^{q} \mathbf{1}_{\{b_i(\lambda)>0\}}\left\{1 + \sum_{1\leqslant j\leqslant p_1,\, j\neq i} \frac{\sigma_i(\sigma_i - \lambda)}{\sigma_i^2 - \sigma_j^2} + \sum_{1\leqslant j\leqslant p_2,\, j\neq i} \frac{\sigma_i(\sigma_i - \lambda)}{\sigma_i^2 - \sigma_j^2}\right\}.$$

This formula for the degrees of freedom is interesting in several respects. First it does not involve any information on the singular vectors of the least squares estimate, but requires only singular values. Second, $\widehat{\mathrm{df}}(\lambda)$ is continuous in λ, in contrast with the piecewise constant degrees-of-freedom estimate for the classical lasso (Zou *et al.*, 2007). Third, at any $\lambda > 0$, the effective degrees of freedom are always dominated by the naive count of parameters $r(\lambda)(p_1 + p_2) - r^2(\lambda)$ in the regularized estimate $\hat{\mathbf{B}}_\lambda$ of rank $r(\lambda)$, as manifest from the straightforward inequalities

$$\begin{aligned}
\widehat{\mathrm{df}}(\lambda) &\leqslant \sum_{i=1}^{q} \mathbf{1}_{\{b_i(\lambda)>0\}}\left(1 + 2\sum_{r(\lambda)<j\leqslant q} \frac{\sigma_i}{\sigma_i + \sigma_j} + \sum_{q<j\leqslant \max\{p_1, p_2\}} \frac{\sigma_i}{\sigma_i + 0}\right) \\
&\leqslant \sum_{i=1}^{q} \mathbf{1}_{\{b_i(\lambda)>0\}}[1 + 2\{q - r(\lambda)\} + \max\{p_1, p_2\} - q] \\
&= \sum_{i=1}^{q} \mathbf{1}_{\{b_i(\lambda)>0\}}\{p_1 + p_2 + 1 - 2r(\lambda)\} \leqslant r(\lambda)(p_1 + p_2) - r^2(\lambda).
\end{aligned}$$

This reflects the overwhelming shrinkage effect of the nuclear norm regularization over model searching. At $\lambda = 0$, $\hat{\mathbf{B}}_0 = \hat{\mathbf{B}}_{\mathrm{LS}}$ almost surely has a full rank and the number of degrees of freedom is $q + q(q-1) + q(p_1 + p_2 - 2q) = p_1 p_2$, which is exactly the number of parameters without any regularization and reflects the effect of no shrinkage. Fig. 2 plots the effective degrees of freedom $\widehat{\mathrm{df}}(\lambda)$ and the naive count $r(\lambda)(p_1 + p_2) - r^2(\lambda)$ for a matrix parameter of size 64×64.

Finally we note that the degrees-of-freedom formula in theorem 3 is limiting as it requires the least squares estimate $\hat{\mathbf{B}}_{\mathrm{LS}}$, which is not unique when $n < p_1 p_2$. In this case we may use a ridge estimate $\hat{\mathbf{B}}_{\mathrm{ridge}(\tau)}$ where

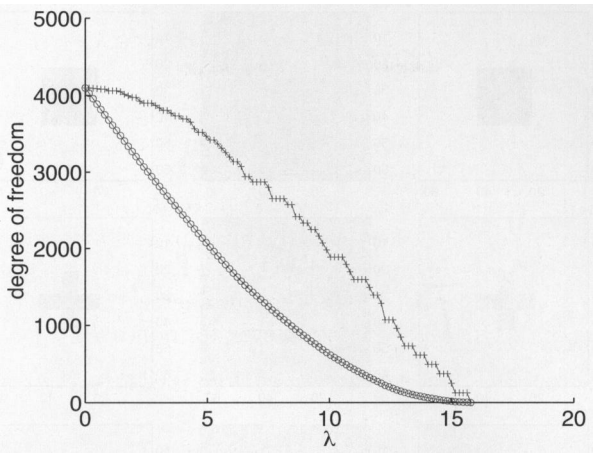

Fig. 2. Degrees-of-freedom estimate $\widehat{\mathrm{df}}(\lambda)$ *versus* the number of parameters in the estimated model $\hat{\mathbf{B}}_\lambda$ with a $\hat{\mathbf{B}}_{\mathrm{LS}} \in \mathbb{R}^{64\times 64}$: ○, $\mathrm{df}(\lambda)$; +, $r(\lambda)^*(p_1+p_2)-r^2(\lambda)$

$$\mathrm{vec}(\hat{\mathbf{B}}_{\mathrm{ridge}(\tau)}) = (\boldsymbol{\Xi}^{\mathrm{T}}\boldsymbol{\Xi} + \tau \mathbf{I}_{p_1 p_2})^{-1}\boldsymbol{\Xi}^{\mathrm{T}}\mathbf{y},$$

which always exists and is unique. Assume that $\hat{\mathbf{B}}_{\mathrm{ridge}(\tau)}$ admits a singular value decomposition $\hat{\mathbf{B}}_{\mathrm{ridge}(\tau)} = \mathbf{U}\,\mathrm{diag}(\boldsymbol{\sigma})\mathbf{V}^{\mathrm{T}}$. The degrees-of-freedom formula

$$\widehat{\mathrm{df}}(\tau) = \sum_{i=1}^{q} \mathbf{1}_{\{b_i(\tau)>0\}}\left[1 + \frac{1}{1+\tau}\sum_{1\leqslant j\leqslant p_1,\, j\neq i}\frac{\sigma_i\{(1+\tau)\sigma_i-\lambda\}}{\sigma_i^2-\sigma_j^2} + \frac{1}{1+\tau}\sum_{1\leqslant j\leqslant p_2,\, j\neq i}\frac{\sigma_i\{(1+\tau)\sigma_i-\lambda\}}{\sigma_i^2-\sigma_j^2}\right]$$

generalizes theorem 3 and is unbiased for the true degrees of freedom under the same assumptions as theorem 3.

5. Numerical examples

We have conducted extensive numerical studies with two aims: first, we investigate the empirical performance of the proposed spectral regularized regression with matrix covariates and, second, we compare with the corresponding classical regularization solutions. Four methods are under comparison: a matrix regression with nuclear norm regularization (since it takes the form $f(\mathbf{w}) = \lambda\Sigma_{j=1}^{q}|w_j|$ in the regularization problem (2), we call this solution the matrix lasso), a usual vector regression after vectorizing the matrix covariate with a lasso penalty (the lasso), a matrix regression with power spectral regularization (matrix power) and the corresponding vector regression with a power penalty (power). For the power penalty, we fix the coefficient $\eta = 0.5$. Our goal here is not how best to tune η. Instead, we examine this penalty since it yields a nearly unbiased estimate. Other non-convex penalties such as SCAD yield very similar results, which are not reported here for brevity. We summarize our findings in three examples. First, we elaborate on the illustrative example by examining some different geometric and natural shapes. Second, we generate some synthetic data and compare different regularization solutions under varying ranks and sparsity levels. Last, we revisit the motivating electroencephalography data that were mentioned in Section 1.

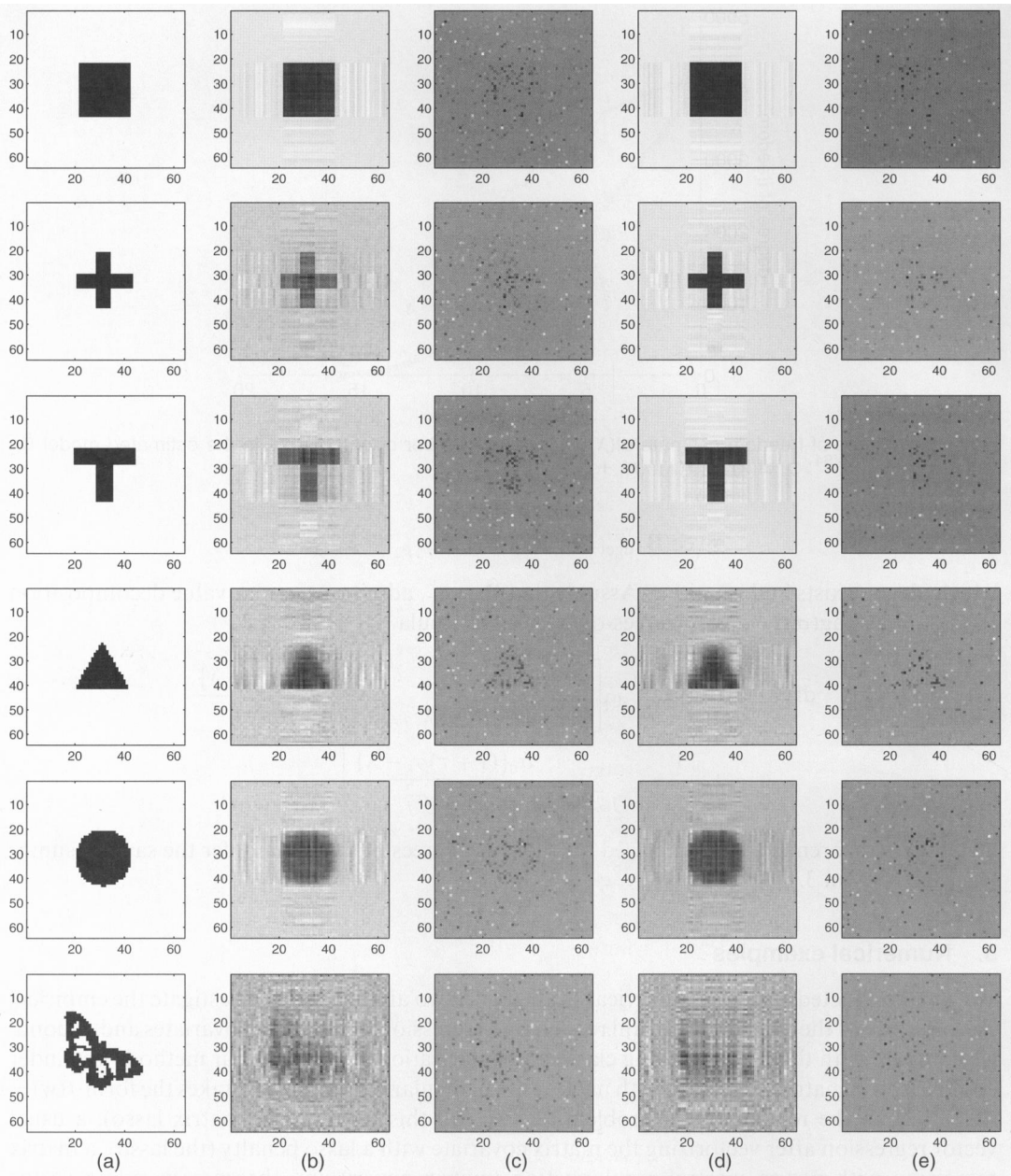

Fig. 3. Comparison of the matrix and vector version of regularized estimators: (a) true signal; (b) matrix lasso estimate; (c) lasso estimate; (d) matrix power estimate; (e) power estimate

5.1. Example 1: two-dimensional shapes

We first elaborate on the illustrative example in Section 1, by employing the same model set-up, but examining a variety of signal shapes. We present in Fig. 3 the true signal followed by the estimates from the four aforementioned regularization methods, where the regularization parameter λ is tuned by BIC. It is seen that the matrix versions of regularized estimators clearly

outperform their vector version counterparts, for both lasso and power penalties. Comparing the matrix lasso with matrix power, the two yield comparable results, whereas the former is better for the high rank signals, and the latter is better for the low rank signals. This observation is further verified in the next simulation example.

5.2. Example 2: synthetic data

We consider a class of synthetic signals to compare various regularization methods under different ranks and sparsity levels. Specifically, we generate the matrix covariates $\mathbf{X}$ of size 64×64 and the five-dimensional vector covariates $\mathbf{Z}$, both of which consist of independent standard normal entries. We set the sample size at $n = 500$, whereas the number of parameters is $64 \times 64 + 5 = 4101$. We set $\gamma = (1, \ldots, 1)^{\mathrm{T}}$ and generate the true array signal as $\mathbf{B} = \mathbf{B}_1 \mathbf{B}_2^{\mathrm{T}}$, where $\mathbf{B}_d \in \mathbb{R}^{p_d \times R}$, $d = 1, 2$. R controls the rank of the signal generated. Moreover, each entry of $\mathbf{B}$ is 0 or 1, and the percentage of non-zero entries is controlled by a sparsity level constant s, i.e. each entry of $\mathbf{B}_d$ is a Bernoulli distribution with probability of 1 equal to $\sqrt{\{1 - (1-s)^{1/R}\}}$. We vary the rank $R = 1, 5, 10, 20$, and the level of (non-)sparsity $s = 0.01, 0.05, 0.1, 0.2, 0.5$. (So $s = 0.05$ means that about 5% of entries of $\mathbf{B}$ are 1s and the rest are 0s.) We generate both a normal and a binomial response Y with the systematic part as in equation (1), with identity link for the normal model, and logistic link for the binomial model.

We evaluate the performance of each method from two aspects: parameter estimation and prediction. For the former, we employ BIC for tuning and the root-mean-squared error as the evaluation criterion. For the latter, we use independent validation data to tune the parameter and testing data to evaluate the prediction error measured by the root-mean-square error of the response for the normal case, and the misclassification error rate for the binomial model. We report the prediction results in Tables 2 and 3. The estimation results show a similar qualitative pattern and thus are not reported for brevity.

We make the following observations. First, the proposed matrix version of estimators almost always outperform the corresponding vector version in terms of both parameter estimation and prediction. Second, the regular version of regularized estimators perform better when the signal is more sparse (a smaller s), whereas it is relatively insensitive to the rank R. In contrast, the matrix version estimators proposed perform better when the rank is smaller and are insensitive to the coefficient sparsity. These patterns agree with our expectations since the former penalizes directly on the coefficient count, whereas the latter on the rank. Third, comparing the lasso and power penalty, the two yield similar results, whereas the lasso usually performs better when the rank is high, and the power is better at low ranks. Such patterns agree with what we observed in example 1 and provide some useful guidelines when choosing a penalty given the data. Finally, we comment on the run time. The run time of the matrix solution is longer than that of the vector solution. But overall it is reasonably fast. For instance, for $R = 1$, $s = 1\%$ and the normal response, the average run time for obtaining the solution path with 40 grid points of the matrix lasso, lasso, the matrix power and power are 11.27, 2.90, 11.91 and 4.29 s respectively. For $R = 20$, $s = 50\%$ and the binary response, the average run times of the four methods are 4.30, 0.40, 13.80 and 0.41 s respectively. We feel that this run time increase is fully justified by the pronounced better performance. The run time of the lasso is similar to that of the power method for the normal response and faster for the binary response, and it only varies slightly for different ranks and levels of sparsity.

5.3. Example 3: electroencephalography data analysis

We next analyse the motivating electroencephalography data. The data arise from a study to

Table 2. Prediction of a normal model†

Sparsity s (%)	*Method*	*Results for the following ranks:*			
		R=1	*R=5*	*R=10*	*R=20*
1	Matrix lasso	1.71 (0.20)	4.60 (1.09)	5.53 (1.01)	5.87 (0.83)
	Lasso	1.61 (0.98)	1.54 (0.49)	1.62 (0.76)	1.53 (0.23)
	Matrix power	*1.18* (0.04)	4.00 (1.16)	5.53 (1.05)	5.99 (0.87)
	Power	1.24 (1.24)	*1.13* (0.74)	*1.17* (0.96)	*1.07* (0.04)
5	Matrix lasso	2.28 (0.31)	11.03 (1.42)	*12.84* (1.58)	*13.75* (1.35)
	Lasso	13.49 (2.92)	13.63 (2.12)	13.69 (2.05)	13.86 (1.61)
	Matrix power	*1.18* (0.04)	*10.25* (1.48)	13.01 (1.59)	14.12 (1.41)
	Power	14.71 (3.38)	14.88 (2.11)	14.93 (2.02)	15.11 (1.60)
10	Matrix lasso	2.54 (0.42)	16.10 (1.94)	*18.59* (1.57)	*19.81* (1.51)
	Lasso	19.60 (2.66)	21.26 (2.49)	21.17 (2.05)	21.19 (1.74)
	Matrix power	*1.18* (0.05)	*15.39* (2.16)	18.84 (1.59)	20.24 (1.60)
	Power	21.14 (2.88)	22.80 (2.78)	22.71 (2.32)	22.55 (1.78)
20	Matrix lasso	3.19 (0.66)	22.72 (2.16)	*26.78* (2.26)	*28.87* (2.26)
	Lasso	28.69 (3.02)	31.86 (3.22)	32.59 (3.18)	32.45 (2.77)
	Matrix power	*1.17* (0.04)	*21.40* (2.31)	27.10 (2.17)	29.27 (2.19)
	Power	30.88 (3.28)	34.35 (3.50)	35.01 (3.60)	35.00 (3.00)
50	Matrix lasso	4.57 (1.03)	35.65 (2.15)	45.13 (2.99)	50.92 (3.05)
	Lasso	45.82 (2.75)	62.83 (4.90)	66.49 (5.52)	67.48 (4.44)
	Matrix power	*1.18* (0.05)	*29.64* (2.08)	*42.10* (2.38)	*48.89* (2.95)
	Power	49.80 (3.09)	68.07 (5.72)	71.63 (6.06)	72.73 (4.99)

†Reported are the mean and standard deviation (in parentheses) of the root-mean-squared error for y out of 100 data replications. Values in italics denote the best method in each group.

examine electroencephalography correlates of genetic predisposition to alcoholism (`http://kdd.ics.uci.edu/datasets/eeg/eeg.data.html`). They consist of 77 alcoholic individuals and 44 controls. For each subject, 64 electrodes were placed on the scalp that were sampled at 256 Hz (3.9-ms epoch) for 1 s. The electrode positions were located at standard sites (standard electrode position nomenclature; American Electroencephalographic Association (1991)). In addition each subject performed 120 trials under three types of stimuli: a single stimulus, two matched stimuli and two unmatched stimuli, and the electroencephalogram measurements were collected for each trial. Detailed information about this data collection can be found in Zhang *et al.* (1995). The same data set was also analysed in Li *et al.* (2010) and Hung and Wang (2013). Following their practice, we focus on the average of all trials under the single-stimulus condition for each subject. The resulting covariates $\mathbf{x}_i$ are 256×64 matrices, and the response y_i is a binary variable indicating whether the ith subject is alcoholic ($y_i = 1$) or not ($y_i = 0$).

Given that the true signal is rarely of an exact low rank structure and the experience from simulations, we focus on comparing the matrix lasso and regular lasso in this data analysis. Evaluations are twofold: the plot of matrix coefficient estimates, and the accuracy of classification for 'future' observations. First, we fit the full data, tuned on the basis of fivefold cross-validation, and show the matrix coefficient estimates in Fig. 4. The matrix-regularized estimate suggests an interesting outcome; for instance, the group of channels 20–30 shows similar time varying patterns. In contrast, the vector version estimate yields no useful information. Given the potential regions of interest, additional experiments are warranted for validation. Second, we evaluate the cross-validation-based misclassification error to compare classification accuracy for future data, i.e. we divide the full data into a training and a testing sample by using k-fold cross-validation.

Table 3. Prediction of a binomial model†

Sparsity s (%)	*Method*	*Results for the following ranks:*			
		R = 1	*R = 5*	*R = 10*	*R = 20*
1	Matrix lasso	*0.23* (0.03)	*0.34* (0.03)	*0.36* (0.02)	*0.37* (0.02)
	Lasso	0.41 (0.04)	0.42 (0.04)	0.41 (0.04)	0.42 (0.03)
	Matrix power	0.25 (0.03)	0.35 (0.03)	0.37 (0.03)	0.38 (0.03)
	Power	0.76 (0.08)	0.76 (0.07)	0.78 (0.08)	0.77 (0.07)
5	Matrix lasso	*0.20* (0.02)	*0.36* (0.02)	*0.39* (0.02)	*0.41* (0.03)
	Lasso	0.46 (0.02)	0.46 (0.03)	0.46 (0.03)	0.46 (0.02)
	Matrix power	0.22 (0.03)	0.37 (0.02)	0.40 (0.02)	0.42 (0.03)
	Power	0.85 (0.09)	0.85 (0.09)	0.82 (0.08)	0.85 (0.08)
10	Matrix lasso	*0.19* (0.02)	*0.36* (0.03)	*0.40* (0.02)	*0.40* (0.03)
	Lasso	0.47 (0.02)	0.47 (0.02)	0.47 (0.02)	0.46 (0.02)
	Matrix power	0.22 (0.03)	0.37 (0.03)	0.40 (0.02)	0.41 (0.02)
	Power	0.86 (0.08)	0.87 (0.10)	0.86 (0.09)	0.87 (0.08)
20	Matrix lasso	*0.19* (0.03)	*0.34* (0.03)	*0.38* (0.03)	*0.40* (0.03)
	Lasso	0.47 (0.02)	0.47 (0.03)	0.47 (0.03)	0.47 (0.02)
	Matrix power	0.22 (0.03)	0.36 (0.02)	0.38 (0.03)	0.40 (0.02)
	Power	0.89 (0.09)	0.87 (0.09)	0.86 (0.09)	0.86 (0.09)
50	Matrix lasso	*0.19* (0.02)	*0.28* (0.03)	*0.31* (0.03)	*0.34* (0.03)
	Lasso	0.47 (0.02)	0.47 (0.03)	0.47 (0.02)	0.47 (0.02)
	Matrix power	0.21 (0.02)	0.31 (0.03)	0.34 (0.03)	0.36 (0.02)
	Power	0.87 (0.09)	0.86 (0.09)	0.87 (0.08)	0.87 (0.09)

†Reported are the mean and standard deviation (in parentheses) of the misclassification error for y out of 100 data replications. Values in italics denote the best method in each group.

We fit the training data, with another fivefold cross-validation to tune the shrinkage parameter. We then apply the tuned model to the testing data and report the overall testing misclassification error in Table 4. The results again show the superior performance of the matrix estimate compared with the vector counterpart, whereas the chance estimate yields a 0.465 misclassification rate. We also emphasize that a key advantage of our proposed approach is its ability to suggest a useful model and potentially interesting regions rather than the classification accuracy *per se*. This is different from black-box-type machine-learning-based imaging classifiers.

We briefly compare our analysis with those of Li *et al.* (2010) and Hung and Wang (2013) in terms of classification accuracy, parameter tuning and preprocessing. We make the following remarks. First, Li *et al.* (2010) reported a leave-one-out classification error rate of 0.205 but did not report how the dimension reduction parameters were tuned. Hung and Wang (2013) reported the best leave-one-out classification error rate of 0.139, whereas their tuning parameter was chosen so that the classification accuracy was maximized. We believe, however, that the result could be overoptimistic, whereas a fair evaluation of prediction should have the parameter tuning solely based on the training data. Second, both Li *et al.* (2010) and Hung and Wang (2013) preprocessed the data by using a different version of matrix principal components analysis for dimension reduction. Part of the reason was that their proposed numerical methods cannot directly handle the 256×64 size of the electroencephalography data. In contrast, our proposal is not limited by the matrix size and was directly applied to the original data, given that the Nesterov algorithm is highly efficient and scalable. However, we agree that such preprocessing could potentially improve the overall classification accuracy by removing noisy irrelevant information. However, principal components analysis is known as unsupervised

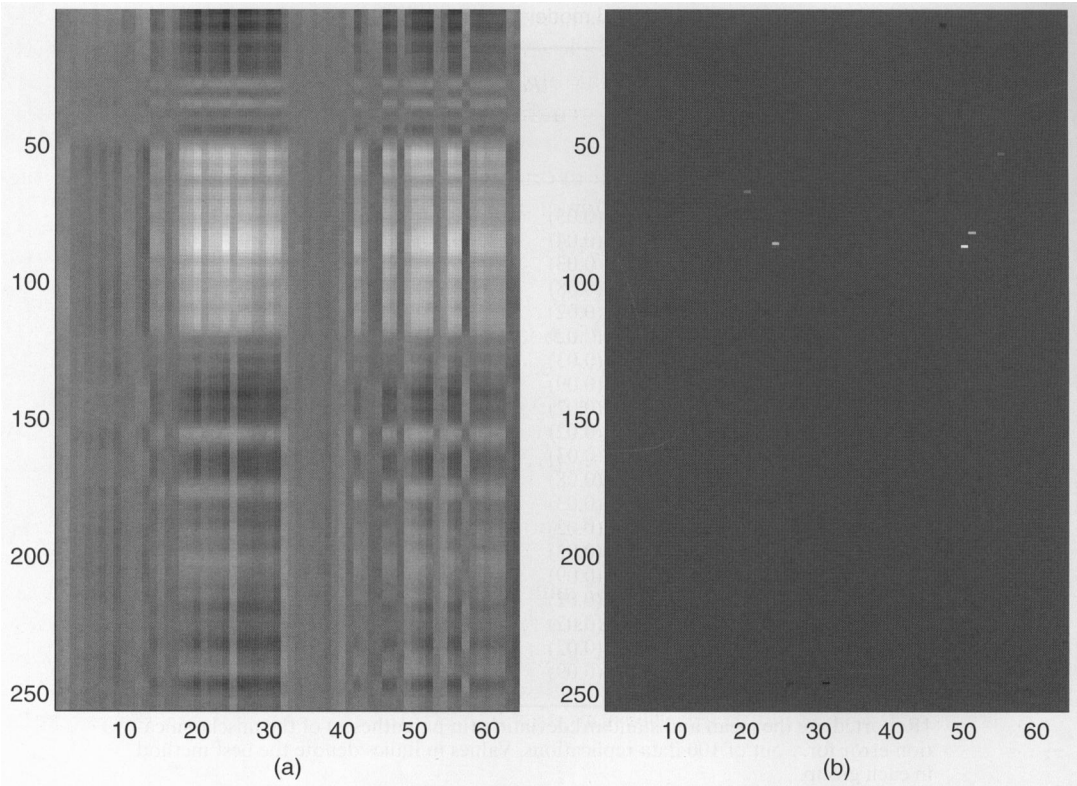

Fig. 4. Electroencephalography data matrix coefficient estimates: (a) matrix lasso estimate; (b) lasso estimate

Table 4. Misclassification error rate for the electroencephalography data

Method	*Leave-one-out rate*	*5-fold rate*	*10-fold rate*	*20-fold rate*
Matrix lasso	*0.230*	*0.214*	*0.222*	*0.181*
Lasso	0.246	0.287	0.271	0.264

dimension reduction, and it introduces another layer of tuning, including choosing the right version of principal components analysis for matrices, and determining the optimal number of principal components. Our goal of this data analysis has primarily been the comparison of the regularized matrix regression estimators with the vector counterparts, so we have chosen not to include any preprocessing.

6. Discussion

Motivated by modern scientific data arising in areas such as brain imaging, we study in this paper the problem of regressions with matrix covariates. Regularization is bound to play a crucial role in such regressions owing to the ultrahigh dimensionality and complex structure of the matrix data. We have proposed a class of regularized matrix regression models by penalizing

the spectrum of the matrix parameters. It is based on the observation that the matrix signals are often of, or can be well approximated by, a low rank structure. Consequently, the new method focuses on the sparsity in terms of the rank of the matrix parameters rather than the number of non-zero entries and is intrinsically different from the classical lasso and related penalization approaches.

In many applications, sparsity is sought in certain prespecified *basis* systems rather than the original co-ordinates. Specifically, the systematic component takes the form

$$q(\mu) = \gamma^{\mathrm{T}}\mathbf{Z} + \langle \mathbf{B}, \mathbf{S}_1^{\mathrm{T}}\mathbf{X}\mathbf{S}_2 \rangle,$$

where $\mathbf{S}_j \in \mathbb{R}^{p_j \times q_j}$ contains q_j basis vectors for $j = 1, 2$. For two-dimensional images, overcomplete wavelet bases ($q_j > p_j$) are often used in both dimensions. For the electroencephalography data that are channels by time points, a wavelet or Fourier basis can be applied to the time dimension. It is of direct interest to seek a sparse low rank representation of the signal $\mathbf{B}$ in the basis system by solving the regularized regression. In that sense, our proposed regularized matrix regression can be considered as the matrix analogue of the classical basis pursuit problem (Chen *et al.*, 2001).

We have concentrated on problems with matrix covariates throughout this paper. In applications such as anatomical magnetic resonance imaging and functional magnetic resonance imaging, the covariates are in the form of multi-dimensional arrays, i.e. tensors. It is natural to extend the work here to regularized tensor regressions. However, the problem formulation requires an appropriate norm for tensors that is analogous to the nuclear norm for matrices, and the regularization and optimization involved are fundamentally different from the methods for matrices. We are currently pursuing this line of extension and will report the results elsewhere.

Acknowledgements

The authors thank the Joint Editor, the Associate Editor and two referees for their insightful and constructive comments. The work is partially supported by National Institutes of Health grant HG006139 (to HZ) and National Science Foundation grants DMS-1106668 (to LL) and DMS-1310319 (to HZ and LL).

Appendix A

A.1. Proof of lemma 1

To show lemma 1, we first need a version of Fan–von Neuman inequality for singular values. See Marshall *et al.* (2011) for a proof of this inequality.

Lemma 2 (Fan's inequality). Let $\mathbf{A} \in \mathbb{R}^{p \times q}$ and $\mathbf{B} \in \mathbb{R}^{q \times r}$ be two matrices with ordered singular values $\mu_1 \geqslant \ldots \geqslant \mu_q$ and $\sigma_1 \geqslant \ldots \geqslant \sigma_q$. Then

$$\mathrm{tr}(\mathbf{AB}) \leqslant \sum_i \mu_i \sigma_i.$$

The equality holds if and only if $\mathbf{A}$ and $\mathbf{B}$ are simultaneously diagonalizable, i.e. $\mathbf{A} = \mathbf{U}\mathbf{D}_{\mathbf{A}}\mathbf{V}^{\mathrm{T}}$ and $\mathbf{B} = \mathbf{U}\mathbf{D}_{\mathbf{B}}\mathbf{V}^{\mathrm{T}}$ for some orthogonal matrices $\mathbf{U}$ and $\mathbf{V}$ and diagonal matrices $\mathbf{D}_{\mathbf{A}}$ and $\mathbf{D}_{\mathbf{B}}$ with entries ordered from largest to smallest.

Recall that the *Fenchel conjugate* of a function $f(\mathbf{x})$ is defined as $f^*(\mathbf{y}) = \sup_{\mathbf{y} \in \mathrm{dom}(f)} \mathbf{x}^{\mathrm{T}}\mathbf{y} - f(\mathbf{x})$. By the Fenchel–Moreau theorem (Borwein and Lewis (2006), theorem 4.2.1), f is convex and lower semicontinuous if and only if f is Fenchel biconjugate $f = f^{**}$. Therefore the first statement of lemma 1 follows from the following result. Its proof follows Borwein and Lewis (2006), theorem 5.2.2, except replacing Fan's inequality for eigenvalues by the version for singular values (lemma 2).

Lemma 3. Let f be a function of the ordered singular values and have Fenchel conjugate f^*. Then $J(\mathbf{X}) = (f \circ \sigma)(\mathbf{X})$ has Fenchel conjugate $J^*(\mathbf{Y}) = f^* \circ \sigma(\mathbf{Y})$.

Proof. By lemma 2,

$$(f \circ \sigma)^*(\mathbf{Y}) = \sup_{\mathbf{X}}[\mathrm{tr}(\mathbf{X}^{\mathrm{T}}\mathbf{Y}) - f\{\sigma(\mathbf{X})\}] \leqslant \sup_{\mathbf{X}}[\sigma(\mathbf{X})^{\mathrm{T}}\sigma(\mathbf{Y}) - f\{\sigma(\mathbf{X})\}]$$
$$\leqslant \sup_{\mathbf{x}}\{\mathbf{x}^{\mathrm{T}}\sigma(\mathbf{Y}) - f(\mathbf{x})\} = f^*\{\sigma(\mathbf{Y})\}.$$

In contrast, assume that $\mathbf{Y}$ has singular value decomposition $\mathbf{U}\,\mathrm{diag}(\mathbf{y})\mathbf{V}^{\mathrm{T}}$. Then

$$f^* \circ \sigma(\mathbf{Y}) = \sup_{\mathbf{x}}\{\sigma(\mathbf{Y})^{\mathrm{T}}\mathbf{x} - f(\mathbf{x})\} = \sup_{\mathbf{x}}[\mathrm{tr}\{\mathbf{U}^{\mathrm{T}}\mathbf{Y}\mathbf{V}\,\mathrm{diag}(\mathbf{x})\} - f(\mathbf{x})]$$
$$= \sup_{\mathbf{x}}[\mathrm{tr}\{\mathbf{Y}\mathbf{V}\,\mathrm{diag}(\mathbf{x})\mathbf{U}^{\mathrm{T}}\} - f \circ \sigma\{\mathbf{U}\,\mathrm{diag}(\mathbf{x})\mathbf{V}^{\mathrm{T}}\}]$$
$$\leqslant \sup_{\mathbf{X}}\{\mathrm{tr}(\mathbf{Y}\mathbf{X}^{\mathrm{T}}) - f \circ \sigma(\mathbf{X})\} = (f \circ \sigma)^*(\mathbf{Y}).$$

Therefore we have the identity $f^* \circ \sigma = (f \circ \sigma)^*$. □

To compute the subdifferential $\partial J(\mathbf{X}) = \partial(f \circ \sigma)(\mathbf{X})$ for a convex f. By lemma 3 and the Fenchel–Young inequality (Borewein and Lewis (2006), proposition 3.3.4),

$$(f \circ \sigma)(\mathbf{X}) + (f \circ \sigma)^*(\mathbf{Y}) = f \circ \sigma(\mathbf{X}) + f^* \circ \sigma(\mathbf{Y}) \geqslant \sigma(\mathbf{X})^{\mathrm{T}}\sigma(\mathbf{Y}) \geqslant \mathrm{tr}(\mathbf{X}^{\mathrm{T}}\mathbf{Y}).$$

Recall that, for a convex function f, $f(\mathbf{x}) + f^*(\mathbf{y}) = \mathbf{x}^{\mathrm{T}}\mathbf{y}$ if and only if $\mathbf{y} \in \partial f(\mathbf{x})$. Therefore a matrix $\mathbf{Y} \in \partial J(\mathbf{X})$ if and only if equality holds throughout in the above inequalities. Then, by lemma 2, $\mathbf{Y} \in \partial J(\mathbf{X})$ if and only if $\mathbf{X}$ and $\mathbf{Y}$ are simultaneously diagonalizable, $\mathbf{X} = \mathbf{U}\,\mathrm{diag}(\mathbf{x})\mathbf{V}^{\mathrm{T}}$ and $\mathbf{Y} = \mathbf{U}\,\mathrm{diag}(\mathbf{y})\mathbf{V}^{\mathrm{T}}$, and $\mathbf{y} \in \partial f(\mathbf{x})$.

A.2. Proof of proposition 1

The objective function is equal to $\frac{1}{2}\|\mathbf{B}\|_{\mathrm{F}}^2 - \mathrm{tr}(\mathbf{B}\mathbf{A}^{\mathrm{T}}) + \frac{1}{2}\|\mathbf{A}\|_{\mathrm{F}}^2 + f \circ \sigma(\mathbf{B})$. Suppose that the matrix $\mathbf{B}$ has singular value decomposition $\mathbf{B} = \mathbf{P}\,\mathrm{diag}(\mathbf{b})\mathbf{Q}^{\mathrm{T}}$. According to lemma 2, $-\mathrm{tr}(\mathbf{B}\mathbf{A}^{\mathrm{T}}) \geqslant -\Sigma_j b_j a_j$ with equality if and only if $\mathbf{P} = \mathbf{U}$ and $\mathbf{Q} = \mathbf{V}$. Neither the Frobenius norm $\|\mathbf{B}\|_{\mathrm{F}}^2$ nor the regularization term $f \circ \sigma(\mathbf{B})$ depends on the orthogonal matrices $\mathbf{P}$ and $\mathbf{Q}$. Therefore the optimal $\mathbf{B}$ has $\mathbf{P} = \mathbf{U}$ and $\mathbf{Q} = \mathbf{V}$ and its singular values are the minimizer of $\|\mathbf{b} - \mathbf{a}\|_2^2/2 + f(\mathbf{b})$.

A.3. Proof of theorem 3

The following standard calculus notation is used. For a scalar function f, ∇f is the (column) gradient vector, $\mathrm{d}f = [\nabla f]^{\mathrm{T}}$ is the differential and $\mathrm{d}^2 f$ is the Hessian matrix. For a multivariate function $g: \mathbb{R}^p \mapsto \mathbb{R}^q$, $\mathrm{D}g \in \mathbb{R}^{q \times p}$ denotes the Jacobian matrix holding partial derivatives $\partial g_i/\partial x_j$. For a matrix function $h: \mathbb{R}^{m \times n} \mapsto \mathbb{R}^{p \times q}$, $\mathrm{D}h \in \mathbb{R}^{pq \times mn}$ denotes the Jacobian matrix.

By the chain rule, the Jacobian of fitted values with respect to responses is

$$\mathrm{D}\hat{\mathbf{Y}}(\mathbf{Y}) = \mathrm{D}\hat{\mathbf{Y}}(\hat{\mathbf{B}}_\lambda)\,\mathrm{D}\hat{\mathbf{B}}_\lambda(\hat{\mathbf{B}}_{\mathrm{LS}})\,\mathrm{D}\hat{\mathbf{B}}_{\mathrm{LS}}(\mathbf{Y}).$$

Let $\boldsymbol{\Xi} = (\mathrm{vec}(\mathbf{X}_1) \ldots \mathrm{vec}(\mathbf{X}_n))^{\mathrm{T}}$ be the design matrix with vectorized matrix covariates. Then $\mathrm{D}\hat{\mathbf{Y}}(\hat{\mathbf{B}}_\lambda) = \boldsymbol{\Xi}$ and $\mathrm{D}\hat{\mathbf{B}}_{\mathrm{LS}}(\mathbf{Y}) = (\boldsymbol{\Xi}^{\mathrm{T}}\boldsymbol{\Xi})^{-1}\boldsymbol{\Xi}^{\mathrm{T}} = \boldsymbol{\Xi}^{\mathrm{T}}$. Therefore the degree of freedom of the fit is

$$\mathrm{tr}\{\mathrm{D}\hat{\mathbf{Y}}(\mathbf{Y})\} = \mathrm{tr}\{\boldsymbol{\Xi}\,\mathrm{D}\hat{\mathbf{B}}_\lambda(\hat{\mathbf{B}}_{\mathrm{LS}})\,\boldsymbol{\Xi}^{\mathrm{T}}\} = \mathrm{tr}\{\boldsymbol{\Xi}^{\mathrm{T}}\boldsymbol{\Xi}\,\mathrm{D}\hat{\mathbf{B}}_\lambda(\hat{\mathbf{B}}_{\mathrm{LS}})\} = \mathrm{tr}\{\mathrm{D}\hat{\mathbf{B}}_\lambda(\hat{\mathbf{B}}_{\mathrm{LS}})\}.$$

Denote the singular value decomposition of the usual least squares estimate $\hat{\mathbf{B}}_{\mathrm{LS}} \in \mathbb{R}^{p_1 \times p_2}$ by $\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{\mathrm{T}} = \Sigma_{i=1}^q \sigma_i \mathbf{u}_i \mathbf{v}_i^{\mathrm{T}}$ where $\mathbf{U} \in \mathbb{R}^{p_1 \times p_1}$, $\boldsymbol{\Sigma} \in \mathbb{R}^{p_1 \times p_2}$, $\mathbf{V} \in \mathbb{R}^{p_2 \times p_2}$ and $q = \min\{p_1, p_2\}$. Then the nuclear norm estimate is $\hat{\mathbf{B}}_\lambda = \mathbf{U}\boldsymbol{\Sigma}_\lambda\mathbf{V}$ where $\boldsymbol{\Sigma}_\lambda$ has diagonal entries $(\sigma_i - \lambda)_+$. Thus we have

$$\mathrm{tr}\{\mathrm{D}\,\hat{\mathbf{B}}_\lambda(\hat{\mathbf{B}}_{\mathrm{LS}})\} = \mathrm{tr}\left[\sum_{i=1}^q \{\mathrm{D}\hat{\mathbf{B}}_\lambda(\mathbf{v}_i)\mathrm{D}\mathbf{v}_i(\hat{\mathbf{B}}_{\mathrm{LS}}) + \mathrm{D}\hat{\mathbf{B}}_\lambda(\mathbf{u}_i)\,\mathrm{D}\mathbf{u}_i(\hat{\mathbf{B}}_{\mathrm{LS}}) + \mathrm{D}\hat{\mathbf{B}}_\lambda(\sigma_i)\,\mathrm{D}\sigma_i(\hat{\mathbf{B}}_{\mathrm{LS}})\}\right].$$

We then tackle this piece by piece.

Lemma 4. Let $\sigma_i > 0$ be a singular value of $\hat{\mathbf{B}}_{\mathrm{LS}}$ with multiplicity 1, left singular vector $\mathbf{u}_i$ and right singular vector $\mathbf{v}_i$. Then

$$\mathrm{tr}\{\mathrm{D}\hat{\mathbf{B}}_\lambda(\mathbf{v}_i)\,\mathrm{D}\mathbf{v}_i(\hat{\mathbf{B}}_{\mathrm{LS}})\} = \mathbf{1}_{\{\sigma_i > \lambda\}} \sum_{1 \leqslant j \leqslant p_2,\, j \neq i} \frac{\sigma_i(\sigma_i - \lambda)}{\sigma_i^2 - \sigma_j^2}, \tag{5}$$

$$\operatorname{tr}\{\mathrm{D}\hat{\mathbf{B}}_\lambda(\mathbf{u}_i)\,\mathrm{D}\mathbf{u}_i(\hat{\mathbf{B}}_{\mathrm{LS}})\} = \mathbf{1}_{\{\sigma_i>\lambda\}} \sum_{1\leqslant j\leqslant p_1, j\neq i} \frac{\sigma_i(\sigma_i-\lambda)}{\sigma_i^2-\sigma_j^2}. \tag{6}$$

Proof. Since $\hat{\mathbf{B}}_{\mathrm{LS}} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{\mathrm{T}}$, the eigenvectors of the symmetric matrix $\hat{\mathbf{B}}_{\mathrm{LS}}^{\mathrm{T}}\hat{\mathbf{B}}_{\mathrm{LS}} = \mathbf{V}\boldsymbol{\Sigma}^2\mathbf{V}^{\mathrm{T}}$ coincide with the right singular vectors of $\hat{\mathbf{B}}_{\mathrm{LS}}$. Then, by the chain rule,

$$\mathrm{D}\hat{\mathbf{B}}_\lambda(\mathbf{v}_i)\mathrm{D}\mathbf{v}_i(\hat{\mathbf{B}}_{\mathrm{LS}}) = \mathrm{D}\hat{\mathbf{B}}_\lambda(\mathbf{v}_i)\,\mathrm{D}\mathbf{v}_i(\hat{\mathbf{B}}_{\mathrm{LS}}^{\mathrm{T}}\hat{\mathbf{B}}_{\mathrm{LS}})\,\mathrm{D}(\mathbf{B}_{\mathrm{LS}}^{\mathrm{T}}\hat{\mathbf{B}}_{\mathrm{LS}})(\hat{\mathbf{B}}_{\mathrm{LS}}).$$

Now $\mathrm{D}\hat{\mathbf{B}}_\lambda(\mathbf{v}_i) = (\sigma_i-\lambda)\,\mathbf{1}_{\{\sigma_i>\lambda\}}\mathbf{I}_{p_2}\otimes\mathbf{u}_i$. By the well-known formula for the differential of eigenvectors (Magnus and Neudecker (1999), section 8.8), $\mathrm{D}\mathbf{v}_i(\hat{\mathbf{B}}_{\mathrm{LS}}^{\mathrm{T}}\hat{\mathbf{B}}_{\mathrm{LS}}) = \mathbf{v}_i^{\mathrm{T}}\otimes(\sigma_i^2\mathbf{I}_{p_2}-\hat{\mathbf{B}}_{\mathrm{LS}}^{\mathrm{T}}\hat{\mathbf{B}}_{\mathrm{LS}})^{+}$, where $\mathbf{A}^{+}$ is the Moore–Penrose generalized inverse of a matrix $\mathbf{A}$. The Jacobian of the symmetric product is $\mathrm{D}(\mathbf{B}_{\mathrm{LS}}^{\mathrm{T}}\hat{\mathbf{B}}_{\mathrm{LS}})(\hat{\mathbf{B}}_{\mathrm{LS}}) = (\mathbf{I}_{p_2^2}+\mathbf{K}_{p_2p_2})(\mathbf{I}_{p_2}\otimes\hat{\mathbf{B}}_{\mathrm{LS}}^{\mathrm{T}})$, where $\mathbf{K}_{p_2p_2}\in\mathbb{R}^{p_2^2\times p_2^2}$ is the commutation matrix (Magnus and Neudecker (1999), section 3.7). Now, by cyclic permutation invariance of the trace function,

$$\begin{aligned}\operatorname{tr}[(\sigma_i-\lambda)(\mathbf{I}_{p_2}\otimes\mathbf{u}_i)\{\mathbf{v}_i^{\mathrm{T}}\otimes(\sigma_i^2\mathbf{I}_{p_2}-\hat{\mathbf{B}}_{\mathrm{LS}}^{\mathrm{T}}\hat{\mathbf{B}}_{\mathrm{LS}})^{+}\}\mathbf{I}_{p_2}(\mathbf{I}_{p_2}\otimes\hat{\mathbf{B}}_{\mathrm{LS}}^{\mathrm{T}})]&\\ = (\sigma_i-\lambda)\operatorname{tr}\{\mathbf{v}_i^{\mathrm{T}}\otimes(\sigma_i^2\mathbf{I}_{p_2}-\hat{\mathbf{B}}_{\mathrm{LS}}^{\mathrm{T}}\hat{\mathbf{B}}_{\mathrm{LS}})^{+}\hat{\mathbf{B}}_{\mathrm{LS}}^{\mathrm{T}}\mathbf{u}_i\} = \sigma(\sigma_i-\lambda)\operatorname{tr}(\mathbf{v}_i^{\mathrm{T}}\otimes\mathbf{0}_{p_2}) = 0&\end{aligned}$$

and, recalling that $\mathbf{K}_{p_21} = \mathbf{I}_{p_2}$,

$$\begin{aligned}\operatorname{tr}[\{(\sigma_i-\lambda)\mathbf{I}_{p_2}\otimes\mathbf{u}_i\}\{\mathbf{v}_i^{\mathrm{T}}\otimes(\sigma_i^2\mathbf{I}_{p_2}-\hat{\mathbf{B}}_{\mathrm{LS}}^{\mathrm{T}}\hat{\mathbf{B}}_{\mathrm{LS}})^{+}\}\mathbf{K}_{p_2p_2}(\mathbf{I}_{p_2}\otimes\hat{\mathbf{B}}_{\mathrm{LS}}^{\mathrm{T}})]&\\ = (\sigma_i-\lambda)\operatorname{tr}\{(\sigma_i^2\mathbf{I}_{p_2}-\hat{\mathbf{B}}_{\mathrm{LS}}^{\mathrm{T}}\hat{\mathbf{B}}_{\mathrm{LS}})^{+}\otimes\mathbf{u}_i\mathbf{v}_i^{\mathrm{T}}\hat{\mathbf{B}}_{\mathrm{LS}}^{\mathrm{T}}\}&\\ = \sigma_i(\sigma_i-\lambda)\operatorname{tr}(\sigma_i^2\mathbf{I}_{p_2}-\hat{\mathbf{B}}_{\mathrm{LS}}^{\mathrm{T}}\hat{\mathbf{B}}_{\mathrm{LS}})^{+}\operatorname{tr}(\mathbf{u}_i\mathbf{u}_i^{\mathrm{T}})&\\ = \sum_{1\leqslant j\leqslant p_2, j\neq i}\frac{\sigma_i(\sigma_i-\lambda)}{\sigma_i^2-\sigma_j^2}\operatorname{tr}(\mathbf{v}_j\mathbf{v}_j^{\mathrm{T}}) = \sum_{1\leqslant j\leqslant p_2, j\neq i}\frac{\sigma_i(\sigma_i-\lambda)}{\sigma_i^2-\sigma_j^2}.&\end{aligned}$$

Combining pieces, we have equation (5). By symmetry, we also have equation (6). □

Lemma 5. Let $\sigma_i > 0$ be a singular value of $\hat{\mathbf{B}}_{\mathrm{LS}}$ with multiplicity 1. Then

$$\operatorname{tr}\{\mathrm{D}\hat{\mathbf{B}}_\lambda(\sigma_i)\,\mathrm{D}\sigma_i(\hat{\mathbf{B}}_{\mathrm{LS}})\} = \mathbf{1}_{\{\sigma_i>\lambda\}}.$$

Proof. Again we utilize the fact that σ_i is the positive square root of the eigenvalues λ_i of the symmetric matrix $\hat{\mathbf{B}}_{\mathrm{LS}}^{\mathrm{T}}\hat{\mathbf{B}}_{\mathrm{LS}}$. Then, by the chain rule,

$$\mathrm{D}\hat{\mathbf{B}}_\lambda(\sigma_i)\,\mathrm{D}\sigma_i(\hat{\mathbf{B}}_{\mathrm{LS}}) = \mathrm{D}\hat{\mathbf{B}}_\lambda(\sigma_i)\,\mathrm{D}\sigma_i(\lambda_i)\,\mathrm{D}\lambda_i(\hat{\mathbf{B}}_{\mathrm{LS}}^{\mathrm{T}}\hat{\mathbf{B}}_{\mathrm{LS}})\,\mathrm{D}(\hat{\mathbf{B}}_{\mathrm{LS}}^{\mathrm{T}}\hat{\mathbf{B}}_{\mathrm{LS}})(\hat{\mathbf{B}}_{\mathrm{LS}}).$$

Now combining $\mathrm{D}\hat{\mathbf{B}}_\lambda(\sigma_i) = \mathbf{1}_{\{\sigma_i>\lambda\}}\mathbf{v}_i\otimes\mathbf{u}_i$, $\mathrm{D}\sigma_i(\lambda_i) = 1/2\sqrt{\lambda_i} = 1/2\sigma_i$, $\mathrm{D}\lambda_i(\hat{\mathbf{B}}_{\mathrm{LS}}^{\mathrm{T}}\hat{\mathbf{B}}_{\mathrm{LS}}) = \mathbf{v}_i^{\mathrm{T}}\otimes\mathbf{v}_i^{\mathrm{T}}$ and

$$\mathrm{D}(\hat{\mathbf{B}}_{\mathrm{LS}}^{\mathrm{T}}\hat{\mathbf{B}}_{\mathrm{LS}})(\hat{\mathbf{B}}_{\mathrm{LS}}) = (\mathbf{I}_{p_2}+\mathbf{K}_{p_2p_2})(\mathbf{I}_{p_2}\otimes\hat{\mathbf{B}}_{\mathrm{LS}}^{\mathrm{T}})$$

shows that

$$\begin{aligned}\operatorname{tr}\{\mathrm{D}\hat{\mathbf{B}}_\lambda(\sigma_i)\mathrm{D}\sigma_i(\hat{\mathbf{B}}_{\mathrm{LS}})\} &= \mathbf{1}_{\{\sigma_i>\lambda\}}\frac{1}{2\sigma_i}\operatorname{tr}\{(\mathbf{v}_i\otimes\mathbf{u}_i)(\mathbf{v}_i^{\mathrm{T}}\otimes\mathbf{v}_i^{\mathrm{T}})\mathbf{I}_{p_2}(\mathbf{I}_{p_2}\otimes\hat{\mathbf{B}}_{\mathrm{LS}}^{\mathrm{T}})\}\\ &\quad+\mathbf{1}_{\{\sigma_i>\lambda\}}\frac{1}{2\sigma_i}\operatorname{tr}\{(\mathbf{v}_i\otimes\mathbf{u}_i)(\mathbf{v}_i^{\mathrm{T}}\otimes\mathbf{v}_i^{\mathrm{T}})\mathbf{K}_{p_2p_2}(\mathbf{I}_{p_2}\otimes\hat{\mathbf{B}}_{\mathrm{LS}}^{\mathrm{T}})\}\\ &= \mathbf{1}_{\{\sigma_i>\lambda\}}\frac{1}{2\sigma_i}\operatorname{tr}(\mathbf{v}_i\mathbf{v}_i^{\mathrm{T}}\otimes\mathbf{u}_i\mathbf{v}_i^{\mathrm{T}}\hat{\mathbf{B}}_{\mathrm{LS}}^{\mathrm{T}})+\mathbf{1}_{\{\sigma_i>\lambda\}}\frac{1}{2\sigma_i}\operatorname{tr}(\mathbf{v}_i\mathbf{v}_i^{\mathrm{T}}\otimes\mathbf{u}_i\mathbf{v}_i^{\mathrm{T}}\hat{\mathbf{B}}_{\mathrm{LS}}^{\mathrm{T}})\\ &= \mathbf{1}_{\{\sigma_i>\lambda\}}\frac{1}{\sigma_i}\operatorname{tr}(\sigma_i\mathbf{v}_i\mathbf{v}_i^{\mathrm{T}}\otimes\mathbf{u}_i\mathbf{u}_i^{\mathrm{T}}) = \mathbf{1}_{\{\sigma_i>\lambda\}}.\end{aligned}$$

Finally, combining lemmas 4 and 5 yields the degrees of freedom in theorem 3.

References

Akaike, H. (1974) A new look at the statistical model identification. *IEEE Trans. Autom. Control*, **19**, 716–723.

Allen, G., Grosenick, L. and Taylor, J. (2011) A generalized least squares matrix decomposition. Rice University, Houston. *Preprint arXiv:1102.3074*.

American Electroencephalographic Society (1991) American Electroencephalographic Society guidelines for standard electrode position nomenclature. *J. Clin. Neurophysiol.*, **8**, 200–202.

Armagan, A., Dunson, D. and Lee, J. (2013) Generalized double Pareto shrinkage. *Statist. Sin.*, **23**, 119–143.

Beck, A. and Teboulle, M. (2009a) A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imgng Sci.*, **2**, 183–202.
Beck, A. and Teboulle, M. (2009b) Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems. *Trans. Image Process.*, **18**, 2419–2434.
Borwein, J. M. and Lewis, A. S. (2006) *Convex Analysis and Nonlinear Optimization*, 2nd edn. New York: Springer.
Caffo, B. S., Crainiceanu, C. M., Verduzco, G., Joel, S., Mostafsky, S. H., Bassett, S. S and Pekar, J. J. (2010) Two-stage decompositions for the analysis of functional connectivity for fMRI with application to Alzheimer's disease risk. *NeuroImage*, **51**, 1140–1149.
Cai, J.-F., Candès, E. J. and Shen, Z. (2010) A singular value thresholding algorithm for matrix completion. *SIAM J. Optimizn*, **20**, 1956–1982.
Candès, E. J. and Recht, B. (2009) Exact matrix completion via convex optimization. *Found. Comput. Math.*, **9**, 717–772.
Candès, E. J., Wakin, M. B. and Boyd, S. P. (2008) Enhancing sparsity by reweighted l_1 minimization. *J. Four. Anal. Appl.*, **14**, 877–905.
Chen, S. S., Donoho, D. L. and Saunders, M. A. (2001) Atomic decomposition by basis pursuit. *SIAM Rev.*, **43**, 129–159.
Chiquet, J., Grandvalet, Y. and Charbonnier, C. (2012) Sparsity with sign-coherent groups of variables via the cooperative-lasso. *Ann. Appl. Statist.*, **6**, 795–830.
Crainiceanu, C. M., Caffo, B. S., Luo, S., Zipunnikov, V. M. and Punjabi, N. M. (2011) Population value decomposition, a framework for the analysis of image populations. *J. Am. Statist. Ass.*, **106**, 775–790.
Donoho, D. L. and Johnstone, I. M. (1994) Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, **81**, 425–455.
Efron, B. (2004) The estimation of prediction error: covariance penalties and cross-validation (with comments). *J. Am. Statist. Ass.*, **99**, 619–642.
Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004) Least angle regression (with discussion). *Ann. Statist.*, **32**, 407–499.
Fan, J. and Li, R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Statist. Ass.*, **96**, 1348–1360.
Frank, I. E. and Friedman, J. H. (1993) A statistical view of some chemometrics regression tools. *Technometrics*, **35**, 109–135.
Golub, G. H. and Van Loan, C. F. (1996) *Matrix Computations*, 3rd edn. Baltimore: Johns Hopkins University Press.
Hoerl, A. E. and Kennard, R. W. (1970) Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, **12**, 55–67.
Hung, H. and Wang, C.-C. (2013) Matrix variate logistic regression model with application to eeg data. *Biostatistics*, **14**, 189–202.
Johnstone, I. M. (2001) On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.*, **29**, 295–327.
Lange, K. (2004) *Optimization*. New York: Springer.
Li, B., Kim, M. K. and Altman, N. (2010) On dimension folding of matrix- or array-valued statistical objects. *Ann. Statist.*, **38**, 1094–1121.
Magnus, J. R. and Neudecker, H. (1999) *Matrix Differential Calculus with Applications in Statistics and Econometrics*. Chichester: Wiley.
Marshall, A. W., Olkin, I. and Arnold, B. C. (2011) *Inequalities: Theory of Majorization and Its Applications*, 2nd edn. New York: Springer.
Mazumder, R., Hastie, T. and Tibshirani, R. (2010) Spectral regularization algorithms for learning large incomplete matrices. *J. Mach. Learn. Res.*, **11**, 2287–2322.
McCullagh, P. and Nelder, J. A. (1983) *Generalized Linear Models*. London: Chapman and Hall.
Nemirovski, A. (1994) Efficient methods in convex programming. (Available from `http://www2.isye.gatech.edu/~nemirovs/Lect_ECMO.pdf`.)
Nesterov, Y. (1983) A method of solving a convex programming problem with convergence rate $O(1/k^2)$. *Sov. Math. Dokl.*, **27**, 372–376.
Nesterov, Y. (2004) *Introductory Lectures on Convex Optimization*. Boston: Kluwer.
Nocedal, J. and Wright, S. J. (2006) *Numerical Optimization*, 2nd edn. New York: Springer.
Qiu, P. (2005) *Image Processing and Jump Regression Analysis*. New York: Wiley.
Qiu, P. (2007) Jump surface estimation, edge detection, and image restoration. *J. Am. Statist. Ass.*, **102**, 745–756.
Recht, B., Fazel, M. and Parrilo, P. A. (2010) Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Rev.*, **52**, 471–501.
Schwarz, G. (1978) Estimating the dimension of a model. *Ann. Statist.*, **6**, 461–464.
Stein, C. M. (1981) Estimation of the mean of a multivariate normal distribution. *Ann. Statist.*, **9**, 1135–1151.
Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *J. R. Statist. Soc.* B, **58**, 267–288.
Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. and Knight, K. (2005) Sparsity and smoothness via the fused lasso. *J. R. Statist. Soc.* B, **67**, 91–108.
Yuan, M. and Lin, Y. (2006) Model selection and estimation in regression with grouped variables. *J. R. Statist. Soc.* B, **68**, 49–67.

Zhang, C.-H. (2010) Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.*, **38**, 894–942.
Zhang, X., Begleiter, H., Porjesz, B., Wang, W. and Litke, A. (1995) Event related potentials during object recognition tasks. *Brain Res. Bull.*, **38**, 531–538.
Zhou, H., Li, L. and Zhu, H. (2013) Tensor regression with applications in neuroimaging data analysis. *J. Am. Statist. Ass.*, to be published.
Zou, H. and Hastie, T. (2005) Regularization and variable selection via the elastic net. *J. R. Statist. Soc.* B, **67**, 301–320.
Zou, H., Hastie, T. and Tibshirani, R. (2007) On the "degrees of freedom" of the lasso. *Ann. Statist.*, **35**, 2173–2192.