Risk estimation for matrix recovery with spectral regularization

Charles-Alban Deledalle, Samuel Vaiter, Gabriel Peyré, Jalal M. Fadili, Charles Dossal

Risk estimation for matrix recovery with spectral regularization

Charles-Alban Deledalle, Samuel Vaiter, Gabriel Peyré DELEDALLE@CEREMADE.DAUPHINE.FR

CEREMADE, CNRS, Université Paris-Dauphine, France

Jalal Fadili

GREYC, CNRS-ENSICAEN-Université de Caen, France

Charles Dossal

IMB, Université Bordeaux 1, France

Abstract

In this paper, we develop an approach to recursively estimate the quadratic risk for matrix recovery problems regularized with spectral functions. Toward this end, in the spirit of the SURE theory, a key step is to compute the (weak) derivative and divergence of a solution with respect to the observations. As such a solution is not available in closed form, but rather through a proximal splitting algorithm, we propose to recursively compute the divergence from the sequence of iterates. A second challenge that we unlocked is the computation of the (weak) derivative of the proximity operator of a spectral function. To show the potential applicability of our approach, we exemplify it on a matrix completion problem to objectively and automatically select the regularization parameter.

1. Introduction

Consider the problem of estimating a matrix $X_0 \in \mathbb{R}^{n_1 \times n_2}$ from P noisy observations $y = \mathcal{A}(X_0) + w \in \mathbb{R}^P$, where $w \sim \mathcal{N}(0, \sigma^2 \mathrm{Id}_P)$. The linear bounded operator $\mathcal{A} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^P$ entails loss of information such that the problem is ill-posed. This problem arises in various research fields. Because of ill-posedness, side information through a regularizing term is necessary. We thus consider the problem

$$X(y) \in \operatorname*{Argmin}_{X \in \mathbb{R}^{n_1 \times n_2}} \frac{1}{2}\|y - \mathcal{A}(X)\|^2 + \lambda J(X) \tag{1}$$

where the set of minimizers is assumed non-empty, $\lambda > 0$ is a regularization parameter and $J : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R} \cup \{\infty\}$ is a proper lower semi-continuous (lsc) convex regularizing function that imposes the desired structure on $X(y)$. In this paper, we focus on the case where J is a convex spectral function, that is a symmetric convex function of the singular values of its argument. Spectral regularization can account for prior knowledge on the spectrum of X_0, typically low-rank (see e.g. Fazel, 2002).

In practice, the choice of the regularization parameter λ in (1) remains an important problem largely unexplored. Typically, we want to select λ minimizing the quadratic risk $\mathbb{E}_w\|X(y) - X_0\|^2$. Since X_0 is unknown and $X(y)$ is non-unique, one can instead consider an unbiased estimate of the *prediction* risk $\mathbb{E}_w\|\mathcal{A}(X(y)) - \mathcal{A}(X_0)\|^2$, where it can be easily shown that $\mu(y) = \mathcal{A}(X(y))$ is a single-valued mapping. With the proviso that $\mu(y)$ is weakly

This work was presented at the *ICML Workshop Sparsity, Dictionaries and Projections in Machine Learning and Signal Processing*, Edinburgh, Scotland, UK, 2012.

Acknowledgement: We would like to thank E. Candès, C. A. Sing-Long and J. D. Trzasko for bringing their work to our attention and for fruitful discussions.

differentiable, the SURE (for Stein unbiased risk estimator, Stein, 1981)

$$\mathrm{SURE}(y) = \|y - \mu(y))\|^2 - P\sigma^2 + 2\sigma^2 \operatorname{div} \mu(y) \tag{2}$$

is an unbiased estimate of the prediction risk, where $\operatorname{div} \mu(y) = \operatorname{Tr}(\partial\mu(y))$, and $\partial\mu(y)$ stands for the (weak) Jacobian of $\mu(y)$. The SURE depends solely on y, without prior knowledge of X_0 and then can prove very useful as a basis for automatic ways to choose the regularization parameters λ.

Contributions. Our main contribution is to provide the derivative of matrix-valued spectral functions where the matrices have distinct singular values which extends the result of Lewis & Sendov (2001) to non-symmetric square matrices. This result is used to recursively compute the derivative of any solution of spectrally regularized inverse problems by solving (1). This is achieved by computing the derivatives of the iterates provided by a proximal splitting algorithm. In particular, this provides an estimate of $\operatorname{div} \mu(y)$ in (2) which allows to compute $\mathrm{SURE}(y)$. A Numerical example on a matrix completion problem is given to support our findings.

2. Recursive risk estimation

Proximal splitting Proximal splitting algorithms have become extremely popular to solve non-smooth convex optimization problems that arise often in inverse problems, e.g. (1). These algorithms provide a sequence of iterates $X^{(\ell)}(y)$ that provably converges to a solution $X(y)$. A practical way to compute $\operatorname{div} \mu(y)$, hence $\mathrm{SURE}(y)$, as initiated by Vonesch et al. (2008), and that we pursue here, consists in differentiating this sequence of iterates. This methodology has been extended to a wide class of proximal splitting schemes in (Deledalle et al., 2012). For the sake of clarity, and without loss of generality, we focus on the case of the forward-backward (FB) splitting algorithm (Combettes & Wajs, 2005).

The FB scheme is a good candidate to solve (1) if J is simple, meaning that its proximity operator has a closed-form. Recall that the proximity operator of a lsc proper convex function G on $\mathbb{R}^{n_1 \times n_2}$ is

$$\mathrm{Prox}_G(X) = \operatorname*{argmin}_{Z \in \mathbb{R}^{n_1 \times n_2}} \frac{1}{2}\|X - Z\|_F^2 + G(Z).$$

The FB algorithm iteration reads

$$X^{(\ell+1)} = \mathrm{Prox}_{\tau\lambda J}(X^{(\ell)} + \tau\mathcal{A}^*(y - \mathcal{A}(X^{(\ell)}))) \tag{3}$$

where $\mathcal{A}^*$ denotes the adjoint operator of $\mathcal{A}$, $\tau > 0$ is chosen such that $\tau\|\mathcal{A}^*\mathcal{A}\| < 2$, the dependency of the iterate $X^{(\ell)}$ to y is dropped to lighten the notation.

Risk estimation The divergence term $\operatorname{div} \mu(y)$ is obtained by differentiating formula (3), which allows, for any vector $\delta \in \mathbb{R}^P$ to compute iteratively $\xi^{(\ell)} = \partial X^{(\ell)}(y)[\delta]$ (the derivative of $y \mapsto X^{(\ell)}(y)$ at y in the direction δ) as

$$\begin{aligned} & \xi^{(\ell+1)} = \partial\, \mathrm{Prox}_{\tau\lambda J}(\Xi^{(\ell)})[\zeta^{(\ell)}] \\ \text{where} \quad & \Xi^{(\ell)} = X^{(\ell)} + \tau\mathcal{A}^*(y - \mathcal{A}(X^{(\ell)})) \\ \text{and} \quad & \zeta^{(\ell)} = \xi^{(\ell)} + \tau\mathcal{A}^*(\delta - \mathcal{A}(\xi^{(\ell)})). \end{aligned}$$

Using the Jacobian trace formula of the divergence, it can be easily seen that

$$\operatorname{div} \mu(y) = \mathbb{E}_\delta\langle \partial\mu(y)[\delta],\, \delta\rangle \approx \frac{1}{k}\sum_{i=1}^{k}\langle \partial\mu(y)[\delta_i],\, \delta_i\rangle \tag{4}$$

where $\delta \sim \mathcal{N}(0, \mathrm{Id}_P)$ and δ_i are k realizations of δ. The $\mathrm{SURE}(y)$ can in turn be iteratively estimated by plugging $\partial\mu(y)[\delta_i] = \mathcal{A}(\partial X^{(\ell)}(y)[\delta_i])$ in (4).

3. Local behavior of spectral functions

This section studies the local behavior of real- and matrix-valued spectral functions. We write the (full) singular value decomposition (SVD) of a matrix $X \in \mathbb{R}^{n_1 \times n_2}$

$$X = V_X \operatorname{diag}(\Lambda_X) U_X^*$$

(which might not be in general unique), where $\Lambda_X \in \mathbb{R}^n$ is the vector of singular values of X with $n = \min(n_1, n_2)$, $\operatorname{diag}(\Lambda_X) \in \mathbb{R}^{n_1 \times n_2}$ denotes the rectangular matrix with entries Λ_X on its main diagonal and 0 otherwise, and $V_X \in \mathbb{R}^{n_1 \times n_1}$ and $U_X \in \mathbb{R}^{n_2 \times n_2}$ are the unitary matrices of left and right singular vectors.

3.1. Scalar-valued Spectral Functions

A real-valued spectral function J can by definition be written as

$$J(X) = \varphi(\Lambda_X) \tag{5}$$

where $\varphi : \mathbb{R}^n \to \mathbb{R}$ is a symmetric function of its argument, meaning $\varphi(P\Lambda) = \varphi(\Lambda)$ for any permutation matrix $P \in \mathbb{R}^{n \times n}$ and Λ in the domain of φ. We extend φ to the negative half-line as $\varphi(\Lambda) = \varphi(|\Lambda|)$.

We then consider J a scalar-valued spectral function as defined in (5). From subdifferential calculus on spectral functions Lewis (1995), we get the following.

Proposition 1. *A spectral function $J(X) = \varphi(\Lambda_X)$ is convex if and only if φ is convex, and then*

$$\forall \gamma > 0, \quad \operatorname{Prox}_{\gamma J}(X) = V_X \operatorname{diag}(\operatorname{Prox}_{\gamma\varphi}(\Lambda_X)) U_X^*.$$

3.2. Matrix-valued Spectral Functions

We now turn to matrix-valued spectral functions

$$F(X) = V_X \operatorname{diag}(\Phi(\Lambda_X)) U_X^*, \tag{6}$$

where $\Phi : \mathbb{R}^n \to \mathbb{R}^n$ is symmetric in its arguments, meaning $\Phi \circ P = P \circ \Phi$ for any permutation matrix $P \in \mathbb{R}^{n \times n}$. We extend Φ to negative numbers as $\Phi(\Lambda) = \operatorname{sign}(\Lambda) \odot \Phi(|\Lambda|)$ and $\odot$ is the entry-wise matrix multiplication. One can observe that for $F(X) = \operatorname{Prox}_{\gamma J}(X)$ with $\Phi = \operatorname{Prox}_{\gamma\varphi}$, the proximity operator of a convex scalar-valued spectral function is a matrix-valued spectral function.

The following theorem provides a closed-form expression of the derivative of F when X is square, i.e. $n_1 = n_2 = n$, with distinct singular values.

Theorem 1. *For any matrix-valued spectral function F in (6), let the quantity*

$$\forall\, \delta \in \mathbb{R}^{n_1 \times n_2}, \quad D(X)[\delta] = V_X(\mathcal{M}(\Lambda_X)[\bar{\delta}] + \Gamma_S(\Lambda_X) \odot \mathcal{P}_S(\bar{\delta}) + \Gamma_A(\Lambda_X) \odot \mathcal{P}_A(\bar{\delta})) U_X^*$$

where $\bar{\delta} = V_X^* \delta U_X \in \mathbb{R}^{n_1 \times n_2}$, the symmetric and anti-symmetric parts are defined, for $1 \le i \le n_1$ and $1 \le j \le n_2$, as

$$\mathcal{P}_S(Y)_{i,j} = \begin{cases} \frac{Y_{i,j} + Y_{j,i}}{2} & \text{if} \quad i \le n \quad \text{and} \quad j \le n \\ \frac{Y_{i,j}}{2} & \text{otherwise} \end{cases}$$

$$\text{and} \quad \mathcal{P}_A(Y)_{i,j} = \begin{cases} \frac{Y_{i,j} - Y_{j,i}}{2} & \text{if} \quad i \le n \quad \text{and} \quad j \le n \\ \frac{Y_{i,j}}{2} & \text{otherwise} \end{cases}$$

and $\mathcal{M}(\Lambda) : \mathbb{R}^{n_1 \times n_2} \mapsto \mathbb{R}^{n_1 \times n_2}$ is

$$\mathcal{M}(\Lambda) = \operatorname{diag} \circ \partial\Phi(\Lambda) \circ \operatorname{diag}.$$

The matrices $\Gamma_S(\Lambda)$ *and* $\Gamma_A(\Lambda)$ *are defined, for all* $1 \leqslant i \leqslant n_1$ *and* $1 \leqslant j \leqslant n_2$, *as*

$$\Gamma_S(\Lambda)_{i,j} = \begin{cases} 0 & \text{if} \quad i = j \\ \frac{\Phi(\Lambda)_i - \Phi(\Lambda)_j}{\Lambda_i - \Lambda_j} & \text{if} \quad \Lambda_i \neq \Lambda_j \\ \partial\Phi(\Lambda)_{i,i} - \partial\Phi(\Lambda)_{i,j} & \text{otherwise,} \end{cases}$$

$$\Gamma_A(\Lambda)_{i,j} = \begin{cases} 0 & \text{if} \quad i = j \\ \frac{\Phi(\Lambda)_i + \Phi(\Lambda)_j}{\Lambda_i + \Lambda_j} & \text{if} \quad \Lambda_i > 0 \quad \text{or} \quad \Lambda_j > 0 \\ \partial\Phi(\Lambda)_{i,i} - \partial\Phi(\Lambda)_{i,j} & \text{otherwise.} \end{cases}$$

where for $i > n$ *we have extended* Λ *and* $\Phi(\Lambda)$ *as* $\Lambda_i = 0$ *and* $\Phi(\Lambda)_i = 0$.
Assume that X *is a square matrix, i.e.* $n_1 = n_2 = n$, *and with distinct singular values, such that* $\Lambda_i \neq \Lambda_j$ *for all* $i \neq j$. *Then, a matrix-valued spectral function* F *is differentiable at* X *if and only if* Φ *is differentiable at* Λ_X. *Moreover,*

$$\forall\, \delta \in \mathbb{R}^{n \times n}, \quad \partial F(X)[\delta] = D(X)[\delta]$$

The proof is given in Appendix A.

Theorem 1 generalizes the result of Lewis & Sendov (2001) to square matrices that are not necessarily symmetric, and we recover their formula when X and δ are symmetric matrices and X has distinct singular values. Regularity properties and expression of the directional derivative of symmetric matrix-valued separable spectral functions (possibly non-smooth) over non-necessarily symmetric matrices were also derived in Sun & Sun (2003). Before revising the previous version of this manuscript, Candès et al. (2012) brought to our attention their recent work on the SURE framework for parameter selection in denoising low-rank matrix data. Towards this goal, they provided closed-form expressions for the directional derivative and divergence of matrix-valued spectral functions over rectangular matrices with distinct singular values. They also addressed the case of complex-valued matrices.

Although our proof of Theorem 1 is rigorously valid only for square matrices with distinct singular values, we conjecture that the formula of the directional derivative holds for rectangular matrices with repeated singular values. For the symmetric case with repeated eigenvalues, this assertion was formally proved in Lewis & Sendov (2001). As stated above, the full-rank rectangular case with distinct singular values was proved in Candès et al. (2012), where it was also shown that the divergence formula has a continuous extension to all matrices.

4. Numerical applications

4.1. Nuclear norm regularization

We here consider the problem of recovering a low-rank matrix $X_0 \in \mathbb{R}^{n_1 \times n_2}$. To this end, J is taken as the nuclear norm (a.k.a., trace or Schatten 1-norm) which is in some sense the tightest convex relaxation to the NP-hard rank minimization problem (Candès & Recht, 2009). The nuclear norm is defined by

$$J(X) = \|X\|_* \triangleq \|\Lambda_X\|_1. \tag{7}$$

Taking $J(\cdot)$ as $\|\cdot\|_*$ and φ as $\|.\|_1$ in Proposition 1 gives:

Corollary 1. *The proximal operator of* $\gamma\|\cdot\|_*$ *is*

$$\forall \gamma > 0, \quad \mathrm{Prox}_{\gamma\|\cdot\|_*}(X) = V_X \,\mathrm{diag}(T_\gamma(\Lambda_X))U_X^* \ , \tag{8}$$

where $T_\gamma = \mathrm{Prox}_{\gamma\|.\|_1}$ *is the component-wise soft-thresholding, defined for* $i = 1, \ldots, n$ *as*

$$T_\gamma(t)_i = \max(0, 1 - \gamma/\|t_i\|)t_i.$$

We now turn to the derivative of $F = \mathrm{Prox}_{\gamma\|\cdot\|_*}$. A straightforward attempt is to take $\Phi = \mathrm{Prox}_{\gamma\|.\|_1} = T_\gamma$ and apply Theorem 1 with

$$\partial\Phi(X)[\delta]_i = \partial T_\gamma(t)[\delta]_i = \begin{cases} 0 & \text{if} \quad \|t_i\| \leqslant \gamma \\ \delta_i & \text{otherwise.} \end{cases} \tag{9}$$

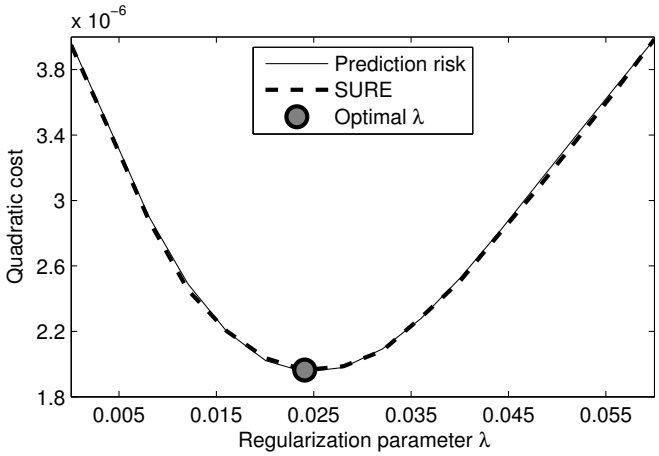

Figure 1. Predicted risk and its SURE estimate[1].

However, strictly speaking, Theorem 1 does not apply since a proximity mapping is 1-Lipschitz in general, hence not necessarily differentiable everywhere. Thus, its derivative may be set-valued, as is the case for soft-thresholding at $\pm\gamma$.

A direct consequence of Corollary 1 is that J is a simple function allowing for the use of the FB algorithm. Moreover, the expression of the derivative (9) provides an estimation of the SURE as explained in Section 2.

4.2. Application to matrix completion

We now exemplify the proposed SURE computation approach on a matrix completion problem encountered in recommendation systems such as the popular Netflix problem. We therefore consider the forward model $y = \mathcal{A}(X_0) + w \in \mathbb{R}^P$, $w \sim \mathcal{N}(0, \sigma^2 \mathrm{Id}_P)$, where X_0 is a dense but low-rank (or approximately so) matrix and A is binary masking operator.

We have taken $(n_1, n_2) = (1000, 100)$ and $P = 25000$ observed entries (i.e., 25%). The underlying dense matrix X_0 has been chosen to be approximately low-rank with a rapidly decaying spectrum $\Lambda_{X_0} = \{k^{-1}\}_{k=1}^n$. The standard deviation σ has been set such that the resulting minimum least-square estimate has a relative error $\|X_{LS} - X_0\|_F / \|X_0\|_F = 0.9$. Figure 1 depicts the prediction risk and its SURE estimate as a function of λ. For each value of λ in the tested range, SURE(y) in (2) has been computed for a single realization of y with $k = 4$ realizations δ_i in (4)[1]. At the optimal λ value, $X(y)$ has a rank of 55 with a relative error of 0.46 (i.e., a gain of about a factor 2 w.r.t. the least-square estimator).

5. Conclusion

The core theoretical contribution of this paper is the derivative of square matrix-valued spectral functions. This was a key step to compute the derivative of the proximal operator associated to the nuclear norm, and finally to use the SURE to recursively estimate the quadratic prediction risk of matrix recovery problems involving the nuclear norm regularization. The SURE was also used to automatically select the optimal regularization parameter.

A. Summary of the proof of Theorem 1

The following lemma derives the expression of the derivative of the SVD mapping $X \mapsto (V_X, \Lambda_X, U_X)$. Note that this mapping is not well defined because even if the Λ_X are distinct, one can apply arbitrary sign changes and permutations to the set of singular vectors. The lemma should thus be interpreted in the sense that one can locally write a Taylor expansion using the given differential for any particular choice of SVD. We point out that

[1]Without impacting the optimal choice of λ, the two curves have been vertically shifted for visualization.

a proof of this lemma using wedge products can be found in (Edelman, 2005), but for the sake of completeness, we provide our own proof here.

Lemma 1. *We consider $X_0 \in \mathbb{R}^{n\times n}$ with distinct singular values. For any matrix X in a neighborhood of X_0, we can define without ambiguity the SVD mapping $X \mapsto (V_X, \Lambda_X, U_X)$ by sorting the values in Λ_X and imposing sign constraints on U_X. The singular value mapping $X \mapsto (V_X, \Lambda_X, U_X)$ is C^1 and for a given matrix δ, its directional derivative is*

$$\begin{aligned}\partial\Lambda_X[\delta] &= \operatorname{diag}(V_X^*\delta U_X)\ ,\\ \partial V_X[\delta] &= V_X\iota_V,\\ \partial U_X[\delta] &= U_X\iota_U\end{aligned}$$

where $\iota_V \in \mathbb{R}^{n\times n}$ and $\iota_U \in \mathbb{R}^{n\times n}$ are defined, for all $1 \leqslant i \leqslant n$ and $1 \leqslant j \leqslant n$, as

$$(\iota_V)_{i,j} = \frac{(\Lambda_X)_j\bar\delta_{i,j} + (\Lambda_X)_i\bar\delta_{j,i}}{(\Lambda_X)_j^2 - (\Lambda_X)_i^2} \quad \text{and} \quad (\iota_U)_{i,j} = \frac{(\Lambda_X)_i\bar\delta_{i,j} + (\Lambda_X)_j\bar\delta_{j,i}}{(\Lambda_X)_j^2 - (\Lambda_X)_i^2}\ . \tag{10}$$

and where $\bar\delta = V_X^*\delta U_X \in \mathbb{R}^{n\times n}$.

Proof. Let S_n be the sub-space of Hermitian matrix in $\mathbb{R}^{n\times n}$. Let $\psi : \mathbb{R}^{n\times n}\times\mathcal{Y} \to \mathbb{R}^{n\times n}\times S_n\times S_n$, where $\mathcal{Y} = \mathbb{R}^{n\times n}\times\mathbb{R}^{n\times n}\times\mathbb{R}^n$, be defined for $Y = (V, U, \Lambda) \in \mathcal{Y}$ as

$$\psi(X, Y) = \begin{pmatrix} X - V\operatorname{diag}(\Lambda)U^*, & V^*V - \mathrm{Id}, & U^*U - \mathrm{Id}\end{pmatrix}.$$

We have for any vector $\zeta_X \in \mathbb{R}^{n\times n}$

$$\partial_1\psi(X, Y)[\zeta_X] = \begin{pmatrix}\zeta_X, & 0, & 0\end{pmatrix} \tag{11}$$

and for any vector $\zeta_Y = (\zeta_U, \zeta_V, \zeta_\Lambda) \in \mathcal{Y}$

$$\partial_2\psi(X, Y)[\zeta_Y] = \begin{pmatrix} -\zeta_V\operatorname{diag}(\Lambda)U^* - V\operatorname{diag}(\Lambda)\zeta_U^* - V\operatorname{diag}(\zeta_\Lambda)U^*, & V^*\zeta_V + \zeta_V^*V, & U^*\zeta_U + \zeta_U^*U\end{pmatrix}.$$

Let X_0 have distinct singular values and any of its SVD $Y_0 = (U_0, V_0, \Lambda_0)$. We have $\psi(X_0, Y_0) = 0$. Moreover, denoting $\iota_V = V_0^*\zeta_V \in \mathbb{R}^{n\times n}$ and $\iota_U = U_0^*\zeta_U \in \mathbb{R}^{n\times n}$, for any $z = (z_1, z_2, z_3) \in \mathbb{R}^{n\times n}\times S_n\times S_n$, solving $\partial_2\psi(X_0, Y_0)[\zeta_Y] = z$ is equivalent to solving

$$\iota_V\operatorname{diag}(\Lambda_0) + \operatorname{diag}(\Lambda_0)\iota_U^* + \operatorname{diag}(\zeta_\Lambda) = -V_0^*z_1U_0 = -\bar z_1 \tag{12}$$

where $\iota_V + \iota_V^* = z_2$ and $\iota_U + \iota_U^* = z_3$. Considering $z_2 = 0$ and $z_3 = 0$ shows that ι_V and ι_U are antisymmetric. In particular they are zero along the diagonal. Thus applying the operator diag to both sides of (12) shows $\zeta_\Lambda = -\operatorname{diag}(V_0^*z_1U_0)$. Now considering the entries (i, j) and (j, i) of the linear system (12) shows that for any $1 \leqslant i \leqslant n$ and $1 \leqslant j \leqslant n$

$$\begin{pmatrix}(\Lambda_0)_j & -(\Lambda_0)_i\\ -(\Lambda_0)_i & (\Lambda_0)_j\end{pmatrix}\begin{pmatrix}(\iota_V)_{i,j}\\ (\iota_U)_{i,j}\end{pmatrix} = \begin{pmatrix}-(\bar z_1)_{i,j}\\ -(\bar z_1)_{j,i}\end{pmatrix}. \tag{13}$$

Since for $i \neq j$, $(\Lambda_0)_i \neq (\Lambda_0)_j$, these 2×2 symmetric linear systems can be solved. Then $\partial_2\psi(X_0, Y_0)$ is invertible on $\mathbb{R}^{n\times n}\times 0_n\times 0_n$ and for $z = (z_1, 0, 0)$, its inverse is

$$(\partial_2\psi(X_0, Y_0))^{-1}[z] = \begin{pmatrix} V_0\iota_V, & U_0\iota_U, & -\operatorname{diag}(V_0^*z_1U_0)\end{pmatrix} \tag{14}$$

where ι_V and ι_U are given by the solutions of the above series of 2×2 symmetric linear systems.

Since $\operatorname{Im}(\partial_1\psi(X, Y(X))) \subset \mathbb{R}^{n\times n}\times 0_n\times 0_n$, we can apply the implicit function theorem (Rockafellar & Wets, 2005). Hence, for any $X \in \mathbb{R}^{n\times n}$ in the neighborhood of X_0, there exists a function $Y(X) = (U_X, V_X, \Lambda_X)$ such that $\psi(X, Y(X)) = 0$, i.e. X admits an SVD. Moreover, this function is C^1 in the neighborhood of X_0 and its differential is

$$\partial Y(X) = -\partial_2\psi(X, Y(X))^{-1}\circ\partial_1\psi(X, Y(X)).$$

Injecting (11) and (14) gives the desired formula by solving (13) in closed form. Since X_0 is any matrix with distinct singular values, we can conclude. □

We now turn to the proof of the theorem.

Proof. Since the singular values of X are all distinct, by composition of differentiable functions, we can derive the relationship (6) that defines F which gives

$$V_X^* \partial F(X)[\delta] U_X = \iota_V \operatorname{diag}(\Phi(\Lambda_X)) + \operatorname{diag}(\Phi(\Lambda_X))\iota_U^* + \mathcal{M}(\Lambda_X)[\bar{\delta}]$$

where we have used the notation introduced in Lemma 1. Using the expression (10) for ι_U and ι_V shows that the matrix $W = \iota_V \operatorname{diag}(\Phi(\Lambda_X)) + \operatorname{diag}(\Phi(\Lambda_X))\iota_U^*$ is computed as

$$W_{i,j} = \frac{1}{\Lambda_j^2 - \Lambda_i^2} \left(\varphi_j(\Lambda_j \bar{\delta}_{i,j} + \Lambda_i \bar{\delta}_{j,i}) - \varphi_i(\Lambda_i \bar{\delta}_{i,j} + \Lambda_j \bar{\delta}_{j,i}) \right)$$

where $\varphi = \Phi(\Lambda)$. Rearranging this expression using the symmetric and anti-symmetric parts shows the desired formula. □

References

Candès, E. J. and Recht, B. Exact matrix completion via convex optimization. *Found. Comput. Math.*, 9(6):717–772, 2009.

Candès, E. J., Sing-Long, C. A., and Trzasko, J. D. Unbiased risk estimates for singular value thresholding and spectral estimators. Technical Report arXiv:1210.4139v1, October 2012.

Combettes, P. L. and Wajs, V. R. Signal recovery by proximal forward-backward splitting. *Math. Mod. Sim.*, 4(4):1168, 2005.

Deledalle, C., Vaiter, S., Peyré, G., Fadili, J., and Dossal, C. Proximal splitting derivatives for risk estimation. In *Journal of Physics: Conference Series*, volume 386, pp. 012003. IOP Publishing, 2012.

Edelman, A. Matrix jacobians with wedge products. *MIT Handout for 18.325*, 2005.

Fazel, M. *Matrix Rank Minimization with Applications*. PhD thesis, Stanford University, 2002.

Lewis, A.S. The convex analysis of unitarily invariant matrix functions. *Journal of Convex Analysis*, 2(1/2):173–183, 1995.

Lewis, A.S. and Sendov, H.S. Twice differentiable spectral functions. *SIAM Journal on Matrix Analysis on Matrix Analysis and Applications*, 23:368–386, 2001.

Rockafellar, R. Tyrrell and Wets, Roger J-B. *Variational Analysis*. Fundamental Principles of Mathematical Sciences. Berlin: Springer-Verlag, third corrected printing edition, 2005.

Stein, C.M. Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6):1135–1151, 1981.

Sun, D. and Sun, J. Nonsmooth matrix valued functions defined by singular values. Technical report, Department of Decision Sciences, National University of Singapore, 2003.

Vonesch, C., Ramani, S., and Unser, M. Recursive risk estimation for non-linear image deconvolution with a wavelet-domain sparsity constraint. In *ICIP*, pp. 665–668. IEEE, 2008.