Supplementary Material to "Anomaly Detection in Images with Smooth Background Via Smooth-Sparse Decomposition"

Hao Yan, Kamran Paynabar, Jianjun Shi

September 24, 2015

Appendix A

Proposition 1. *If B_a is orthogonal, in iteration k, the subproblem $\theta_a^{(k)} = \text{argmin}_{\theta_S} \|y - B\theta^{(k)} - B_a\theta_a\|^2 + \gamma\|\theta_a\|_1$ has a closed-form solution in the form of $\theta_a^{(k)} = S_{\frac{\gamma}{2}}(B_a^T(y - B\theta^{(k)}))$, in which $S_\gamma(x) = \text{sgn}(x)(|x| - \gamma)_+$ is the soft-thresholding operator, and $\text{sgn}(x)$ is the sign function and $x_+ = \max(x, 0)$.*

Proof. If B_a is orthogonal, in each iteration k, we solve $\theta_a^{(k)} = \text{argmin}_{\theta_S} \|y - B\theta^{(k)} - B_a\theta_a\|^2 + \gamma\|\theta_a\|_1$. The first Karush–Kuhn–Tucker (KKT) condition of this optimization problem can be expressed as: $\nabla\|y - B\theta^{(k)} - B_a\theta_a\|^2 + \gamma g = 0$, where ∇ is the gradient operator and $g = [g_i] = \begin{cases} \text{sgn}(\theta_{ai}) & \theta_{ai} \neq 0 \\ [-1, 1] & \theta_{ai} = 0 \end{cases}$. The square is $\|y - B\theta^{(k)} - B_a\theta_a\|^2 = \theta_a^T B_a^T B_a\theta_a - 2\theta_a^T B_a^T(y - B\theta^{(k)}) + \|y - B\theta^{(k)}\|^2$. Since $B_a^T B_a = I$, the loss function can be simplified to $\|y - B\theta^{(k)} - B_a\theta_a\|^2 = \theta_a^T\theta_a - 2\theta_a^T B_a^T(y - B\theta^{(k)}) + \|y - B\theta^{(k)}\|^2$. Consequently, after simplification, the KKT condition gives $\theta_a = B_a^T(y - B\theta^{(k)}) - \frac{\gamma}{2}g$. We consider two cases for this solution, if $\theta_{ai} \neq 0$, then $\theta_{ai} + \frac{\gamma}{2}\text{sgn}(\theta_{ai}) = B_S^T(y - B\theta^{(k)})$. If $\theta_{ai} = 0$, then $B_a^T(y - B\theta^{(k)}) = \frac{\gamma}{2}g \in [-\frac{\gamma}{2}, \frac{\gamma}{2}]$. The solution can be given in a compact form of $\theta_a^{(k)} = \text{sgn}(B_a^T(y - B\theta^{(k)}))(\left|B_a^T(y - B\theta^{(k)})\right| - \frac{\gamma}{2})_+$, which is a soft-thresholding operator denoted by $S_{\frac{\gamma}{2}}(B_a^T(y - B\theta^{(k)}))$. □

Appendix B

Proposition 2. *The BCD algorithm attains the global optimum of the SSD loss function in (1).*

$$\underset{\theta,\theta_a}{argmin}\ \|e\|^2 + \lambda\theta^T R\theta + \gamma\|\theta_a\|_1, \text{ subject to. } y = B\theta + B_a\theta_a + e \tag{1}$$

Proof. (Tseng, 2001) in page 484, Theorem 5.1 proved that if an objective function f can be decomposed into the sum of a continuous function f_0 and some non-differentiable functions $f_i = 1, \cdots, N$, with some basic continuity assumptions on f_0, the BCD algorithm guarantees to attain a local optimum. It is clear that the SSD objective function in (1) is comprised of a continuous function $\|e\|^2 + \lambda\theta' R\theta$ and a non-differentiable penalty term $\gamma\|\theta_a\|_1$. Consequently, the BCD algorithm converges to a local optimum. In addition, since problem (1) is convex, the attained optimum is the global optimum. □

Appendix C:

Proposition 3. *The SSD problem in (1) is equivalent to a weighted LASSO problem in the form of*

$$\arg\min_{\theta_a} F(\theta_a) = (y - B_a\theta_a)^T(I - H)(y - B_a\theta_a) + \gamma\|\theta_a\|_1 \tag{2}$$

with $H = B(B^TB + \lambda R)^{-1}B^T$.

Proof. We first solve (1) for θ by fixing θ_a. That is $\hat{\theta} = argmin_\theta \|y-B\theta-B_a\theta_a\|^2+\lambda\theta^T R\theta+\gamma\|\theta_a\|_1$, which can be solved via $\hat{\theta} = (B^TB+\lambda R)^{-1}B^T(y-B_a\theta_a)$. Thus, it can be written that $B\hat{\theta} = B(B^TB+\lambda R)^{-1}B^T(y-B_a\theta_a) = H(y-B_a\theta_a)$. By plugging in this into (1), we have $\hat{\theta} = argmin_\theta \|y-H(y-B_a\theta_a)-B_a\theta_a\|^2+\lambda(y-B_a\theta_a)^T H^T RH(y-B_a\theta_a)+\gamma\|\theta_a\|_1$. After simplification and since $(I-H)^2+\lambda BK_\lambda^{-1}RK_\lambda^{-1}B^T(y-B_a\theta_a) = I-H$, where $K_\lambda = B^TB+\lambda R$, we can show that $\|y-B\theta-B_a\theta_a\|^2+\lambda\theta^T R\theta+\gamma\|\theta_a\|_1 = (y-B_a\theta_a)^T(I-H)(y-B_a\theta_a)+\gamma\|\theta_a\|_1$, which is the weighted LASSO formulation. □

Appendix D:

Claim 4. The $f(\theta_a) = (y-B_a\theta_a)^T(I-H)(y-B_a\theta_a)$. is convex for θ_a.

Proof. $f(\theta_a) = (y-B_a\theta_a)^T(I-H)(y-B_a\theta_a)$. To prove $f(\theta_a)$ is convex, we only need to show that $I-H$ is a positive semi-definite matrix. From Appendix C, it is given that $I-H = (I-H)^2+\lambda B(B^TB+\lambda R)^{-1}R(B^TB+\lambda R)^{-1}B^T$. Clearly, the first term $(I-H)^2$ is a positive semi-definite matrix. For the second term, since R is a positive semi-definite matrix, $B(B^TB+\lambda R)^{-1}R(B^TB+\lambda R)^{-1}B^T$ is also positive semi-definite. Consequently, $f(\theta_a)$ is a convex function. □

Appendix E:

Claim 5. $f(\cdot)$ is Lipschitz continuos, in which satisfies$\|\nabla f(\alpha)-\nabla f(\beta)\| \le L\|\alpha-\beta\|$ for any $\alpha,\beta \in R$ with $L = 2\|B_a\|_2^2$

Proof. We first show that H is positive semidefinite matrix. $H = B(B^TB+\lambda R)^{-1}B^T$. Since $B^TB+\lambda R$ is positive definite matrix, $(B^TB+\lambda R)^{-1}$ is also positive definite matrix, and H is positive semi-definite matrix.

We then prove that $\|I-H\|_2 \le 1$. Notice that $\|X\|_2$ refers to the spectrum norm of matrix X. This is because that $\|I-H\|_2 = \sqrt{\lambda_{max}[(I-H)^2]} = \lambda_{max}(I-H) = 1-\lambda_{min}(H) \le 1$. The last equation hold because $\lambda_{min}(H) \ge 0$ since H is positive semi-definite matrix. Note that $\lambda_{max}(X)$ refers to the largest eigenvalue of matrix$X$ and $\lambda_{min}(X)$ refers to the smallest eigenvalue of matrix X.

Consequently, $\nabla f(\alpha) = \nabla(y-B_a\alpha)^T(I-H)(y-B_a\alpha) = 2B_a^T(I-H)(B_a\alpha-y)$. $\|\nabla f(\alpha)-\nabla f(\beta)\| = \|2B_a^T(I-H)B_a(\alpha-\beta)\| \le \|2B_a^T(I-H)B_a\|_2 \cdot \|\alpha-\beta\| \le L\|\alpha-\beta\|$, in which $L = 2\|B_a\|_2^2$

The last equation holds because $\|2B_a^T(I-H)B_a\|_2 \le \|2B_a^T\|_2\|(I-H)\|_2\|B_a\|_2 \le \|2B_a^T\|_2\|B_a\|_2 = 2\|B_a\|_2^2$. □

Appendix F:

Proposition 6. *The proximal gradient method for the SSD problem in (1), given by* $\theta_a^{(k)} = argmin_{\theta_a}\{f(\theta_a^{(k-1)})+\left\langle \theta_a-\theta_a^{(k-1)}, \nabla f(\theta_a^{(k-1)})\right\rangle + \frac{L}{2}\|\theta_a-\theta_a^{(k-1)}\|^2+\gamma\|\theta_a\|_1\}$*, has a closed-form solution in each iteration* k*, in the form of a soft-thresholding function as follows:*

$$\theta_a^{(k)} = S_{\frac{\gamma}{L}}(\theta_a^{(k-1)} + \frac{2}{L}B_a^T(y-B_a\theta_a^{(k-1)}-\mu^{(k)})) \tag{3}$$

with $L = 2\|B_a\|_2^2$.

Proof. Since $\nabla f(\theta_a^{(k-1)}) = 2B_a^TB_a\theta_a^{(k-1)} - 2B_a^T(y-B\theta^{(k)})$, in each iteration given $\theta_a^{(k-1)}$, the $\theta_a^{(k)} = argmin_{\theta_a}\|\theta_a-\theta_a^{(k-1)}-\frac{2}{L}B_a^T(y-B\theta^{(k)}-B_a\theta_a^{(k-1)})\|^2+\gamma\|\theta_a\|_1\}\}$. Thus, from the result of Appendix A, it is straightforward to show that this problem can be solved using a soft thresholding operator in the form of $\theta_a^{(k)} = S_{\frac{\gamma}{L}}(\theta_a^{(k-1)}+\frac{2}{L}B_a^T(y-B_a\theta_a^{(k-1)}-\mu^{(k)}))$. □

Appendix G:

Claim 7. Suppose the Cholesky decomposition of $B_i^T B_i$ is given as $B_i^T B_i = Z_i Z_i^T$, the eigen decomposition $Z_i^{-1} D_i^T D_i (Z_i^{-1})^T$ is $U_i \text{diag}(s_i) U_i^T$ and $V_i = B_i (Z_i^{-1})^T U_i$. It can be shown that $H_i(\lambda) = V_i^T \text{diag}(\frac{1}{1+\lambda s_1}, \cdots, \frac{1}{1+\lambda s_n}) V_i$, and its trace is given by $\text{tr}(H_i) = \sum_{i=1}^n \frac{1}{1+\lambda s_i}$

Proof. The proof of the first part is given below:

$$\begin{aligned} H_i(\lambda) &= B_i(B_i^T B_i + \lambda D_i^T D_i)^{-1} B_i^T = B_i(Z_i Z_i^T + \lambda D_i^T D_i)^{-1} B_i^T \\ &= B_i (Z_i^{-1})^T (I + \lambda Z_i^{-1} D_i^T D_i (Z_i^{-1})^T)^{-1} (Z_i^{-1}) B_i^T \\ &= B_i (Z_i^{-1})^T (I + \lambda U_i \text{diag(s}_\text{i}) U_i^T)^{-1} (Z_i^{-1}) B_i^T \\ &= B_i (Z_i^{-1})^T U_i (I + \lambda \text{diag(s}_\text{i}))^{-1} U_i^T (Z_i^{-1}) B_i^T \\ &= V_i (I + \lambda \text{diag(s}_\text{i}))^{-1} V_i^T \\ &= V_i^T \text{diag}(\frac{1}{1+\lambda s_1}, \cdots, \frac{1}{1+\lambda s_n}) V_i \end{aligned}$$

To compute the trace of H_i, we first show that $V_i^T V_i = U_i^T Z_i^{-1} B_i^T B_i (Z_i^{-1})^T U_i = U_i^T U_i = I$. Thus the trace of H_i becomes $\text{tr}(H_i) = \text{tr}(V_i (I + \lambda \text{diag(s}_\text{i}))^{-1} V_i^T) = \text{tr}(V_i^T V_i (I + \lambda \text{diag(s}_\text{i}))^{-1}) = \text{tr}((I + \lambda \text{diag(s}_\text{i}))^{-1}) = \sum_{i=1}^n \frac{1}{1+\lambda s_i}$ □

Appendix H:

"In this appendix, we applied the extended-maxima transformationmethod to the simulated images with line anomalies, clustered anomalies and scattered anomalies. The detection results are reportedin Figure . Moreover, the FPR, FNR, and computational time for all the benchmark methods are reported in Table 1. "

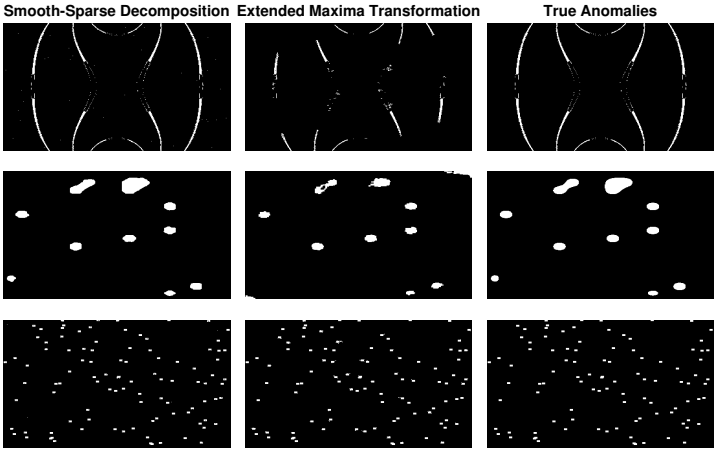

Figure 1: Anomalies detection comparison result for SSD and extended maxima transformation when $\delta = 3$

References

Tseng, P. (2001). Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications*, 109(3):475–494.

Table 1: FPR, FNR, and computation time for line , clustered and scattered anomalies with $\delta = 0.1, 0.2, 0.3$

δ	Defect Type	Criterion	SSD	Edge	Jump	Local	Global	Maxima
0.1	Line	FPR	0.108	0.012	0.022	0.066	0.202	0.045
		FNR	0.234	0.989	0.908	0.492	0.591	0.791
	Clustered	FPR	0.016	0.0003	0.086	0.539	0.211	0.008
		FNR	0.035	0.979	0.837	0.756	0.799	0.868
	Scattered	FPR	0.011	0.008	0.179	0.019	0.204	0.018
		FNR	0.076	0.858	0.722	0.567	0.752	0.984
0.2	Line	FPR	0.027	0.016	0.037	0.058	0.202	0.005
		FNR	0.021	0.900	0.126	0.181	0.507	0.792
	Clustered	FPR	0.017	0.0003	0.083	0.052	0.213	0.002
		FNR	0.005	0.89	0.127	0.462	0.673	0.657
	Scattered	FPR	0.0114	0.005	0.138	0.02	0.203	0.004
		FNR	0.0153	0.293	0.108	0.251	0.595	0.038
0.3	Line	FPR	0.001	0.015	0.035	0.054	0.195	0.001
		FNR	0.003	0.783	0.111	0.063	0.456	0.557
	Clustered	FPR	0.018	0.001	0.081	0.046	0.211	0.007
		FNR	0.001	0.754	0.054	0.289	0.572	0.268
	Scattered	FPR	0.012	0.003	0.11	0.02	0.203	0.001
		FNR	0.007	0.257	0.063	0.087	0.407	0.012
Computational Time			0.19s	0.667s	38.43s	0.043s	0.048s	0.039s

'SSD' for Smooth Sparse Decomposition, 'Edge' for edge detection, 'Jump' for jump regression, 'Local' for local thresholding, 'Global' for global thresholding, and 'Maxima' for extented maxima transformation.