Multi-Task Learning for Subspace Segmentation: Supplementary File

Yu Wang * YW323@CAM.AC.UK
David Wipf † DAVIDWIP@MICROSOFT.COM
Qing Ling ♯ QINGLING@MAIL.USTC.EDU.CN
Wei Chen * WC253@CAM.AC.UK
Ian Wassell * IJW24@CAM.AC.UK
∗ Computer Laboratory, University of Cambridge, Cambridge, UK
† Microsoft Research, Beijing, China
♯ University of Science and Technology of China, Hefei, Anhui, China

1. Multi-Task Subspace Clustering (MTSC) Algorithm Summary

Throughout this supplementary file, equations from the main paper will be prefixed by an 'M', e.g., (M.12) would denote equation (12) from the main paper. Regarding MTSC, the input is the $D \times N$ data matrix $\boldsymbol{X}$. We must first inialize the hyperparameter matrices $\boldsymbol{\Lambda}$ and $\boldsymbol{W}$ and choose some β sufficiently large. We then iterate the following updates:

$$
\begin{aligned}
\boldsymbol{\Gamma}_j &\leftarrow \left[\sum_i w_{ij}\boldsymbol{\Lambda}_i^{-1}\right]^{-1}, \\
\boldsymbol{u}_j &\leftarrow \text{diag}\left[\left(\sum_i w_{ij}\boldsymbol{\Lambda}_i^{-1} + \frac{1}{\nu}\bar{\boldsymbol{X}}_j^\top \bar{\boldsymbol{X}}_j\right)^{-1}\right], \\
\boldsymbol{z}_j &\leftarrow \boldsymbol{\Gamma}_j \bar{\boldsymbol{X}}_j^\top \left(\nu I + \bar{\boldsymbol{X}}_j \boldsymbol{\Gamma}_j \bar{\boldsymbol{X}}_j^\top\right)^{-1} \boldsymbol{x}_j, \\
\lambda_{ik} &\leftarrow \frac{\sum_j w_{kj}\left(z_{ij}^2 + u_{ij}\right)}{\sum_j w_{kj}}, \\
w_{kj} &\leftarrow \exp\left(\frac{1}{\beta}\left[-\left(\sum_i \frac{z_{ij}^2}{\lambda_{ik}} + \sum_i \log\lambda_{ik} + \sum_i \frac{u_{ij}}{\lambda_{ik}}\right)\right] - 1\right), \\
w_{ij} &\leftarrow \frac{w_{ij}}{\sum_i w_{ij}}.
\end{aligned}
\tag{1}
$$

Note that the index k above is used for convenience, and is unrelated to the subspace number which frequently uses the same index in the main text. These rules are guaranteed to reduce or leave unchanged (M.16) at every iteration by construction (as a majorization-minimization algorithm); however, there is admittedly no formal guarantee of convergence to a stationary point.

The underlying general derivations are based on the selection

$$\rho(\boldsymbol{W}) = \beta \sum_{ij} w_{ij} \log w_{ij}, \tag{2}$$

which we advocate in (Wang et al., 2015) as a convex function that leads to convenient iterations that closely resemble those from (Qi et al., 2008). However, the exact form of this function is likely not that important as long as it favors sharing of basis functions within the constraint set. Note that Lemma 2 is based upon the alternative selection $\rho(\boldsymbol{W}) = \beta\|\boldsymbol{W}\|_2$ which accomplishes more or less the same thing but is slightly easier to analyze.

2. Proof Sketches

Full proof details will be deferred to a subsequent journal publication; here we provide the basic high-level constructions of Lemma 1 and Lemma 2.

Proof of Lemma 1: Let $\boldsymbol{\Gamma}_j^*$ denote the value of $\boldsymbol{\Gamma}_j$ computed via (M.12) at any stationary point, and let $\boldsymbol{z}^*$ indicate the associated value of $\boldsymbol{z}_j$ computed via (M.13). In the limit as $\nu \to 0$, any stationary point of $\mathcal{L}(\boldsymbol{\Lambda}, \boldsymbol{W})$ will produce a point $\boldsymbol{z}_j^*$ feasible to $\boldsymbol{x}_j = \bar{\boldsymbol{X}}\boldsymbol{z}_j$.[1] If this were not the case, it is easily shown that $\mathcal{L}(\boldsymbol{\Lambda}, \boldsymbol{W})$ would be driven to infinity via straightforward extension of the analysis in (Wipf et al., 2011). Because of the independent subspace assumption and the left multiplication by $\boldsymbol{\Gamma}_j^*$ in (M.13), the only way that $\boldsymbol{z}_j^*$ can be feasible is if $\boldsymbol{\Gamma}_j^*$ has sufficient nonzero diagonal values aligned with points from the same subspace; additional nonzero values of $\boldsymbol{\Gamma}_j^*$ may have arbitrary posi-

[1] Note that (M.13) is still well-defined in the limit $\nu \to 0$ by using the appropriate pseudo-inverse.

tions.[2]

Now if $\{\boldsymbol{\Lambda}^*, \boldsymbol{W}^*\}$ is truly a stationary point of $\mathcal{L}(\boldsymbol{\Lambda}, \boldsymbol{W})$, then the remaining values of the corresponding $\boldsymbol{z}_j^*$ must be a stationary point of

$$\min_{\widetilde{\boldsymbol{z}}_j} \frac{1}{\nu}\|\boldsymbol{x}_j - \widetilde{\boldsymbol{X}}_j\widetilde{\boldsymbol{z}}_j\|_2^2 + \widetilde{\boldsymbol{z}}_j^\top \left(\widetilde{\boldsymbol{\Gamma}}_j^*\right)^{-1}\widetilde{\boldsymbol{z}}_j + C, \quad (3)$$

where C is some constant independent of $\boldsymbol{z}_j$, $\widetilde{\boldsymbol{\Gamma}}_j^*$ denotes the nonzero elements of $\boldsymbol{\Gamma}_j^*$, and $\widetilde{\boldsymbol{z}}_j$ and $\widetilde{\boldsymbol{X}}_j$ are the corresponding nonzero elements of $\boldsymbol{z}_j$ and columns of $\bar{\boldsymbol{X}}_j$ respectively. This occurs because (3), with the appropriate choice of C, represents a convex upper bound on $\mathcal{L}(\boldsymbol{\Lambda}, \boldsymbol{W})$, with equality iff $\boldsymbol{z}_j$ is given by (M.13).[3] When $\nu \to 0$, (3) reduces to

$$\min_{\widetilde{\boldsymbol{z}}_j} \widetilde{\boldsymbol{z}}_j^\top (\widetilde{\boldsymbol{\Gamma}}_j^*)^{-1}\widetilde{\boldsymbol{z}}_j \;\text{ s.t. } \boldsymbol{x}_j = \widetilde{\boldsymbol{X}}_j\widetilde{\boldsymbol{z}}_j. \quad (4)$$

This represents a weighted ℓ_2 norm penalty on $\widetilde{\boldsymbol{z}}_j$ being minimized over a restricted feasible set, which includes a sufficient number of samples within subspace $\mathcal{S}_k$ per the arguments above. It also represents a slightly modified version of the LSR objective function from (Lu et al., 2012) evaluated over a reduced feasible set. Moreover, the LSR algorithm has already been shown to produce an ideal block-sparse solution using a standard ℓ_2 norm penalty; however, the proof provided in (Lu et al., 2012) applies equally well with a generalized weighted norm. ■

Proof of Lemma 2: The basic strategy is to first show that the cost function in (M.15) satisfies the conditions of the theorem. Furthermore, we show that the optimal solution is such that each column of $\boldsymbol{W}^*$ has elements that are either zero or a constant value, and that columns of $\boldsymbol{\Lambda}^*$ associated with these nonzero elements are equal. Next, we note that the upper bound from (M.16) follows from the determinant identity

$$\log|\Sigma_{x_j}| \quad \equiv \quad (5)$$

$$\log\left|\sum_i w_{ij}\Lambda_i^{-1} + \frac{1}{\nu}\bar{\boldsymbol{X}}_j^\top\bar{\boldsymbol{X}}_j\right| - \log\left|\sum_i w_{ij}\Lambda_i^{-1}\right|,$$

and then application of Jensen's inequality via

$$\sum_i w_{ij}|\Lambda_i| \geq -\log\left|\sum_i w_{ij}\Lambda_i^{-1}\right|. \quad (6)$$

However, given the stipulated conditions on $\boldsymbol{W}^*$ and $\boldsymbol{\Lambda}^*$ from above, the above inequality collapses to a strict equality. Given that (M.16) is an upper bound, this ensures that $\boldsymbol{\Lambda}^*$ and $\boldsymbol{W}^*$ also globally optimize (M.16).

We now fill in some of the missing ingredients. The technical introduction of randomness into the generic subspace model definition ensures that the maximally sparse feasible $\boldsymbol{z}_j$ will have at most D_k nonzero elements with probability one (assuming $\boldsymbol{x}_j \in \mathcal{S}_k$). This follows from minor adaptation of Theorem 1 in (Baron et al.). Now if $\beta = 0$ then (M.15) decouples completely across tasks, and we may achieve the global optimum by simply setting $\boldsymbol{\Gamma}_j = \boldsymbol{\Lambda}_j$ for all j without loss of generality, and we may optimize each task j individually over $\boldsymbol{\Lambda}_j$. In the limit $\nu \to 0$, Theorem 4 in (Wipf et al., 2011) guarantees that a maximally sparse solution will be found for each $\boldsymbol{\Lambda}_j$, with the cost function dominated by an $O\left([D - D_k]\log\nu\right)$ factor which is unbounded from below. Additionally, the resulting $\boldsymbol{P}^{-1}\boldsymbol{\Gamma}_j^*$ will be block-sparse an aligned with the proper subspace, and likewise for $\boldsymbol{z}_j$ by virtue of (M.13).

Let $\Omega_k \subset \{1, \ldots, N\}$ denote the set of column indeces associated with $\boldsymbol{X}_k$. Now consider the optimal solution to (M.15) in the restricted case where $\boldsymbol{\Gamma}_j = \boldsymbol{\Gamma}_j'$ for all $j, j' \in \Omega_k$. Upon careful inspection however, we can show that this restriction only alters the objective function value by an inconsequential $O(1)$ factor independent of ν. Consequently, there is marginal advantage to individually optimizing each $\boldsymbol{\Gamma}_j$ within a subspace block, which ultimately leads to the desired grouping effect.

Now we reintroduce the $\boldsymbol{W}$ penalty factor by allowing $\beta > 0$. The overall objective function is still dominated by the $O\left([D - D_k]\log\nu\right)$ factor as long as each $\boldsymbol{\Gamma}_j$ maintains the proper block-sparsity; however, from above we know that there is a bounded advantage to optimizing each $\boldsymbol{\Gamma}_j$ individually via $\boldsymbol{\Lambda}$ and $\boldsymbol{W}$. In contrast, as we make β large, there is an increasing incentive to minimize $\|\boldsymbol{W}\|_2$. Within the constraint set, and the additional block-sparse restriction on each $\boldsymbol{\Gamma}_j$ which must be maintained at any optimum, $\|\boldsymbol{W}\|_2$ will achieve its minimum $\boldsymbol{W}^*$ when each column $\boldsymbol{w}_j$ is populated by either zero or some C_j such that $\sum_i w_{ij} = 1$. To seen this, consider any other $\boldsymbol{W}'$. To achieve the global optimum of (M.15), we must have $w_{ij}' = 0$ whenever $\boldsymbol{\Lambda}_i$ does not match the proper subspace-aligned block-sparsity. It follows then that $\|\boldsymbol{W}'\|_2 > \|\boldsymbol{W}^*\|_2$, and as β grows this gap can become arbitrarily wide. However, we know from above that we cannot compensate for this increase by modulating the magnitudes of each individual $\boldsymbol{\Gamma}_j$. Hence $\boldsymbol{W}'$ cannot be optimal.

All of this implies that for sufficiently large β, a common $\boldsymbol{\Gamma}_j^*$ across all $j \in \Omega_k$ optimizes (M.15) and likewise (M.16). The remainder of the theorem directly follows from related arguments surrounding the effectiveness

[2] For convenience, if the ℓ-th diagonal element of $\boldsymbol{\Lambda}_i$ is equal to zero with associated weight $w_{ij} \neq 0$, then we simply define $(\boldsymbol{\Gamma}_j)_{\ell\ell} = 0$. This avoids the singularity of dividing by zero and satisfies our present purposes.

[3] The generic form of this bound is $\boldsymbol{y}^\top\left(\boldsymbol{A}\boldsymbol{\Gamma}\boldsymbol{A}^\top + \nu\boldsymbol{I}\right)^{-1}\boldsymbol{y} \leq 1/\nu\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \boldsymbol{x}^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{x}$ for all $\boldsymbol{x}$

of (M.8). Obviously there are some gaps in the above derivation, but we prefer to leave a detailed treatment for a subsequent journal publication along with more extensive simulation results. ■

References

Baron, D., Duarte, M. F., and Wakin, M. B. Distributed compressive sensing. *arXiv:0901.3403v1.4729v2*.

Lu, C., Min, H., Zhao, Z., Zhu, L., Huang, D., and Yan, S. Robust and efficient subspace segmentation via least squares regression. In *ECCV*, 2012.

Qi, Y., Liu, D., Dunson, D., and Carin, L. Multi-task compressive sensing with Dirichlet process priors. In *ICML*, 2008.

Wang, Y., Wipf, D., Yun, J. M., Chen, W., and Wassell, I. J. Clustered sparse Bayesian learning. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2015.

Wipf, D. P., Rao, B. D., and Nagarajan, S. Latent variable Bayesian models for promoting sparsity. *IEEE Transactions on Information Theory*, 57(9):6236–6255, 2011.