Supplementary Material for: Multimodal Sparse Bayesian Dictionary Learning

Igor Fedorov, Bhaskar D. Rao

I. Incremental EM

One stochastic inference alternative is called Incremental EM, which we review next for the case of $J = 1$, omitting modality subscripts for brevity [1]. Let $F(\tilde{p}, \theta) = E_{\tilde{p}} \left[\log p(X, Y, \theta)\right] + H(\tilde{p})$, where $H(\tilde{p})$ is the entropy of $\tilde{p}(\cdot)$. It can be shown that $p(X|Y, \theta)$ is the unique maximizer of $F(\tilde{p}, \theta)$, given θ. It can also be shown that $F(\tilde{p}, \theta) = \log p(Y|\theta)$ for $\tilde{p}(X) = p(X|Y, \theta)$. It then follows that the E and M steps of EM can be re-stated in the following form:

$$\tilde{p}^{t+1} = \arg\max_{\tilde{p}} F(\tilde{p}, \theta^t) \tag{48}$$

$$\theta^{t+1} = \arg\max_{\theta} F(\tilde{p}^{t+1}, \theta). \tag{49}$$

When the posterior factors over the data points in the dataset, it is reasonable to consider only distributions of the form $\tilde{p}(X) = \prod_{i\in[L]} \tilde{p}^i(x^i)$ in (48). Although the factorization constraint may seem restrictive, it should be noted that the maximizer of $F(\tilde{p}, \theta)$ must also factor [1] [1]. It then follows that $F(\tilde{p}, \theta^t) = \sum_{i\in[L]} F^i(\tilde{p}^i, \theta^t)$. This leads to a class of algorithms which perform the E-step in an incremental fashion by modifying (48) to

$$\left(\tilde{p}^i\right)^{t+1} = \begin{cases} \arg\max_{\tilde{p}_i} F^i(\tilde{p}^i, \theta^t) \text{ if } i \in \phi \\ \left(\tilde{p}^i\right)^t \text{ else.} \end{cases} \tag{50}$$

II. Computation of (21)

It can be shown that $\partial \log p\left(Y_j|\theta^t, \sigma_j^t\right)/\partial\sigma_j^t$ is given by

$$-\sum_{i\in[L]} \left(y_j^i\right)^T V_j^i \left(\Lambda_j^i\right)^{-2} \left(V_j^i\right)^T y_j^i + \sum_{n\in[N_j]} \left(\Lambda_j[n, n] + \sigma_j^2\right)^{-1}$$

where $V_j^i \Lambda_j^i \left(V_j^i\right)^T$ is the eigen-decomposition of $\Sigma_{y,j}^i$ and $\left(\Lambda_j^i\right)^{-2}$ represents a diagonal matrix whose $[n, n]$'th entry is $\left(\Lambda_j^i[n, n]\right)^{-2}$.

III. Complete TD-MSBDL Algorithm

The complete TD-MSBDL algorithm is shown in Fig. 11.

IV. Taxonomy

The taxonomy of MSBDL algorithms is visualized in Fig. 12.

```
Require: $\boldsymbol{Y}, \boldsymbol{Y^V}, \boldsymbol{\sigma^0}, \boldsymbol{\beta^0}, \sigma^\infty, \beta^\infty, H, H^V, \alpha_\sigma, \alpha_\beta, T^V$
while $\boldsymbol{\beta}$ not converged do
    while $\boldsymbol{D}$ and $\boldsymbol{W}$ not converged do
        for $i \in [L]$ do
            Update $\boldsymbol{\Sigma_x^{TD,i}}$ using (36)
            Update $\boldsymbol{\mu^{TD,i}}$ using (37)
            Update $\gamma^i$ using (15)
        end for
        $\{$ Update $D_j$ using (16) if $\sigma_j$ not converged$\}_{j\in[J]}$
        $\{$ Update $W_j$ using (42) if $\beta_j$ not converged$\}_{j\in[J]}$
    end while
    $\{$Update $\sigma_j$ using (21) if $\sigma_j$ not converged$\}_{j\in[J]}$
    if modulo$(t, T^V) = 0$ then
        for $j \in [J]$ do
            if $\beta_j$ not converged then
                $\{\gamma^{*,V,i}\}_{i\in L^V} = MSBDL\left(Y_j^V, \sigma_j^0, \sigma_j, \sigma^\infty, \alpha_\sigma\right)$
                Update $\beta_j$ using (45)
            end if
        end for
    end if
end while
```

Fig. 11: Complete TD-MSBDL algorithm for the prior in (8).

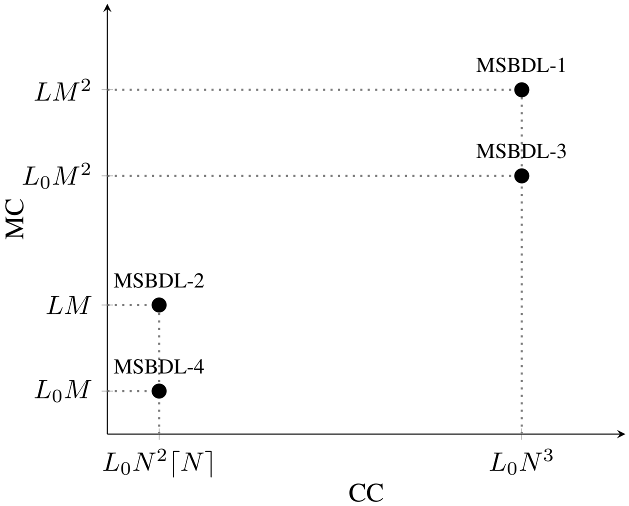


Fig. 12: Visualization of worst-case computational and memory complexity per modality and EM iteration of the proposed approaches.

The authors are with the Department of Electrical and Computer Engineering, University of California, San Diego, 9500 Gilman Dr, San Diego, CA, 92103

[1] In other words, if $\{p^*, \theta^*\}$ is the maximizer of $F(\cdot, \cdot)$, then p^* factors.

V. Marginal Priors for Models Introduced in Section V

.

A. Atom-to-Subspace Sparsity Prior

The marginal prior on $\boldsymbol{x}$ under $\gamma_B[k] \sim \mathsf{IGa}\left(\frac{\tau}{2}, \frac{\tau}{2}\right)$ takes the form:

$$p\left(\boldsymbol{x}\right) = \prod_{k \in [K]} ST\left(\left\| \begin{bmatrix} x_1[k] \\ x_2\left[\mathscr{T}^k\right] \end{bmatrix} \right\|_2^2 ; \tau\right) \tag{51}$$

where $x_2\left[\mathscr{T}^k\right]$ is shorthand for the the elements of x_2 indexed by $\mathscr{T}^k$ and $ST(\cdot)$ denotes the Student's-t distribution.

B. Hierarchical Sparsity Prior

Marginalizing the conditional prior in Section V-B over $\boldsymbol{\gamma_j}$ for $\gamma_j[m] \sim \mathsf{IGa}\left(\frac{\tau}{2}, \frac{\tau}{2}\right)$, the prior on $\boldsymbol{x}$ turns out to be

$$p\left(\boldsymbol{x}\right) = \prod_{k \in [K], m \in \mathscr{T}^k} ST\left(\left\| \begin{bmatrix} x_1[k] \\ x_2\left[m\right] \end{bmatrix} \right\|_2^2 ; \tau\right) ST\left(x_2[m]^2 ; \tau\right). \tag{52}$$

VI. Additional Results for Synthetic Data Experiments

Fig. 13 shows histograms of results for the synthetic data experiments in Section VIII under the atom-to-subspace prior. Fig. 14 shows histograms of results for the synthetic data experiments in Section VIII under the hierarchical sparsity prior.

VII. Proof of Theorem 1

By definition, the log-likelihood function is coercive if

$$\lim_{\|\{\theta, \boldsymbol{\sigma}\}\| \to \infty} -\log p\left(\boldsymbol{Y} | \theta, \boldsymbol{\sigma}\right) = \infty \tag{53}$$

where we define the norm of $\{\theta, \boldsymbol{\sigma}\}$ to be

$$\|\{\theta, \boldsymbol{\sigma}\}\| = \sqrt{\sum_{m \in [M], j \in [J]} \left\|d_j^m\right\|_2^2 + \sum_{i \in [L]} \|\gamma^i\|_2^2 + \sum_{j \in [J]} \sigma_j^2}. \tag{54}$$

The negative log-likelihood can be written as

$$-\log p\left(Y | \theta, \boldsymbol{\sigma}\right) \doteq \sum_{i \in [L], j \in [J]} \left(y_j^i\right)^T \left(\Sigma_{y,j}^i\right)^{-1} y_j^i + \log\left|\Sigma_{y,j}^i\right| \tag{55}$$

where $\doteq$ refers to dropping terms which do not depend on θ or $\boldsymbol{\sigma}$.

Next, we establish several results about $\Sigma_{y,j}^i$, which is defined in (11). Let $\left(\Gamma^i\right)^{0.5}$ be a diagonal matrix whose $[m, m]$'th entry is given by $\left(\gamma^i[m]\right)^{0.5}$. Then, $D_j \Gamma^i D_j^T$ is the Gramian matrix of $\left(\Gamma^i\right)^{0.5} D_j^T$. Since Gramian matrices are positive semi-definite (PSD), $D_j \Gamma D_j^T$ must be PSD. Since $\sigma_j^2 \mathsf{I}$ is PSD, $\Sigma_{y,j}^i$ is PSD. Finally, since $\Sigma_{y,j}^i$ is PSD, then $\left(\Sigma_{y,j}^i\right)^{-1}$ is also PSD. Therefore,

$$\left(y_j^i\right)^T \left(\Sigma_{y,j}^i\right)^{-1} y_j^i \geq 0 \tag{56}$$

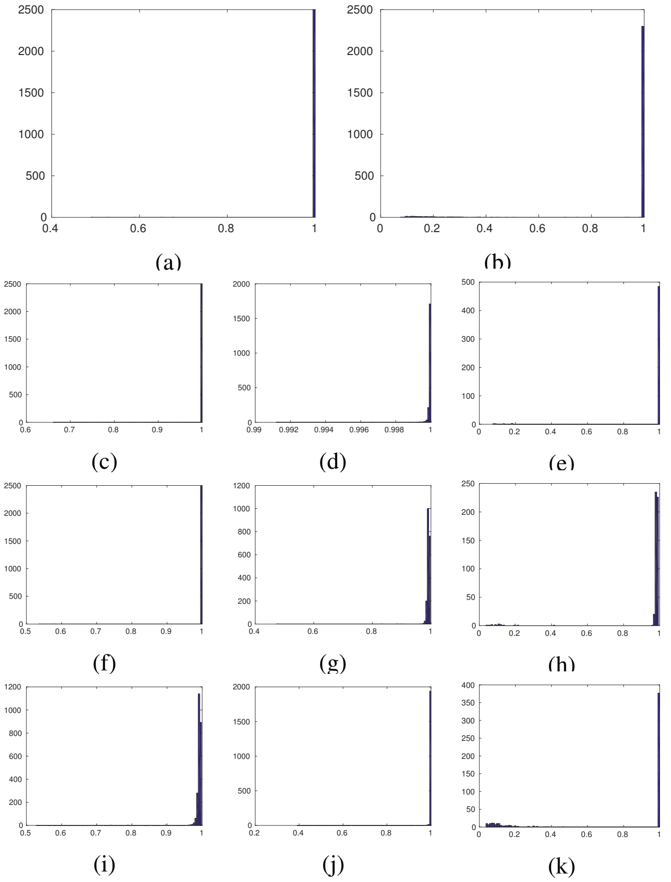


Fig. 13: Histograms of dictionary recovery results for the atom-to-subspace model and test cases in Table II. (Fig.'s 13a, 13c, 13f, 13i): distribution of $\iota\left(D_1[:,m], \hat{D}_1\right)$ $\forall m$ for cases A-D, respectively. (Fig.'s 13d, 13g, 13j): distribution of $\iota\left(D_2[:,k], \hat{D}_2\right)$ $\forall k : |\mathscr{T}^k| = 1$ for cases B-D, respectively. (Fig.'s 13b, 13e, 13h, 13k): distribution of $\iota\left(D_2[:,k], \hat{D}_2\right)$ $\forall k : |\mathscr{T}^k| > 1$ for cases A-D, respectively.

in general.

Turning to the second term in (55), we can re-write it as

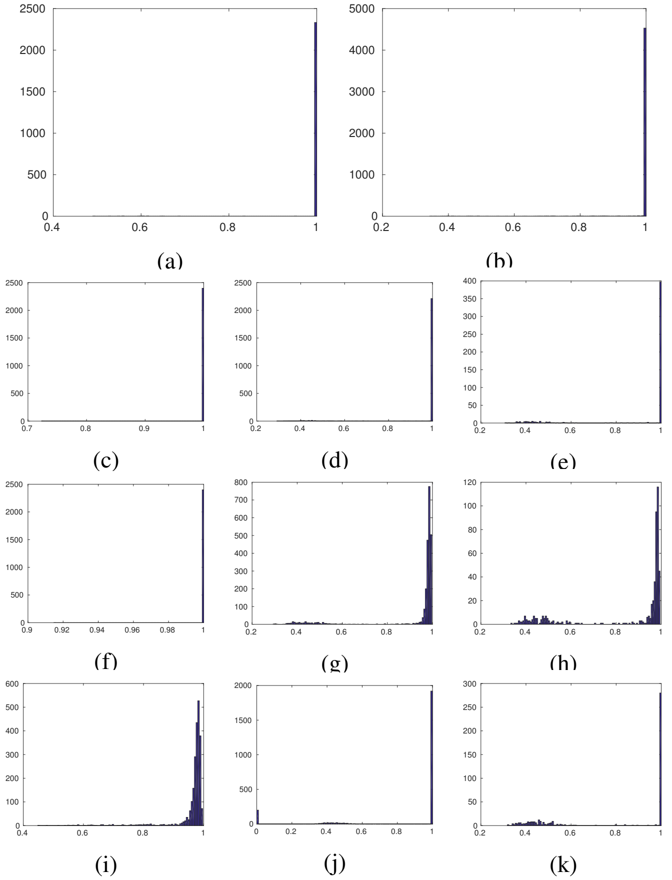
$$\log|\Sigma_{y,j}^i| = \sum_{n \in [N_j]} \log\left(\sigma_j^2 + \lambda_j^i[n]\right) \tag{57}$$

where $\lambda_j^i[n] \geq 0$ denotes the n'th eigenvalue of $D_j \Gamma^i D_j^T$. Combining (59) with (55), we have that

$$-\log p\left(\boldsymbol{Y} | \theta, \boldsymbol{\sigma}\right) \geq \sum_{j \in [J], i \in [L], n \in [N_j]} \log\left(\sigma_j^2 + \lambda_j^i[n]\right). \tag{58}$$

Note that $\sigma_j^2 + \lambda_j^i[n] > 0$ for all j, i, n, such that the right hand side of (58) tends to ∞ if $\sigma_{j^*}^2 + \lambda_{j^*}^{i^*}[n^*]$ tends to ∞ for some j^*, i^*, n^*.

In order for $\|\{\theta, \boldsymbol{\sigma}\}\| \to \infty$, one or more of the terms under the square root in (54) must also approach infinity. Starting with $\boldsymbol{\sigma}$, we observe that in order for $\sum_{j \in [J]} \sigma_j^2 \to \infty$, there must be at least one j^* such that $\sigma_{j^*} \to \infty$. Since $\lambda_j^i[n] \geq 0$



(a) (b) (c) (d) (e) (f) (g) (h) (i) (j) (k)

Fig. 14: Histograms of dictionary recovery results for the atom-to-subspace model and test cases in Table III. (Fig.'s 14a, 14c, 14f, 14i): distribution of $\iota\left(D_1[:,m],\hat{D}_1\right)$ $\forall m$ for cases A-D, respectively. (Fig.'s 14d, 14g, 14j): distribution of $\iota\left(D_2[:,k],\hat{D}_2\right)$ $\forall k : |\mathscr{T}^k| = 1$ for cases B-D, respectively. (Fig.'s 14b, 14e, 14h, 14k): distribution of $\iota\left(D_2[:,m],\hat{D}_2\right)$ $\forall m \in T^k : |\mathscr{T}^k| > 1$ for cases A-D, respectively.

for all j, i, n, at least one of the terms in the right hand side of (58) must tend to ∞, leading to the result

$$\lim_{\sum_{j\in[J]}\sigma_j^2\to\infty} -\log p\left(\boldsymbol{Y}|\theta,\boldsymbol{\sigma}\right) = \infty.$$

Turning to γ, we observe that in order for $\sum_{i\in[L]}\left\|\gamma^i\right\|_2^2 \to \infty$, there must exist at least one i^* and m^* such that $\gamma^{i^*}[m^*] \to \infty$. We also know that

$$\begin{aligned}
\sum_{j\in[J],n\in[N_j]} \lambda_j^i[n] &= \sum_{j\in[J]} trace\left(D_j\Gamma^i D_j^T\right) && (59)\\
&= \sum_{j\in[J],n\in[N_j],m\in[M]} \gamma^i[m]\left(d_j^m[n]\right)^2 && (60)\\
&= \sum_{j\in[J],m\in[M]} \gamma^i[m]\left\|d_j^m\right\|_2^2 && (61)\\
&\geq \gamma^{i^*}[m^*]\left\|d_{j^*}^{m^*}\right\|_2^2 && (62)\\
&=^a \infty. && (63)
\end{aligned}$$

where step (a) follows from the assumption that for each m, there exists a j^* such that $\left\|d_{j^*}^m\right\|_2^2 > 0$. Since N_j, J are finite, (63) implies that there must exist j', n^* such that $\lambda_{j'}^i[n^*] \to \infty$. In addition, since $\sigma_j^2 > 0$ for all j, $\lambda_{j'}^i[n^*] \to \infty$ implies that at least one of the terms in the right hand side of (58) tends to ∞, leading to the result

$$\lim_{\sum_{i\in[L]}\|\gamma^i\|_2^2\to\infty} -\log p\left(\boldsymbol{Y}|\theta,\boldsymbol{\sigma}\right) = \infty$$

Finally, in order for $\sum_{j\in[J],m\in[M]}\left\|d_j^m\right\|_2^2 \to \infty$, there must exist at least one j^* and m^* such that $\left\|d_{j^*}^{m^*}\right\|_2^2 \to \infty$. We can apply the same argument as in (59)-(63) to conclude that $\left\|d_{j^*}^{m^*}\right\|_2^2 \to \infty$ implies that there exists j', n^* such that $\lambda_{j'}^i[n^*] \to \infty$, where in this case step (a) in (63) follows from the assumption that at least one of $\left\{\gamma^i[m]\right\}_{i\in[L]}$ is non-zero for all m. As a result, we have:

$$\lim_{\sum_{j\in[J],m\in[M]}\left\|d_j^m\right\|_2^2\to\infty} -\log p\left(\boldsymbol{Y}|\theta,\boldsymbol{\sigma}\right) = \infty.$$

VIII. Proof of Corollary 1

This proof follows closely to the first part of the proof of (Theorem 1, [2]). Let

$$\mathscr{S}_0 = \left\{\theta : -\log p\left(\boldsymbol{Y}|\theta,\boldsymbol{\sigma}\right) \leq -\log p\left(\boldsymbol{Y}|\theta^0,\boldsymbol{\sigma}\right)\right\}, \qquad (64)$$

where θ^0 denotes the initial value of θ. Theorem 1 established that $-\log p\left(\boldsymbol{Y}|\theta,\boldsymbol{\sigma}\right)$ is coercive. In addition, assume, for now, that $-\log p\left(\boldsymbol{Y}|\theta,\boldsymbol{\sigma}\right)$ is a continuous function of θ. Under these conditions, $\mathscr{S}_0$ is a compact set (Theorem 1.2, [3]). In addition, we have that

$$-\log p\left(\boldsymbol{Y}|\theta^{t+1},\boldsymbol{\sigma}\right) \leq -\log p\left(\boldsymbol{Y}|\theta^t,\boldsymbol{\sigma}\right)$$

because θ is updated using EM, which guarantees monotonicity of the log-likelihood [4]. Therefore, the sequence $\{-\log p\left(\boldsymbol{Y}|\theta^t,\boldsymbol{\sigma}\right)\}_{t=1}^{\infty}$ is a monotonically decreasing sequence. This monotonicity property guarantees that $\{\theta^t\}_{t=1}^{\infty} \subseteq \mathscr{S}_0$. Since $\mathscr{S}_0$ is compact, $\{\theta^t\}_{t=1}^{\infty}$ admits at least one limit point.

What remains is to show that $-\log p\left(\boldsymbol{Y}|\theta,\boldsymbol{\sigma}\right)$ is continuous. The continuity of the negative log-likelihood follows directly from the fact that both the determinant and matrix inverse functions are continuous[2].

IX. Proof of Theorem 2

From Corollary 1, we know that $\{\theta^t\}_{t=1}^{\infty}$ admits a limit point. What remains to be shown is that all limit points are stationary and that $\{\log p\left(\boldsymbol{Y}|\theta,\boldsymbol{\sigma}\right)\}_{t=1}^{\infty}$ converges monotincally to $\log p\left(\boldsymbol{Y}|\theta^*,\boldsymbol{\sigma}\right)$ for stationary point θ^*. These follow

[2] See [Theorem 5.19 [5]] and [Theorem 5.20 [5]] for continuity of the matrix determinant and inverse functions, respectively.

directly from [Theorem 2, [4]] if it can be shown that $Q\left(\theta, \theta^t\right)$ is continuous in both θ and θ^t, where $Q\left(\theta, \theta^t\right)$ is given by

$$\begin{aligned} Q\left(\theta, \theta^t\right) &= \left\langle \log p\left(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{D}, \left\{\gamma^i\right\}_{i\in[L]}\right)\right\rangle & (65)\\ &= \sum_{j\in[J]} \left\langle \log p\left(Y_j|X_j, D_j\right) + \log p\left(X_j|\left\{\gamma^i\right\}_{i\in[L]}\right)\right\rangle & (66)\\ &= \sum_{j\in[J]} Q_j\left(\theta, \theta^t\right). & (67)\end{aligned}$$

We will proceed by showing that $Q_j\left(\theta, \theta^t\right)$ is continuous in both θ and θ^t for all j.

First, consider the dependence of $Q_j\left(\theta, \theta^t\right)$ on $\gamma^i[m]$, which is given by

$$-\frac{\log \gamma^i[m]}{2} - \left(\frac{\Sigma^i_{x,j}[m,m] + \left({\mu_j}^i[m]\right)^2}{2\gamma^i[m]}\right) \tag{68}$$

where $\Sigma^i_{x,j}$ and ${\mu_j}^i$ are given by (13) and (14), respectively, and depend on θ^t. It suffices to show that (68) is continuous on the open interval $(0, \infty]$. Since both $\log(\cdot)$ and $(\cdot)^{-1}$ are continuous functions on the interval $(0, \infty]$, it follows that (68) is continuous in $\gamma^i[m]$. The dependence of $Q_j\left(\theta, \theta^t\right)$ on D_j is given by

$$\sum_{i\in[L]} \left(y^i_j\right)^T D_j {\mu_j}^i - \frac{\text{tr}\left(\left(D_j^T D_j\left(\Sigma^i_{x,j} + {\mu_j}^i\left(\mu^i_j\right)^T\right)\right)\right.}{2} \tag{69}$$

which is continuous in D_j. Next, we turn to the task of showing that $Q_j\left(\theta, \theta^t\right)$ is continuous in θ^t, which reduces to showing that $\Sigma^i_{x,j}$ is continuous in both D^t_j and γ^t. Let B be the matrix being inverted in (13):

$$B = \sigma_j^2 \mathsf{I} + D^t_j \Gamma^{t,i}\left(D^t_j\right)^T. \tag{70}$$

The task then reduces to showing that $(B)^{-1}$ is continuous in D^t_j and $\gamma^{t,i}$. We first show that $(B)^{-1}$ exists. Using the assumption that D^t_j is full rank, B is full rank over $\gamma \in (0, \infty]^M$ since

$$\begin{aligned} rank(B) &\geq rank\left(D^t_j \Gamma^{t,i}\left(D^t_j\right)^T\right) & (71)\\ &= rank\left(D^t_j\left(\Gamma^{t,i}\right)^{\frac{1}{2}}\left(D^t_j\left(\Gamma^{t,i}\right)^{\frac{1}{2}}\right)^T\right) & (72)\\ &= rank\left(D^t_j\left(\Gamma^{t,i}\right)^{\frac{1}{2}}\right) & (73)\\ &= N_j & (74)\end{aligned}$$

where $\left(\Gamma^{t,i}\right)^{\frac{1}{2}}$ is a diagonal matrix with the m'th diagonal entry given by $\sqrt{\gamma^{t,i}[m]}$. Therefore, B is full rank and admits an inverse.

Since B is continuous in D^t_j and γ^t, what remains to be shown is that $(B)^{-1}$ is continuous in B, which has previously been shown in [6]. Therefore, $Q_j\left(\theta, \theta^t\right)$ is continuous in θ^t.

X. Proof of Theorem 3

The guarantee given in Theorem 3 follows directly from [Proposition 6, [7]] if we can show that $Q\left(\theta, \theta^t\right)$ has a unique maximizer with respect to θ. Consider the optimization of $Q\left(\theta, \theta^t\right)$ with respect to $\boldsymbol{D}$. This optimization problem can be rewritten as

$$\underset{\boldsymbol{D}}{\arg\max} \sum_{j\in[J]} \sum_{i\in[L]} -\left(y^i_j\right)^T D_j \mu^i_j + D_j\left(\Sigma^i_{x,j} + \mu^i_j\left(\mu^i_j\right)^T\right). \tag{75}$$

Since the terms being summed over j in (75) are independent of each other, (75) is equivalent to solving

$$\underset{D_j}{\arg\max} \sum_{i\in[L]} -\left(y^i_j\right)^T D_j \mu^i_j + D_j\left(\Sigma^i_{x,j} + \mu^i_j\left(\mu^i_j\right)^T\right) \tag{76}$$

for all j.

Since (76) is an unconstrained optimization problem, its maxima must occur at points where the gradient of the objective function vanishes. Taking the gradient of the objective function in (76) with respect to D_j and setting the result to zero, we get

$$D_j \underbrace{\left(U_j U_j^T + \sum_{i\in[L]} \Sigma^i_{x,j}\right)}_{B} = Y_j U^t_j \tag{77}$$

If B is invertible, all of the stationary points of the objective function in (76) have the form $Y_j U_j (B)^{-1}$. Since Y_j, U_j are fixed and $(B)^{-1}$ is unique given U_j and $\left\{\Sigma^i_{x,j}\right\}_{i\in[L]}$, we conclude that the objective function in (76) has exactly one, unique stationary point. In order to show that B is invertible, we observe that $\Sigma^i_{x,j}$ is positive semi-definite for all i and U_j is full rank by assumption. Since the sum of a positive semi-definite and positive definite matrix is positive definite, it follows that B is invertible.

We now turn to the optimization of $Q\left(\theta, \theta^t\right)$ with respect to $\gamma^i[m]$ (since $Q\left(\theta, \theta^t\right)$ is separable in the elements of γ^i). This optimization problem can be rewritten as

$$\begin{aligned} &\underset{\gamma^i[m]\geq 0}{\arg\max} \sum_{j\in[J]} -\frac{\log \gamma^i[m]}{2} - \left(\frac{\Sigma^i_{x,j}[m,m] + \left({\mu_j}^i[m]\right)^2}{2\gamma^i[m]}\right)\\ &= \underset{\gamma^i[m]\geq 0}{\arg\min} \log \gamma^i[m] + \frac{\rho}{\gamma^i[m]} \qquad (78)\end{aligned}$$

where

$$\rho = \frac{1}{J} \sum_{j\in[J]} \Sigma^i_{x,j}[m,m] + \left({\mu_j}^i[m]\right)^2 \tag{79}$$

Note that we explicitly state the constraint on $\gamma^i[m]$ in (78). For $\rho = 0$, (78) reduces to

$$\underset{\gamma^i[m]\geq 0}{\arg\min} \log \gamma^i[m] = 0. \tag{80}$$

For $\rho > 0$, we can show that $\log \gamma^i[m] + \frac{\rho}{\gamma^i[m]} \geq \log \rho + 1$:

$$\begin{aligned} \log \rho + 1 - \log \gamma^i[m] - \frac{\rho}{\gamma^i[m]} &= \log \frac{\rho}{\gamma^i[m]} + 1 - \frac{\rho}{\gamma^i[m]}\\ &\leq^a \frac{\rho}{\gamma^i[m]} - 1 + 1 - \frac{\rho}{\gamma^i[m]}\\ &= 0\\ &\downarrow\\ \log \rho + 1 \leq \log \gamma^i[m] &+ \frac{\rho}{\gamma^i[m]}\end{aligned}$$

where step (a) follows from the identity $\log x \le x - 1$ for $x > 0$ [8]. Equality in step (a) is achieved only for $\gamma^i[m] = \rho$. To see this, we observe that $\log x$ is a strictly concave function that is tangent to the function $x - 1$ at $x = 1$. The strict concavity of $\log x$ implies that it can only be tangent to the function $x-1$ at a single point. Therefore, the objective in (78) is lowerbounded by $\log \rho + 1$, with the lowerbound achieved at $\gamma^i[m] = \rho$. Putting this result together with the case when $\rho = 0$, we see that (78) admits a single, unique optimizer given by ρ in (79).

XI. Proof of Corollary 2

This proof follows closely to the first part of the proof of (Theorem 1, [2]). Let

$$\mathscr{S}_0 = \left\{\{\theta, \boldsymbol{\sigma}\} : -\log p\left(\boldsymbol{Y}|\theta, \boldsymbol{\sigma}\right) \le -\log p\left(\boldsymbol{Y}|\theta^0, \boldsymbol{\sigma^0}\right)\right\}, \tag{81}$$

where θ^0 and $\boldsymbol{\sigma^0}$ denote the initial values of θ and $\boldsymbol{\sigma}$, respectively. Theorem 1 established that $-\log p\left(\boldsymbol{Y}|\theta, \boldsymbol{\sigma}\right)$ is coercive. In addition, assume, for now, that $-\log p\left(\boldsymbol{Y}|\theta, \boldsymbol{\sigma}\right)$ is a continuous function of $\{\theta, \boldsymbol{\sigma}\}$. Under these conditions, $\mathscr{S}_0$ is a compact set (Theorem 1.2, [3]). In addition, we have that

$$-\log p\left(\boldsymbol{Y}|\theta^{t+1}, \boldsymbol{\sigma}^t\right) \le -\log p\left(\boldsymbol{Y}|\theta^t, \boldsymbol{\sigma}^t\right)$$

because θ is updated using EM, which guarantees monotonicity of the log-likelihood [4]. Likewise, we have that

$$-\log p\left(\boldsymbol{Y}|\theta^{t+1}, \boldsymbol{\sigma}^{t+1}\right) \le -\log p\left(\boldsymbol{Y}|\theta^{t+1}, \boldsymbol{\sigma}^t\right)$$

by construction of the update rule in (20). Therefore, the sequence $\left\{-\log p\left(\boldsymbol{Y}|\theta^t, \boldsymbol{\sigma^t}\right)\right\}_{t=1}^{\infty}$ is a monotonically decreasing sequence, i.e.

$$-\log p\left(\boldsymbol{Y}|\theta^{t+1}, \boldsymbol{\sigma}^{t+1}\right) \le -\log p\left(\boldsymbol{Y}|\theta^t, \boldsymbol{\sigma}^t\right). \tag{82}$$

This monotonicity property guarantees that $\{\theta^t, \boldsymbol{\sigma^t}\}_{t=1}^{\infty} \subseteq \mathscr{S}_0$. Since $\mathscr{S}_0$ is compact, $\{\theta^t, \boldsymbol{\sigma^t}\}_{t=1}^{\infty}$ admits at least one limit point.

What remains is to show that $-\log p\left(\boldsymbol{Y}|\theta, \boldsymbol{\sigma}\right)$ is continuous. The continuity of the negative log-likelihood follows directly from the fact that both the determinant and matrix inverse functions are continuous[3].

XII. Proof of Theorem 4

We will show that any point in $\Omega_{\sigma_j^1, j}$ must be in $\Omega_{\sigma_j^2}$. Let the operator $\Theta_l : \mathbb{R}^{M\times M} \to \mathbb{R}^{M-N_j \times M-N_j}$ be defined such that $\Theta_l(\Gamma)$ extracts the top left $M - N_j \times M - N_j$ submatrix of Γ. Let the operator $\Theta_h : \mathbb{R}^{M\times M} \to \mathbb{R}^{N_j\times N_j}$ be defined such that $\Theta_h(\Gamma)$ extracts the bottom right $N \times N$ submatrix of Γ. Using these operators, we can express any element of $\Omega_{\sigma_j, j}$ as

$$\Sigma_{y_j} = \left(\sigma^2 \mathsf{I} + \Theta_h(\Gamma)\right) + \check{D}_j \Theta_l(\Gamma) \check{D}_j^T. \tag{83}$$

Let $\Sigma_{y,j}^1(D_j^1, \gamma^1) \in \Omega_{\sigma_j^1, j}$, where $\Sigma_{y,j}(\cdot,\cdot)$ denotes the dependence of $\Sigma_{y,j}$ on $\check{D}_j$ and γ. We can then show that $\Sigma_{y,j}^1(D_j^1, \gamma^1) = \Sigma_{y,j}^2(D_j^2, \gamma^2) \in \Omega_2$ for

$$\gamma^2[m] = \begin{cases} \gamma^1[m] & \text{if } m \le M - N_j \\ \gamma^1[m] + \left(\sigma_j^1\right)^2 - \left(\sigma_j^2\right)^2 & \text{else} \end{cases}$$

and $\check{D}_j^2 = \check{D}_j^1$. Such a choice of $\Sigma_{y,j}^2(D_j^2, \gamma^2)$ is always possible because $\sigma_j^1 > \sigma_j^2$. The converse is not true for arbitrary choices of D_j^1 and γ^1, leading to the relation $\Omega_{\sigma_j^1} \subseteq \Omega_{\sigma_j^2}$.

XIII. Proof of Theorem 5

We begin by studying the shape of

$$\log p\left(Y_j|\theta^t, \sigma_j\right). \tag{84}$$

The log-likelihood in (84) depends on σ_j through the covariance matrices $\left\{\Sigma_{y,j}^i\right\}_{i\in[L]}$ shown in (11). If we parametrize $p\left(y_j^i; \Sigma_{y,j}^i\right)$ by the precision matrix $\Lambda_j^i = \left(\Sigma_{y,j}^i\right)^{-1}$, then it can be shown that $\log p\left(y_j^i; \Lambda_j^i\right)$ is a strictly concave function of Λ_j^i. Since the log-likelihood of Y_j is a sum of such functions, $\log p\left(Y_j | \left\{\Lambda_j^i\right\}_{i\in[L]}\right)$ is itself a strictly concave function. Therefore, $\log p\left(Y_j | \{\Lambda_j^i\}_{i\in[L]}\right)$ admits a single local maximum $\left\{\Lambda_j^{i,*}\right\}_{i\in[L]}$, which is also its global maximum. Since the mapping from Λ_j^i to $\Sigma_{y,j}^i$ is one-to-one, we conclude that $\log p\left(Y_j | \left\{\Sigma_{y,j}^i\right\}_{i\in[L]}\right)$ also admits a single local maximum, which is also a global maximum. In other words, $\log p\left(Y_j | \left\{\Sigma_{y,j}^i\right\}_{i\in[L]}\right)$ is a strictly quasiconcave function [9]. Consider maximizing the log-likelihood over the convex set $\Sigma_{y,j}^i \in \{\sigma^2 \mathsf{I} + D_j \Gamma^i D_j^T : \sigma_j > 0\}$. Quasiconcave functions admit a single local maximum, which is also the global maximum, over convex sets [9]. We conclude that $\log p\left(Y_j | \left\{\Sigma_{y,j}^i\right\}_{i\in[L]}\right)$ admits a single maximum with respect to σ_j.

Suppose that the condition

$$\sigma_j^t = \sigma_j^{t-1} \tag{85}$$

is satisfied at iteration t (and not before), meaning that

$$\log p\left(Y_j|\theta^t, \alpha\sigma_j^{t-1}\right) < \log p\left(Y_j|\theta^t, \sigma_j^{t-1}\right). \tag{86}$$

Because α_σ can be arbitrarily close to 1 (86) implies that there exists a neighborhood $\left[\sigma_j^{t-1} - \epsilon_1, \sigma_j^{t-1}\right], \epsilon_1 > 0$, over which (84) is increasing. We now claim that there must also exist a neighborhood $\left[\sigma_j^{t-1}, \sigma_j^{t-1} + \epsilon_2\right], \epsilon_2 > 0$, for which (84) is decreasing. Suppose that the converse is true and that there exists a σ_j^* such that

$$\log p\left(Y_j|\theta^t, \sigma_j^{t-1}\right) < \log p\left(Y_j|\theta^t, \sigma_j^*\right), \sigma_j^* > \sigma_j^{t-1}. \tag{87}$$

Remember that $\theta^t = \left\{\left\{\gamma^{t,i}\right\}_{i\in[L]}, \boldsymbol{D^t}\right\}$ where $D_j^t \in \Psi_j$. The inequality in (87) means that there exists $\theta^* = \left\{\left\{\gamma^{*,i}\right\}_{i\in[L]}, \boldsymbol{D^t}\right\}$ with

$$\gamma^{*,i}[m] = \begin{cases} \gamma^{t,i}[m] & \text{if } m \le M - N_j \\ \gamma^{t,i}[m] + \left(\sigma_j^*\right)^2 - \left(\sigma_j^{t-1}\right)^2 & \text{else} \end{cases}. \tag{88}$$

[3] See [Theorem 5.19 [5]] and [Theorem 5.20 [5]] for continuity of the matrix determinant and inverse functions, respectively.

such that

$$\log p\left(Y_j|\theta^t, \sigma_j^{t-1}\right) < \log p\left(Y_j|\theta^*, \sigma_j^{t-1}\right). \tag{89}$$

But (89) must be a contradiction since we have assumed that

$$\theta^t = \arg\max_{\theta} \log p\left(Y_j|\theta, \sigma_j^{t-1}\right). \tag{90}$$

The condition that $\sigma_j^0 \geq \arg\max_{\sigma_j} \max_\theta \log p\left(Y_j|\theta, \sigma_j\right)$ ensures that the preceding argument holds at iteration $t = 1$.

To conclude, we have shown that if (85) is satisfied, σ_j^{t-1} is a local maximum. Since we have already shown that the objective in (84) only admits a single maximum, this completes the proof.

XIV. Proof of Theorem 6

Since Corollary 2 established that $\{\theta^t, \boldsymbol{\sigma^t}\}_{t=1}^{\infty}$ admits at least one limit point, what remains is to show that all limit points are stationary. Let $\{\theta^*, \boldsymbol{\sigma^*}\}$ denote any limit point of $\{\theta^t, \boldsymbol{\sigma^t}\}_{t=1}^{\infty}$. For any $\boldsymbol{\sigma}$, we know that a limit point $\bar{\theta}$ is stationary from Theorem 2. Therefore, $\bar{\theta} = \theta^*$ must be a stationary point.

For any θ, we know that a limit point $\bar{\boldsymbol{\sigma}}$ must be the global maximizer of the log-likelihood from Theorem 5. Since a global maximizer must be stationary, we conclude that $\bar{\boldsymbol{\sigma}}$ must be stationary. Therefore, $\bar{\boldsymbol{\sigma}} = \boldsymbol{\sigma^*}$ must be a stationary point.

XV. Proof of Thm. 7

This proof is an extension of the proof shown in [10]. Under the assumptions of Theorem 7, we can focus exclusively on recovering $\boldsymbol{D}$ because, given $\{\boldsymbol{Y}, \boldsymbol{D}\}$, the sparse codes $\boldsymbol{X}$ are unique [11]. To prove that $\boldsymbol{D}$ is unique, we will show how $\boldsymbol{D}$ can be recovered by construction.

The construction of $\hat{\boldsymbol{D}}$ proceeds in three steps:

1) Divide the columns of Y_j into $R = \binom{M}{s}$ sets $\{G_j^1, \cdots, G_j^R\}$ for all j, where $G_j^k = \left\{i : y_j^i \in span\left(D_j[:, \Upsilon^k]\right)\right\}$ and Υ^k denotes the k'th subset of size s of $[M]$.
2) Detect pairs $(G_j^{k_1}, G_j^{k_2})$ such that $|G_j^{k_1} \cap G_j^{k_2}| = 1$ for all j.
3) For each j, find the atom common to Υ^{k_1} and Υ^{k_2}. This atom is necessarily one of the atoms of D_j [10]. Repeat for all pairs (k_1, k_2).

We begin by describing how the data is clustered. Starting with modality j^*, we begin by testing every group of $s+1$ data points from Y_{j^*}. The rank of this group of points will be s if and only if the points lie in the subspace spanned by a set of s columns from D_{j^*} [10]. Once $\left\{G_{j^*}^k\right\}_{k=1}^R$ has been established, the remaining points in Y_{j^*} which have not been assigned to a set are combined with one of the groups $G_{j^*}^k$ based on the fact that the rank of the subspace spanned by the columns indexed by $G_{j^*}^k$ and an additional column, y_j^i, from Y_{j^*} is s if an only if $y_j^i \in span\left(D_j[:, \Upsilon_k]\right)$. Finally, due to the nature of the data generation process, we know that $G_j^k = \{i : P[j, i] = 1 \text{ and } i \in G_{j^*}^k\}, \forall j$. Note that the construction of $G_{j^*}^k$ requires $s+1$ data points from modality j^*, but we get G_j^k directly from $G_{j^*}^k$.

Next, we describe the process by which we detect pairs $(G_j^{k_1}, G_j^{k_2})$ such that $|\Upsilon^{k_1} \cap \Upsilon^{k_2}| = 1$. This can be done for each modality j independently. Namely, for each pair $(G_j^{k_1}, G_j^{k_2})$, we test the rank of the subspace spanned by the columns of Y_j indexed by $G_j^{k_1} \cup G_j^{k_2}$. The rank of this subspace will be $2s-1$ if and only if the intersection of $span(G_j^{k_1})$ and $span(G_j^{k_2})$ has dimension 1. This process is guaranteed to produce every atom of D_j at least once [10].

Finally, we describe how to form $\hat{D}_j$. Given a pair $(G_j^{k_1}, G_j^{k_2})$ such that $|\Upsilon^{k_1} \cap \Upsilon^{k_2}| = 1$, we extract any s points from $G_j^{k_1}$ and any s points from $G_j^{k_2}$ and concatenate them into two matrices B^{k_1} and B^{k_2}, respectively. There exist vectors v^1 and v^2 such that both $B^{k_1}v^1$ and $B^{k_2}v^2$ are parallel to the atom of D_j of interest. We can now set up a system of equations given by

$$\underbrace{\begin{bmatrix} B^{k_1} & -B^{k_2} \end{bmatrix}}_{B} \underbrace{\begin{bmatrix} v^1 \\ v^2 \end{bmatrix}}_{v} = 0 \tag{91}$$

and find v, which is guaranteed to exist since $rank(B) = 2s-1$ from the detection step. We can then extract v^1 from v and find one of the columns of D_j (up to scaling) using $B^{k_1}v^1$. This process can be repeated to find all M atoms of D_j (up to scale) [10].

The major difference between the proof given here and the one in [10] for the unimodal dictionary learning problem is that we require $s+1$ data points per every s dimensional subspace for *one* of the modalities and only s data points per subspace for the rest of the modalities. The reason that we can use *less* samples stems from the special structure of the multimodal data generation process. In the end, we require $s+1$ data points from modality j^* to complete the data clustering step and s data points from each of the $J-1$ data modalities to complete the extraction step.

References

[1] R. M. Neal and G. E. Hinton, "A view of the em algorithm that justifies incremental, sparse, and other variants," in *Learning in graphical models*. Springer, 1998, pp. 355–368.

[2] R. Zhao and V. Y. Tan, "A unified convergence analysis of the multiplicative update algorithm for nonnegative matrix factorization," *arXiv preprint arXiv:1609.00951*, 2016.

[3] J. V. Burke, *Undergraduate Nonlinear Continuous Optimization*.

[4] C. J. Wu, "On the convergence properties of the em algorithm," *The Annals of statistics*, pp. 95–103, 1983.

[5] J. R. Schott, *Matrix analysis for statistics*. John Wiley & Sons, 2016.

[6] G. Stewart, "On the continuity of the generalized inverse," *SIAM Journal on Applied Mathematics*, vol. 17, no. 1, pp. 33–45, 1969.

[7] A. Gunawardana and W. Byrne, "Convergence theorems for generalized alternating minimization procedures," *Journal of Machine Learning Research*, vol. 6, no. Dec, pp. 2049–2073, 2005.

[8] E. Love, "64.4 some logarithm inequalities," *The Mathematical Gazette*, vol. 64, no. 427, pp. 55–57, 1980.

[9] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[10] M. Aharon, M. Elad, and A. M. Bruckstein, "On the uniqueness of overcomplete dictionaries, and a practical way to retrieve them," *Linear algebra and its applications*, vol. 416, no. 1, pp. 48–67, 2006.

[11] D. L. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via 1 minimization," *Proceedings of the National Academy of Sciences*, vol. 100, no. 5, pp. 2197–2202, 2003.