A Proof of Theorem 3

Proof. We prove by mapping (4) to an equivalent problem by performing a variable change.

Let $\mathsf{G}_\mathsf{S} := \bigcup_{i\in\mathsf{S}} \mathsf{G}_i$. Note that the inner optimization $\min_{|\mathsf{G}_\mathsf{S}|\leq k, q\in\mathcal{F}_{\mathsf{G}_\mathsf{S}}} \mathrm{KL}(\mathrm{q}\|\mathrm{p}) = -\log p(\mathbf{x}_{\mathsf{G}\setminus\mathsf{G}_\mathsf{S}})$ [Koyejo et al., 2014].

Define the function $J : \mathfrak{p}(r) \to \mathbb{R}$ as $J(\mathsf{S}) := \log \mathrm{p}(\mathbf{x}_{\mathsf{G}\setminus\mathsf{G}_\mathsf{S}} = 0)$, and the function $\tilde{J} : \mathfrak{p}(r) \to \mathbb{R}$ as $\tilde{J}(\mathsf{S}) := J(\mathsf{S}) - J(\emptyset)$.

Define the costs associated with picking G_i as $c_i = |\mathsf{G}_i| \forall i \in [r]$. The cost function of a set $\mathsf{s} \subset \mathsf{G}$ can thus be written as $c(\mathsf{s}) := \sum_{\forall i \text{ s.t. } \mathsf{G}_i\in\mathsf{s}} c_i$ The optimization problem 4 is then equivalent to $\max_{\sum_{i\in\mathsf{S}} c_i\leq k} \tilde{J}(\mathsf{S})$.

The result follows from Theorem 1.

□

B Extension to multiple factors

The factor models defined in Sections 4.3,4.2 are readily extensible to multiple factors using deflation techniques, either by information projection or simple substraction. See the work of [Khanna et al., 2017] for further discussion. An alternative is also to infer for all factors as is typically done in classic non-sparse settings. We present the detailed equations and explicitly derive EM algorithm and the relevant equations next for the case of the sparse PCA model (Section 4.2). Note that because of the partition constraint construction described in Section 4.3, the equations are also relevant as the EM algorithm for the probabilistic CCA. The only difference is in the Estep, where Algorithm 2 is to be used instead of 3.

We have n observations made in a d dimensional space, which are stacked in a matrix $\mathbf{T} \in \mathbb{R}^{n\times d}$. Drawing analogy from traditional PCA, we want to search for a few sparse basis vectors whose linear combination generates the observation matrix with small error. Additionally, the basis vectors themselves may have structure.

We model the observation matrix as a product of a parameter $\mathbf{X} \in \mathbb{R}^{n\times p}$ and a *sparse* $\mathbf{W} \in \mathbb{R}^{p\times d}$. The sparse basis vectors are stacked as rows of $\mathbf{W}$, and their linear combination is modelled by $\mathbf{X}$. We are looking at scenarios with $n \gg d$, so the above factorization is useful for small p, which is set according to the domain. $\boldsymbol{\mu}$ is the matrix of column means generated as, $\boldsymbol{\mu} = \text{columnMeans}(\mathbf{T})^\dagger \otimes \mathbf{1}$, and gaussian noise is represented by $\boldsymbol{\epsilon}_{ij} \sim \mathrm{N}(0, \sigma^2), \forall i \in [n], \forall j \in [d]$.

$$\mathbf{T} = \mathbf{X}\mathbf{W} + \boldsymbol{\mu} + \boldsymbol{\epsilon} \tag{6}$$

We have a normal prior on each row of $\mathbf{W}$. $\forall i \in [p], \mathbf{W}_{i,\cdot} \sim \mathrm{N}(0, \mathbf{C})$, for a known matrix $\mathbf{C}$, while the rows are independent. In the matrix-variate normal form, $\mathbf{W} \sim \mathrm{MVN}(0, \mathbf{C}, \mathbf{I})$. For convenience that will be apparent in coming pages, the above equation can be vectorized and rewritten as

$$\overrightarrow{\mathbf{T}} = (\mathbf{I} \otimes \mathbf{X})\overrightarrow{\mathbf{W}} + \overrightarrow{\boldsymbol{\mu}} + \overrightarrow{\boldsymbol{\epsilon}} \tag{7}$$

Inference and learning can be performed for this model by an Expectation Maximization (EM) algorithm, We introduce sparsity into $\mathbf{W}$ by constraining its support in the variational E-step as detailed next.

C EM Algorithm

Neal and Hinton [1998] give a free energy function based interpretation of the EM algorithm wherein the E-step maximizes the energy function $\mathscr{F}$ in the space of distributions of the missing data and the M-step maximizes it in the parameter space. In our model, $\mathbf{X}$ and σ^2 are the parameters, and $\mathbf{W}$ can be treated as the missing data for the EM algorithm. With $\mathrm{KL}(\cdot,\cdot)$ as the Kullback-Liebler distance, from Neal and Hinton [1998], we have

$$\begin{aligned}\mathscr{F}(\mathrm{q}(\mathbf{W}), \mathbf{X}, \sigma^2) = -\mathrm{KL}\big(\mathrm{q}(\mathbf{W})\|\mathrm{p}(\mathbf{W}|\mathbf{T}; \mathbf{X}, \sigma^2)\big) \\ + \log \mathrm{P}(\mathbf{T}; X, \sigma^2).\end{aligned} \tag{8}$$

E-step can then be viewed as the search for the best $\mathrm{q}(\mathbf{W})$,

$$\text{E-step: } \max_{\mathrm{q}} \mathscr{F}(\mathrm{q}(\mathbf{W}), \mathbf{X}, \sigma^2). \tag{9}$$

The M-step is the search for the best parameters,

$$\text{M-step: } \max_{X,\sigma^2} \mathscr{F}(\mathrm{q}(\mathbf{W}), \mathbf{X}, \sigma^2). \tag{10}$$

This view of the EM algorithm provides the flexibility to design algorithms with any E and M steps that monotonically increase $\mathscr{F}$.

C.1 Variational E-step

Equation 9 finds the best q to maximize the free energy function $\mathscr{F}$. An unconstrained search returns the posterior $\mathrm{p}(\mathbf{W}|\mathbf{T}; \mathbf{X}, \sigma^2)$, making the KL term 0. $\mathbf{W}$ is size $p \times d$, so doing a full E-step is costly. To introduce sparsity, and to make the Estep more tractable, we present a variational E-step. Variational methods perform the search for best q over a constrained set. For sparsity, we introduce the constraint is that $\mathrm{q}(\overrightarrow{\mathbf{W}})$ is k-sparse i.e. it has support only on at most k out of total $n \times d$ dimensions. From Equations 9 and 8, it follows that the variational E-step can be written using the vectorized $\mathbf{W}$ as:

$$\min_{|\mathrm{q}|=k} \mathrm{KL}\Big((\mathrm{q}(\overrightarrow{\mathbf{W}})\|\mathrm{p}(\overrightarrow{\mathbf{W}}|\mathbf{T}; \mathbf{X}, \sigma^2)\Big). \tag{11}$$

Let K be the enumerative set of constraints (group sparsity is a special case, but the equation is valid for other constraint sets as well). Expanding the KL equation, it is straightforward to see that the Equation 11 is equivalent to

$$\max_{\text{Supp}(P)\in\mathsf{K}} \log(\text{p}(\overrightarrow{\text{W}}_{\mathsf{K}^c} = \mathbf{0}_{\mathsf{K}^c}|\mathbf{T};\mathbf{X},\sigma^2)) \tag{12}$$

Equation 12 is the resulting optimization problem to be solved for variational E-step. However, it requires a combinatorial search over the set k which is of size $\frac{(p*d)!}{(p*d-k)!k!}$, where $p*d$ is dimension of the ambient space of $\overrightarrow{\text{W}}$. Searching over *atmost* k-sparse supports (instead of *exactly* k-sparse) is more costly as the size of the set to be searched over increases. Since this is to be done in every iteration of the EM, it is prohibitively expensive. As a workaround, we proposed selecting each of the k dimensions greedily.

C.1.1 E-step Equations

Having defined the optimization function, the algorithm and its gaurantees, we now derive explicit equations for the variational E-step.

For deriving the E-step equations, we first require the posterior $\text{p}(\overrightarrow{\text{W}}|\overrightarrow{\text{T}};\mathbf{X},\sigma^2)$. In this section, we suppress $\{\mathbf{X},\sigma^2\}$ from the equations for brevity since dependence on them is obvious. From standard properties of the gaussians, we know that,

$$\begin{aligned}
&\text{p}(\overrightarrow{\text{W}}|\overrightarrow{\text{T}}) \sim \mathcal{N}(\boldsymbol{m},\boldsymbol{\Sigma})\\
&\text{where,}\\
&\boldsymbol{\Sigma}^{-1} = \frac{1}{\sigma^2}(\mathbf{I}_\text{d}\otimes\mathbf{X}^\dagger\mathbf{X}) + (\mathbf{C}^{-1}\otimes\mathbf{I}_\text{p}),\\
&\mathbf{m} = \boldsymbol{\Sigma}\,\frac{1}{\sigma^2}(\mathbf{I}_\text{d}\otimes\mathbf{X}^\dagger)(\overrightarrow{\text{T}}-\overrightarrow{\mu}).
\end{aligned} \tag{13}$$

We can now expand Equation 12 as

$$\max_{\mathsf{K}}\ \mathbf{m}_{\mathsf{K}^c}^\dagger[\boldsymbol{\Sigma}_{\mathsf{K}^c}]^{-1}\mathbf{m}_{\mathsf{K}^c} - \log\det\boldsymbol{\Sigma}_{\mathsf{K}^c} + \text{const},$$

which is equivalent to:

$$\begin{aligned}
&\text{define}\ \ \mathbf{r} = \boldsymbol{\Sigma}^{-1}\mathbf{m};\\
&\max_{\mathsf{K}}\ \mathbf{r}_\mathsf{K}[[\boldsymbol{\Sigma}^{-1}]_\mathsf{K}]^{-1}\mathbf{r}_\mathsf{K} - \log\det[\boldsymbol{\Sigma}^{-1}]_\mathsf{K}.
\end{aligned} \tag{14}$$

Recall that $\boldsymbol{\Sigma}_\mathsf{K}$ is the submatrix of $\boldsymbol{\Sigma}$ supported on K, similarly for $[\boldsymbol{\Sigma}^{-1}]_\mathsf{K}$. Note that Equation 14 is in space of size $p*d$. Since $[\boldsymbol{\Sigma}^{-1}]_\mathsf{K}$ and $\mathbf{r}_\mathsf{K}$ are easy to calculate by some smart indexing, this version of the equation is convenient to implement and use.

Say after solving the constrained optimization problem specified by Equations [13,14], we obtain the best support as K^*, then the resulting solution density, say q^*, which is known to be the conditional Koyejo et al. [2014], is

$$\begin{aligned}
&q^* \sim \mathcal{N}(\mathbf{c},\mathbf{D})\\
&\text{where,}\\
&\mathbf{D}^{-1} = [\boldsymbol{\Sigma}^{-1}]_{\mathsf{K}^*},\\
&\mathbf{c} = \mathbf{D}\mathbf{r}_{\mathsf{K}^*}
\end{aligned} \tag{15}$$

Recall, q^* has support only on K^*, so in Equation 15, $\mathbf{c}\in\mathbb{R}^{|\mathsf{K}^*|}$, $\mathbf{D}\in\mathbb{R}^{|\mathsf{K}^*|\times|\mathsf{K}^*|}$.

C.2 M-step

Equation 10 optimizes $\mathscr{F}$ in the parameter space to get the argmax parameters that maximize $\mathscr{F}$. However, it turns out solving for $\{X,\sigma^2\}$ over $\mathscr{F}$ directly is costly. Since the free energy view of the EM shows that any M-step that increases $\mathscr{F}$ suffices, we maximize the log likelihood portion of $\mathscr{F}$ instead for the M-step. Recall q^* is the distribution on $\overrightarrow{\text{W}}$ obtained from the E-step, the effective M-step is:

$$\max_{\mathbf{X},\sigma^2} \mathbb{E}_{q^*}\ [\log\text{p}(\overrightarrow{\text{T}}|\overrightarrow{\text{W}};\mathbf{X},\sigma^2)] \tag{16}$$

C.2.1 M-step Equations

Since q^* has measure 0 outside K^*, let $\widehat{\mathbf{c}}$ represent the mean vector $\mathbf{c}$ expanded from $|\mathsf{K}^*|$ to ambient dimension $p*d$, with zeroes padded as needed. Similarly, $\widehat{\mathbf{D}}\in\mathbb{R}^{pd\times pd}$ represents $\mathbf{D}$ with zeroes padded as needed. For brevity, the following equations assume $\overrightarrow{\text{T}}$ to be zero mean i.e. $\overrightarrow{\mu}=0$, the derivation extends to non-zero mean $\overrightarrow{\text{T}}$ trivially by replacing $\overrightarrow{\text{T}}$ with $\overrightarrow{\text{T}}-\overrightarrow{\mu}$. Equation 16 can be written as:

$$\begin{aligned}
&\max_{\mathbf{X},\sigma^2} \mathbb{E}_{q^*}[\frac{-1}{2\sigma^2}(\overrightarrow{\text{T}}-(\mathbf{I}\otimes\mathbf{X})\overrightarrow{\text{W}})^\dagger(\overrightarrow{\text{T}}-(\mathbf{I}\otimes\mathbf{X})\overrightarrow{\text{W}})\\
&\qquad - nd\log\sigma^2]\\
&\equiv \max_{\mathbf{X},\sigma^2}\ \frac{-1}{2\sigma^2}\mathcal{V}(\mathbf{X}) - nd\log\sigma^2,\\
&\text{where,}\\
&\mathcal{V}(\mathbf{X}) = \text{r}(\mathbf{I}\otimes\mathbf{X}^\dagger\mathbf{X})(\widehat{\mathbf{c}}\widehat{\mathbf{c}}^\dagger+\widehat{\mathbf{D}}) - 2\widehat{\mathbf{c}}^\dagger\overrightarrow{(\mathbf{X}^\dagger\mathbf{T})}
\end{aligned}$$

Clearly , $\mathbf{X}$ and σ^2 can be updated separately.

For $\mathbf{X}$, it is easy to take gradient of $\mathcal{V}(\mathbf{X})$ w.r.t $\mathbf{X}$. We can matricize $\widehat{\mathbf{c}}$ as mat$(\widehat{\mathbf{c}})\in\mathbb{R}^{p\times d}$ to rewrite the second term as a trace: rmat$(\widehat{\mathbf{c}})^\dagger\mathbf{X}^\dagger\mathbf{T}$. Note that $\mathbf{I}\otimes\mathbf{X}^\dagger\mathbf{X}$ is a block diagonal matrix with the same block $\mathbf{X}^\dagger\mathbf{X}$ repeating over and over. Thus, we can define $\mathbf{M} = \sum_{\text{blocks}}\widehat{\mathbf{c}}\widehat{\mathbf{c}}^\dagger + \widehat{\mathbf{D}}$, where the summation is over diagonal blocks of size $p\times p$

(sp $\mathbf{M} \in \mathbb{R}^{p \times p}$), to write $\mathcal{V}(\mathbf{X})$ and its gradient as

$$\begin{aligned} \mathcal{V}(\mathbf{X}) &= \mathrm{r}\mathbf{X}^{\dagger}\mathbf{X}\mathbf{M} - \mathrm{rmat}(\widehat{\mathbf{c}})^{\dagger}\mathbf{X}^{\dagger}\mathbf{T} \\ \frac{\partial \mathcal{V}}{\partial \mathbf{X}} &= 2\mathbf{X}\mathbf{M} - \mathbf{T}\mathrm{mat}(\widehat{\mathbf{c}})^{\dagger} \end{aligned} \tag{17}$$

Equation 17 gives closed form solution if $\mathbf{M}$ is invertible. If that is not the case, gradient steps can be taken to update $\mathbf{X}$. Recall that even a single gradient step to update $\mathbf{X}$ suffices as required by the free energy view of the EM.

Once $\mathbf{X}$ is updated, updating σ^2 is more straightforward by taking derivative and setting to 0 to get

$$\sigma^2 = \frac{\mathcal{V}(\mathbf{X})}{2nd} \tag{18}$$

Algorithm 5: EM Algorithm for sparse projections

Input: $k, p, d, \mathbf{C}, \mathbf{T}$
Initialize $\mathbf{X}$ randomly
while not converged **do**
  E-Step
  Init: $\mathsf{K}^* = \{\}$
  for $i = 1 \dots k$ **do**
    Solve Equation 14 with
      $\mathsf{K} = \mathsf{H} \bigcup \{j\},\ \forall j \in [p * d], j \notin \mathsf{K}^*$
    Set $\mathsf{K}^* \bigcup \{j^*\}$, where j^* is argmax from previous step
  end for
  Use Equation 15 to get q^*
  M-Step
  Solve for $\mathbf{X}$ using Equation 17
  Solve for σ^2 using Equation 18
end while
return$(q^*, \mathbf{X}, \sigma^2)$