*Submitted to the Annals of Applied Statistics*

# SUPPLEMENTARY MATERIALS: TRANSPOSABLE REGULARIZED COVARIANCE MODELS WITH AN APPLICATION TO MISSING DATA IMPUTATION

By Genevera I. Allen and Robert Tibshirani

*Stanford University*

## 1. Imputation for Multivariate Data (RCMimpute).

The EM algorithm for the multivariate normal has been traditionally used to impute missing values. This method is based on maximum likelihood estimation of parameters in the presence of missing values. Imputation is simply a side result [12, 9]. Many have noted that this approach is not well suited to high-dimensional data, especially when the number of features is large. Hence, a common remedy to this problem, seen predominately in image processing, is to penalize the log-likelihood, [2], [13], [6], [3].

Our method maximizes the penalized log-likelihood of the observed data given below. First, however, we assume that our data $x_i \sim N(\mu, \mathbf{\Delta})$ with $i = 1 \dots n$ observations and $p$ features, with values missing at random. We let $x_{i,o_i}$ denote the observed components of observation $i$ and $x_{i,m_i}$ denote the missing components. For each observation, we partition the mean and covariance to correspond to the observed parts of observation $i$ and denote these as $\mu_{o_i}$ and $\mathbf{\Delta}_{o_i,o_i}$.

The penalized observed data log-likelihood is:

$$\begin{aligned} \ell_{\text{obs}}(\mu, \mathbf{\Delta}) = \frac{1}{2} \sum_{i=1}^{n} \left[ \log | \, \mathbf{\Delta}_{o_i,o_i} \, | - (x_{o_i} - \mu_{o_i})^T \, \mathbf{\Delta}^{\text{-}1}{}_{o_i,o_i} (x_{o_i} - \mu_{o_i}) \right] \\ - \rho || \, \mathbf{\Delta}^{\text{-}1} \, ||^q. \end{aligned} \tag{1}$$

Notice that except for the last penalty term, this is the same observed log-likelihood that is maximized by the original multivariate normal EM algorithm [9].

Missing data imputation is part of the E step of the algorithm in which the conditional expectation of the complete-data log-likelihood is taken given the current parameter estimates as in un-penalized EM algorithms. This is denoted by $Q(\theta|\theta^{(k)}) = \mathrm{E}\left(\ell(\mu, \mathbf{\Delta}) | X_o, \mu', \mathbf{\Delta}'\right)$, letting $\theta = (\mu, \mathbf{\Delta})$ and where $X_o$ denotes the totality of observed elements of $X$. We can break this down into two parts which we term the imputation and covariance correction steps. These steps come directly from the conditional distribution formulas for the multivariate normal.

Imputation Step:

$$\hat{x}_{i,j} = \mathrm{E}\left(x_{i,j}|x_{i,o_i},\mu',\boldsymbol{\Delta}'\right) = \begin{cases} \mu'_{m_i} + \boldsymbol{\Delta}'_{m_i,o_i}\,\boldsymbol{\Delta}^{\text{-}1'}{}_{o_i,o_i}\left(x_{i,o_i} - \mu'_{o_i}\right), & \text{if } j \in m_i \\ x_{i,j}, & \text{if } j \in o_i. \end{cases}$$

Covariance Correction Step: $\mathrm{E}\left(x_{i,j}x_{i,j'}|x_{i,o_i},\mu',\boldsymbol{\Delta}'\right) = \hat{x}_{i,j}\hat{x}_{i,j'} + c_{i,jj'}$,

$$c_{i,jj'} = \begin{cases} \boldsymbol{\Delta}'_{m_i,m_i} - \boldsymbol{\Delta}'_{m_i,o_i}\,\boldsymbol{\Delta}^{\text{-}1'}{}_{o_i,o_i}\,\boldsymbol{\Delta}'_{o_i,m_i}, & \text{if } j,j' \in m_i \\ 0, & \text{else.} \end{cases}$$

We call $c_{i,jj'}$ the covariance correction term because it is added to the cross products forming the covariance matrix. Notice that $c_{i,jj'}$ is only non-zero if both $j$ and $j'$ are missing. Intuitively, this correction term is needed because in the imputation step, we set the missing values equal to their conditional expectations and hence lose some of the variance associated with these elements.

The Maximization step is where our algorithm differs from that of Little and Rubin [9]. In the M step, we must maximize $Q(\theta|\theta^{(k)})$ with respect to $\theta$ to obtain the new estimate $\theta^{(k+1)}$, giving

$$Q(\theta|\theta^{(k)}) = \frac{n}{2}\log|\,\boldsymbol{\Delta}^{\text{-}1}\,| - \frac{1}{2}\mathrm{tr}\left(\hat{\boldsymbol{\Delta}}'\,\boldsymbol{\Delta}^{\text{-}1}\right) - \rho||\,\boldsymbol{\Delta}^{\text{-}1}\,||^q,$$

$$\text{where} \qquad \hat{\boldsymbol{\Delta}}'_{jj'} = \sum_{i=1}^{n}\left[(\hat{x}_{ij} - \mu_j)(\hat{x}_{ij} - \mu_j) + c_{i,jj'}\right].$$

The computations in the E step come into $Q$ through $\hat{\boldsymbol{\Delta}}'$. Maximizing with respect to $\mu$, gives the estimate $\hat{\mu}_j = \sum_{i=1}^{n} = \hat{x}_{ij}/n$. Replacing $\mu$ with $\hat{\mu}$ in $\hat{\boldsymbol{\Delta}}'$, we see that $Q$ has the structure of the regularized covariance models. Hence, our estimate of $\hat{\boldsymbol{\Delta}}$ is obtained by applying either the $L_1$ or $L_2$ solvers of the RCM problem. We break this M step into two parts which we present in the imputation algorithm, Algorithm 1.

This regularized covariance model imputation approach (RCMimpute) is closely related to other penalized EM methods for missing value estimation. These algorithms give non-singular covariance estimates [6], thus enabling use of the EM framework when $p > n$. Also, our algorithm has a unified theoretical framework based on the regularized covariance models.

**2. Covariance Estimation Results.** We investigate the accuracy of our transposable regularized covariance estimates through simulation using the Kullback-Leibler divergence as the metric. As the focus of this paper is on the application of our models to missing data imputation, we compare our covariance estimates to simple shrinkage estimates for completeness. We

**Algorithm 1** Imputation for Regularized Covariance Models (RCMimpute)

```
1. Initialization:
    (a) Set the missing values to the mean: $\hat{x}_{i,m_i} = \sum_{i \in o_i} x_{ij}/n_i$
    (b) Set $\mu^{(0)}$ and $\mathbf{\Delta}^{(0)}$ to the empirical mean and covariance.
2. E Step:
    (a) Imputation: Compute $\mathrm{E}\left(x_{i,j} | x_{i,o_i}, \mu^{(k)}, \mathbf{\Delta}^{(k)}\right)$.
    (b) Cross Products: Compute $\mathrm{E}\left(x_{i,j} x_{i,j'} | x_{i,o_i}, \mu^{(k)}, \mathbf{\Delta}^{(k)}\right)$.
3. M Step:
    (a) Update Estimates: $\hat{\mu}_j$ & $\hat{\mathbf{\Delta}}'_{jj'}$.
    (b) Maximize penalized log-likelihood with respect to $\mathbf{\Delta}^{\text{-}1}$ to obtain the new
        estimate $\hat{\mathbf{\Delta}}$.
4. Repeat steps 2-3 until convergence.
```

will assume that our data has mean zero an has previously been centered as in Proposition 1, and the Kullback-Leibler (K-L) divergence is given in Proposition 5.

PROPOSITION 5. *The Kullback-Leibler divergence for the mean-restricted matrix-variate normal with mean zero is*

$$\mathrm{E}_{(\mathbf{\Sigma},\mathbf{\Delta})}\left[\ell(\mathbf{\Sigma},\mathbf{\Delta}) - \ell(\hat{\mathbf{\Sigma}},\hat{\mathbf{\Delta}})\right] = \frac{p}{2}\log|\hat{\mathbf{\Sigma}}\,\mathbf{\Sigma}^{\text{-}1}| + \frac{n}{2}\log|\hat{\mathbf{\Delta}}\,\mathbf{\Delta}^{\text{-}1}| - \frac{np}{2} + \frac{1}{2}\mathrm{tr}\left(\hat{\mathbf{\Sigma}}^{-1}\,\mathbf{\Sigma}\right)\mathrm{tr}\left(\hat{\mathbf{\Delta}}^{-1}\,\mathbf{\Delta}\right).$$

*Proof. See Section 7.*

In Figure 1, we compare covariance estimates of the $L_1 : L_1$ and $L_2 : L_2$ transposable regularized covariance models to three other shrinkage covariance estimates parameterized by $\rho$ using the Kullback-Leibler (K-L) distance. The data of dimension $50 \times 50$ was simulated under the matrix-variate normal model with autoregressive covariances, left, and blocked diagonal covariances, right. The covariances are as follows:

- Autoregressive: $\mathbf{\Sigma}_{ij} = 0.8^{|i-j|}$ and $\mathbf{\Delta}_{ij} = 0.6^{|i-j|}$.
- Blocked diagonal: Off-diagonal elements of $5 \times 5$ blocks of $\mathbf{\Sigma}$ are 0.8 and of $\mathbf{\Delta}$, 0.6.

In the TRCM estimates, both penalty parameters, $\rho_r$ and $\rho_c$ are set to $\rho$ for comparability. The three comparison transposable shrinkage covariance estimators are straightforward extensions of simple multivariate shrinkage estimates.

- SE I: Denote the SVD of $\mathbf{X}$ by $\mathbf{X} = \mathbf{U}\,\mathbf{D}\,\mathbf{V}^{\mathrm{T}}$, then

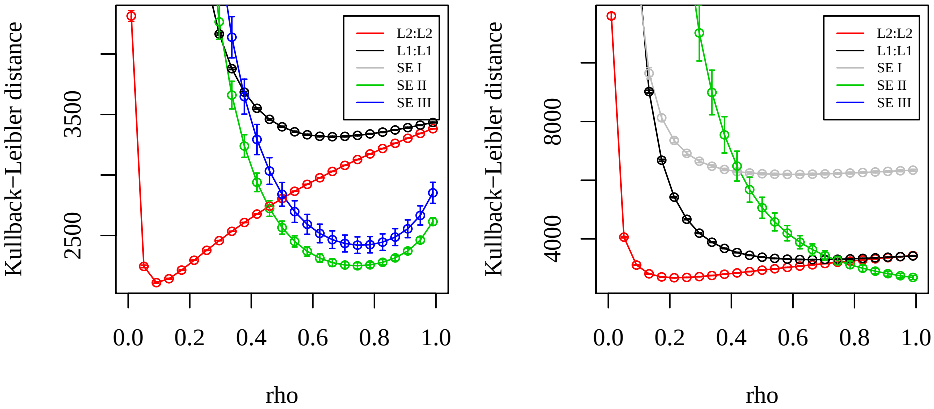

FIG 1. *Mean Kullback-Leibler divergence with standard errors verses the penalty parameter $\rho$ ("rho") for 50 datasets of dimension $50 \times 50$ simulated from the matrix-variate normal distribution. Autoregressive (left) and blocked diagonal (right) covariance matrices are used as given in Section 2. The divergence is compared between two transposable regularized covariance estimates, $L_2 : L_2$ and $L_1 : L_1$,with the penalty parameters on both covariance matrices equal to $\rho$ and three families of shrinkage estimators, SE I, SE II, and SE III described in Section 2 also indexed by $\rho$.*

- $\hat{\mathbf{\Sigma}}_{\text{SE I}} = \mathbf{U}(\mathbf{D}^2 + \rho\mathbf{I})\,\mathbf{U}^{\text{T}}, \quad \& \quad \hat{\mathbf{\Delta}}_{\text{SE I}} = \mathbf{V}(\mathbf{D}^2 + \rho\mathbf{I})\,\mathbf{V}^{\text{T}}.$

- SE II & III: Let $S$ be the unpenalized MLE of $\mathbf{\Sigma}$, i.e. $S = \mathbf{X}\,\mathbf{X}^{\text{T}}/p$ and $D$ be the unpenalized MLE of $\mathbf{\Delta}$, $D = \mathbf{X}^{\text{T}}\,\mathbf{X}/n$, then
  - $\hat{\mathbf{\Sigma}}_{\text{SE II}} = \rho\,\frac{\text{tr}(S)}{n}\mathbf{I} + (1-\rho)S, \quad \& \quad \hat{\mathbf{\Delta}}_{\text{SE II}} = \rho\,\frac{\text{tr}(D)}{p}\mathbf{I} + (1-\rho)D.$
  - $\hat{\mathbf{\Sigma}}_{\text{SE III}} = \rho\,\text{diag}(S) + (1-\rho)S, \quad \& \quad \hat{\mathbf{\Delta}}_{\text{SE III}} = \rho\,\text{diag}(D) + (1-\rho)D.$

These covariance estimates are taken in the spirit of shrinkage estimators presented by Daniels and Kass [1], which either shrink the eigenvalues (SE I) or shrink towards a specific structure(SE II and SE III).

According to the Kullback-Leibler distance, the $L_2 : L_2$ TRCM covariance estimates are more accurate than our comparison estimators and the $L_1 : L_1$ TRCM estimates even with underlying sparsity. This finding is not unexpected since from Theorem 1, we have a unique solution compared with the possibly many sub-optimal stationary points of the $L_1 : L_1$ TRCM solution, Proposition 2. From these results, we conjecture that the $L_2 : L_2$ TRCM estimates may lead to significant improvements in missing data imputation.

## 3. MCECM Algorithm Properties and Comparisons.

We present some properties and the rationale for structuring our imputation approach with transposable models as a Multi-Cycle ECM algorithm. The MCECM algorithm is a special case of the ECM algorithm of Meng and Rubin [11], where the Maximization step is split into several Conditional Maximization (CM) steps. Expectation steps are inserted between the CM steps to form multi-cycles. Our algorithm uses a specific type of conditional maximization, maximization with respect to one block of coordinates, either $\mathbf{\Sigma}^{\text{-1}}$ or $\mathbf{\Delta}^{\text{-1}}$. These ECM-type algorithms are, as Meng and Rubin say, a type of "extended GEM" (Generalized EM) algorithm [11], and thus retain many of the properties of GEM algorithms. Most importantly, our algorithm is monotonic, increasing the observed likelihood each step, and converges to a stationary point of the observed likelihood function.

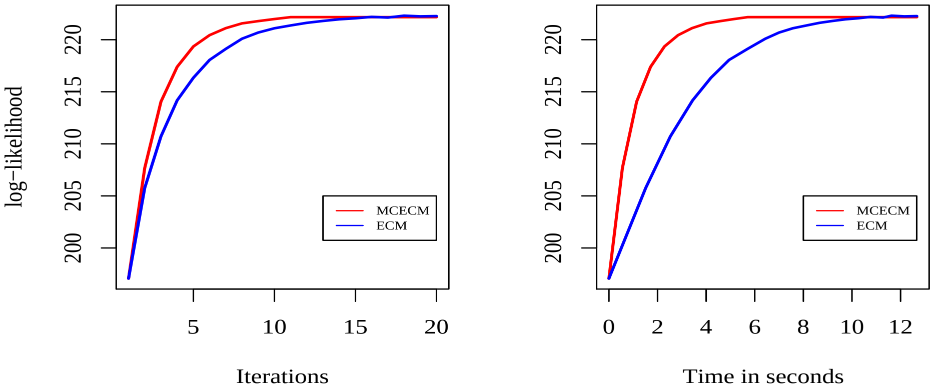

Fig 2. *Comparison of convergence of the observed log-likelihood by iterations (left) and by time in seconds (right) of the two structural approaches to the transposable EM algorithm: the Multi-cycle ECM and the ECM algorithms. The MCECM algorithm took 6.01 seconds of CPU time and the ECM algorithm took 12.375 seconds of CPU time until convergence. The data,* $25 \times 25$ *with 20% missing values, was simulated from the matrix-variate normal distribution with autoregressive covariances given in Section 3.*

We note that the ECM and also the MCECM algorithm can lead to a substantial computational savings in terms of the rate of convergence [10]. We investigate this possibility with simulated data in Figure 2 where we compare the convergence of the MCECM and the ECM approach to imputation with transposable data. The data, $25 \times 25$ with 20% of the data missing, is simulated from the matrix-variate normal with the autoregressive covariance parameters:

- Autoregressive: $\mathbf{\Sigma}_{ij} = 0.8^{|i-j|}$ and $\mathbf{\Delta}_{ij} = 0.6^{|i-j|}$.

For consistency, we used the $L_2 : L_2$ algorithm with $\rho_r = \rho_c = 1$ for both methods. From this simulation, we see that the MCECM algorithm converges in fewer iterations and in less time than the ECM algorithm. The MCECM algorithm reaches its maximum in 11 iterations, whereas the ECM algorithm needs 20 iterations. In our experience, the MCECM algorithm provides substantial computational savings in both real and simulated data.

## 4. Bayesian One-Step Approximation.

In discussing the computations involved in the Expectation step of the MCECM algorithm for imputation, we mentioned the possibility of extending the algorithm to a stochastic or stochastic approximation ECM-type algorithm. Here, we also note that the one-step approximation can be formulated as a stochastic one-step algorithm using Gibbs sampling. First, we present a blocked Gibbs sampler, Algorithm 2, for stochastically generating missing values from their posterior distribution, and then we discuss how this can be used in a Bayesian-type one-step approximation.

**Algorithm 2** Blocked Gibbs Sampler for estimating E step calculations in TRCMimpute or the final step of the Bayesian one-step approximation.

```
1. For each row, $i$, with missing values:
    (a) Generate $\mathbf{X}^{(k+1)}_{i,m_i}$ from $g\left(\mathbf{X}_{i,m_i} \,|\, \mathbf{X}^{(k)}_{i,o_i}, \mathbf{X}^{(k)}_{\neq i,r}\right)$, given in Theorem 2 (a).
2. For each column, $j$, with missing values:
    (a) Generate $\mathbf{X}^{(k+1)}_{m_j,j}$ from $g\left(\mathbf{X}_{m_j,j} \,|\, \mathbf{X}^{(k)}_{o_j,j}, \mathbf{X}^{(k)}_{c,\neq j}\right)$, given in Theorem 2 (b).
3. Repeat Steps 1 and 2 until a stationary distribution is reached.
```

The blocked Gibbs sampler generates all missing values in a row or a column as a group from their conditional distributions given in Theorem 2. Thus, a deterministic overlapping blocking scheme is used to update the elements that can lead to faster convergence [14]. This algorithm converges to the stationary distribution of the missing values given the observed values, and thus can be used in place of the Alternating Conditional Expectation Algorithm in the final step of the approximation. We call this the Bayesian one-step approximation. The conditional distribution of the missing values can also be thought of as the posterior distribution from which repeated draws can form a set of repeated imputations in the multiple imputation framework [15].

## 5. Computations for Alternating Conditional Expectations Algorithm.

The Alternating Conditional Expectations algorithm is the key component in the one-step approximation TRCMAimpute. Thus, computational costs are important especially when applying this imputation algorithm to high dimensional data. We show that using properties of the Schur complement, the order of operations can be reduced from $O(n^3 + p^3)$ to $O(\sum_{i=1}^{n} \min\{|m_i|, |o_i|\}^3 + \sum_{j=1}^{p} \min\{|m_j|, |o_j|\}^3)$, where $|m_i|$ and $|o_i|$ are the number of missing and observed elements of row $i$ respectively.

With both the $L_1$ and $L_2$ TRCM penalties, the covariance estimates and their inverses are easy to obtain from the computations of the graphical lasso [4] and the eigenvalue decompositions. We present the alternative forms using the Schur complements of the row covariance estimate and its inverse of part (i) in Theorem 2. The forms are analogous for the columns in part (ii). Given $\mathbf{\Sigma}$ and its inverse $\mathbf{\Theta}$, we have the following.

$$\begin{pmatrix} \mathbf{\Sigma}_{i,i} & \mathbf{\Sigma}_{i,k} \\ \mathbf{\Sigma}_{k,i} & \mathbf{\Sigma}_{k,k} \end{pmatrix} \begin{pmatrix} \mathbf{\Theta}_{i,i} & \mathbf{\Theta}_{i,k} \\ \mathbf{\Theta}_{k,i} & \mathbf{\Theta}_{k,k} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & \mathbf{I} \end{pmatrix}$$

Thus, $\mathbf{\Sigma}_{i,i} - \mathbf{\Sigma}_{i,k}\, \mathbf{\Sigma}^{-1}{}_{k,k}\, \mathbf{\Sigma}_{k,i} = 1/\,\mathbf{\Theta}_{i,i}$, and $\mathbf{\Sigma}_{i,k}\, \mathbf{\Sigma}_{k,k}^{-1} = -\,\mathbf{\Theta}_{i,k} \,/\, \mathbf{\Theta}_{i,i}$, meaning that $\mathbf{\Gamma} = \mathbf{\Delta} \,/\, \mathbf{\Theta}_{i,i}$ according to the notation on Theorem 2. If we let $\mathbf{\Psi}$ be the inverse of $\mathbf{\Delta}$ and partition it according to $m_i$ and $o_i$, then we have $\mathbf{\Gamma}_{m_i,m_i} - \mathbf{\Gamma}_{m_i,o_i}\mathbf{\Gamma}_{o_i,o_i}^{-1}\mathbf{\Gamma}_{o_i,m_i} = \mathbf{\Psi}_{m_i,m_i}^{-1} \,/\, \mathbf{\Theta}_{i,i}$ and $\mathbf{\Gamma}_{m_i,o_i}\mathbf{\Gamma}_{o_i,o_i} = -\,\mathbf{\Psi}_{m_i,m_i}^{-1}\, \mathbf{\Psi}_{m_i,o_i}$. Thus, in the second step of the theorem, if the number of missing elements in row $i$ is less than the number of observed elements, this alternative form requires less computation.

**6. Cross Validation for TRCMAimpute.** With real data, the TRCM penalty parameters $\rho_r$ and $\rho_c$ must be estimated from the data. To this end, we use 5-fold cross validation in all simulations and examples. This is accomplished by randomly deleting 20% of the observed values in each fold, applying an imputation method, and then measuring the imputation error on the deleted values. When selecting penalty parameters, we always include the possibility of an infinite value giving a marginal, multivariate model obtained in the first step of our approximation method. Thus, for application of TRCMAimpute, a model penalizing only the rows, only the columns, or both may be chosen by the cross-validation procedure. Also, we remind the reader that the four penalty types of the TRCM model are referred to as $L_{q_r} : L_{q_c}$, or the penalty-type placed on the rows and then the penalty-type on the columns.

In Figure 3, we give the estimated MSE from 5-fold cross validation and the true MSE. TRCMAimpute, $L_2 : L_2$ was used with an infinite penalty on the columns. This shows us that while, cross-validation may not estimate

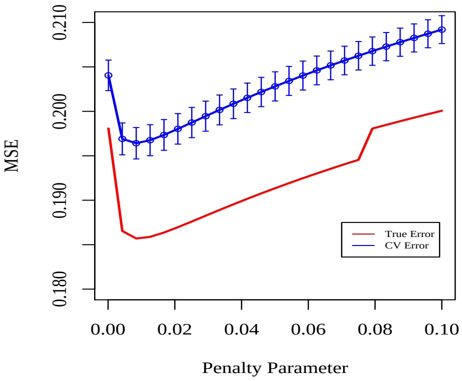

Fig 3. *True MSE and 5-fold cross-validation estimated MSE with standard errors. Genes from the kidney cancer microarray data with all samples observed are taken and deleted at random with 10% missing. TRCMAimpute, $L_2$ : $L_2$ is used with an infinite penalty on the column covariances. Cross-validation works well for determining the best penalty parameter.*

the true MSE well, it is fairly accurate as estimating the correct penalty parameter.

## 7. Proofs.

Proof of Proposition 1. Expanding the trace term of $\ell(\mathbf{M}, \boldsymbol{\Sigma}, \boldsymbol{\Delta})$ in terms of $\mu$ and $\nu$ and then taking partial derivatives, we get

$$\begin{aligned}\frac{\partial \ell}{\partial \nu} &= 2\,\boldsymbol{\Sigma}^{-1}\,\nu\mathbf{1}^T\,\boldsymbol{\Delta}^{-1} - 2\,\boldsymbol{\Sigma}^{-1}(\mathbf{X} - \mathbf{1}\mu^T)\,\boldsymbol{\Delta}^{-1} = 0\\ &\Rightarrow \hat{\nu}\mathbf{1}^T = \mathbf{X} - \mathbf{1}\mu^T\\ &\Rightarrow \hat{\nu} = \frac{\mathbf{1}^T(\mathbf{X} - \mathbf{1}\mu^T)}{p} = \sum_{j=1}^{p} \frac{X_{cj} - \mu_j}{p}\end{aligned}$$

A similar argument gives $\hat{\mu}$. □

Proof of Propoisition 2. Notice that the first three terms of $\ell(\boldsymbol{\Sigma}, \boldsymbol{\Delta})$ are differentiable and $\ell(\boldsymbol{\Sigma}, \boldsymbol{\Delta})$ is continuous. Then, since $\ell(\boldsymbol{\Sigma}, \boldsymbol{\Delta})$ is strictly concave in $\boldsymbol{\Sigma}^{-1}$ with $\boldsymbol{\Delta}^{-1}$ fixed and in $\boldsymbol{\Delta}^{-1}$ with $\boldsymbol{\Sigma}^{-1}$ fixed, maximization with respect to each gradient gives a unique coordinate-wise maximum. Thus, we have satisfied the conditions of Tseng, Theorem 4.1 (c) [16], and block coordinate-wise maximization converges to a stationary point of $\ell(\boldsymbol{\Sigma}, \boldsymbol{\Delta})$. □

Proof of Theorem 1. Our proof that the solution in Theorem 1 maximizes the penalized log-likelihood with $L_2$ penalties begins with the coordinate-wise gradients and shows that there is only one solution to these equations and it is thus the globally optimal solution. Throughout this proof we assume $\mathbf{X}$ is centered and let $\mathbf{X} = \mathbf{U}\,\mathbf{D}\,\mathbf{V}^{\mathrm{T}}$ be the SVD of $\mathbf{X}$ and $d = \mathrm{diag}(\mathbf{D})$.

First, we can take the eigenvectors of $\mathbf{\Sigma}^*$ and $\mathbf{\Delta}^*$ to be the left and right singular vectors of $\mathbf{X}$ respectively, because rearranging the gradients, gives

$$p\,\mathbf{\Sigma} - 4\rho_r\,\mathbf{\Sigma}^{-1} = \mathbf{X}\,\mathbf{\Delta}^{-1}\,\mathbf{X}^{\mathrm{T}}$$
$$n\,\mathbf{\Delta} - 4\rho_c\,\mathbf{\Delta}^{-1} = \mathbf{X}^{\mathrm{T}}\,\mathbf{\Sigma}^{-1}\,\mathbf{X}\,.$$

Thus, the eigenvectors of $\mathbf{\Sigma}$ and $\mathbf{\Delta}$ must be equal to their respective quadratic forms. This gives only one solution for the eigenvectors which is the left and right singular vectors of $\mathbf{X}$. Note that if $\mathrm{rank}(\mathbf{X}) = r$, then the last $n-r$ eigenvectors of $\mathbf{U}$ and the last $p-r$ of $\mathbf{V}$ are not unique.

Now, given the eigenvectors of $\mathbf{\Sigma}^*$ and $\mathbf{\Delta}^*$, we can write our penalized log-likelihood function in terms of the eigenvalues $\beta$ and $\theta$ and the singular values $d$.

$$-\ell(\beta,\theta) = \frac{p}{2}\sum_{i=1}^{n}\log\beta_i + \frac{n}{2}\sum_{j=1}^{p}\log\theta_j + \frac{1}{2}\sum_{j=1}^{p}\frac{d_j^2}{\beta_j\theta_j} + \rho_r\sum_{i=1}^{n}\frac{1}{\beta_i^2} + \rho_c\sum_{j=1}^{p}\frac{1}{\theta_j^2}. \tag{2}$$

Note that this is a biconvex function of $\frac{1}{\beta}$ and $\frac{1}{\theta}$. Hence, we can use sequential convex programming to minimize (2), but this may not converge to the global minimum, which we seek. Instead, we use the coordinate-wise gradients of (2), adding $n-p$ zeros to $\theta$ and $d$ so they are of length $n$. They are

$$\begin{aligned} p\theta_i\beta_i^2 - d_i^2\beta_i - 4\rho_r\theta_i &= 0 \qquad \text{for } i = 1\ldots n. \\ n\beta_i\theta_i^2 - d_i^2\theta_i - 4\rho_c\beta_i &= 0 \qquad \text{for } i = 1\ldots n. \end{aligned} \tag{3}$$

Note that the global minimum must satisfy (3). The last $n-r$ values of $\beta^*$ are immediately known and equal to $2\sqrt{\frac{\rho_r}{p}}$. Also, if $p > r$, then the last $p-r$ values of $\theta^*$ are $2\sqrt{\frac{\rho_c}{n}}$. So, we concentrate on the first $r$ values of $\theta$ and $\beta$.

Since the gradients, (3), are quadratic equations, we can write it as one equation in terms of $\beta_i$ by letting $\theta_i = \frac{d_i^2\beta_i}{p\beta_i^2 - 4\rho_r}$. This gives us the following fifth degree polynomial.

$$\beta_i\left(c_1\beta_i^4 + c_2\beta_i^2 + c_3\right) = 0 \tag{4}$$

for each eigenvalue indexed by $i$ with coefficients

$$c_1^{(i)} = -4\rho_c p^2, \quad c_2^{(i)} = 32\rho_r\rho_c p + d_i^4(n-p), \ \ \& \ \ c_3^{(i)} = 4\rho_r(d_i^4 - 16\rho_r\rho_c).$$

The five solutions to (4) are 0 and $\pm\sqrt{\frac{-c_2^{(i)} \pm \sqrt{{c_2^{(i)}}^2 - 4c_1^{(i)}c_3^{(i)}}}{2c_1^{(i)}}}$. But, we are looking for solutions that are real and positive in terms of both $\beta_i$ and $\theta_i$. Thus, we can immediately dismiss the zero solution and the two negative solutions. Now, notice that if $d_i^4 \geq 16\rho_r\rho_c$, then $c_3^{(i)} \geq 0$, and thus by Descartes' sign rule, there exists only one positive real root. This root is given by $\beta_i = \sqrt{\frac{-c_2^{(i)} - \sqrt{{c_2^{(i)}}^2 - 4c_1^{(i)}c_3^{(i)}}}{2c_1^{(i)}}}$. We now consider the case when $d_i^4 < 16\rho_r\rho_c$ and we have two positive real roots. The above root is obviously still feasible in this case. We check the other positive root given by $\beta_i = \sqrt{\frac{-c_2^{(i)} + \sqrt{{c_2^{(i)}}^2 - 4c_1^{(i)}c_3^{(i)}}}{2c_1^{(i)}}}$ to see if the corresponding $\theta_i > 0$. The numerator of $\theta_i$ is always strictly positive, so for $\theta_i$ to be feasible, $p\beta_i^2 > 4\rho_r$. Substituting in the possible root, $\beta_i$, we see that this inequality does not hold. Hence, this root is infeasible, leaving us with only one feasible root.

Therefore, we have only one feasible solution $(\beta_i, \theta_i)$ for each $i$ to the gradient equations (3). Since the global solution must satisfy these gradient equations, we conclude that the root is the unique global solution $(\beta^*, \theta^*)$. We comment here that using iterative coordinate descent to solve the quadratic equations (3) by taking the positive roots, converges to this globally optimal solution. This is true because any solution to the coordinate procedure must satisfy the coordinate-wise gradient conditions, which we have just proven to have only one solution. Thus, this is a rare instance when the coordinate descent solution to a biconvex problem converges to the global minimum. □

Proof of Proposition 5.

$$
\begin{aligned}
\mathrm{E}_{(\boldsymbol{\Sigma},\boldsymbol{\Delta})}\left[\ell(\boldsymbol{\Sigma},\boldsymbol{\Delta})-\ell(\hat{\boldsymbol{\Sigma}},\hat{\boldsymbol{\Delta}})\right] &= \frac{p}{2}\left(\log|\,\boldsymbol{\Sigma}^{\text{-}1}\,|-\log|\widehat{\boldsymbol{\Sigma}^{\text{-}1}}|\right)+\frac{n}{2}\left(\log|\,\boldsymbol{\Delta}^{\text{-}1}\,|-\log|\widehat{\boldsymbol{\Delta}^{\text{-}1}}|\right)\\
&\quad -\frac{1}{2}\mathrm{E}_{(\boldsymbol{\Sigma},\boldsymbol{\Delta})}\left[\mathrm{tr}\left(\boldsymbol{\Sigma}^{\text{-}1}\,\mathbf{X}\,\boldsymbol{\Delta}^{\text{-}1}\,\mathbf{X}^{\mathrm{T}}\right)-\mathrm{tr}\left(\widehat{\boldsymbol{\Sigma}^{\text{-}1}}\,\mathbf{X}\,\hat{\boldsymbol{\Delta}}^{-1}\,\mathbf{X}^{\mathrm{T}}\right)\right]\\
&= \frac{p}{2}\log|\hat{\boldsymbol{\Sigma}}\,\boldsymbol{\Sigma}^{\text{-}1}\,|+\frac{n}{2}\log|\hat{\boldsymbol{\Delta}}\,\boldsymbol{\Delta}^{\text{-}1}\,|\\
&\quad -\frac{1}{2}\left[\mathrm{tr}\left(\boldsymbol{\Sigma}^{\text{-}1}\,\mathrm{E}_{(\boldsymbol{\Sigma},\boldsymbol{\Delta})}\left[\mathbf{X}\,\boldsymbol{\Delta}^{\text{-}1}\,\mathbf{X}^{\mathrm{T}}\right]\right)-\mathrm{tr}\left(\widehat{\boldsymbol{\Sigma}^{\text{-}1}}\mathrm{E}_{(\boldsymbol{\Sigma},\boldsymbol{\Delta})}\left[\mathbf{X}\,\hat{\boldsymbol{\Delta}}^{-1}\,\mathbf{X}^{\mathrm{T}}\right]\right)\right]\\
&= \ldots-\frac{1}{2}\left[\mathrm{tr}\left(\boldsymbol{\Sigma}^{\text{-}1}\left[\mathrm{tr}(\boldsymbol{\Delta}^{\text{-}1}\,\boldsymbol{\Delta})\,\boldsymbol{\Sigma}\right]\right)-\mathrm{tr}\left(\hat{\boldsymbol{\Sigma}}^{-1}\left[\mathrm{tr}(\boldsymbol{\Delta}\,\hat{\boldsymbol{\Delta}}^{-1})\,\boldsymbol{\Sigma}\right]\right)\right]\\
&= \ldots-\frac{1}{2}\left[\mathrm{tr}(\boldsymbol{\Sigma}^{\text{-}1}\,\boldsymbol{\Sigma})\mathrm{tr}(\boldsymbol{\Delta}^{\text{-}1}\,\boldsymbol{\Delta})-\mathrm{tr}(\hat{\boldsymbol{\Sigma}}^{-1}\,\boldsymbol{\Sigma})\mathrm{tr}(\hat{\boldsymbol{\Delta}}^{-1}\,\boldsymbol{\Delta})\right]\\
&= \frac{p}{2}\log|\hat{\boldsymbol{\Sigma}}\,\boldsymbol{\Sigma}^{\text{-}1}\,|+\frac{n}{2}\log|\hat{\boldsymbol{\Delta}}\,\boldsymbol{\Delta}^{\text{-}1}\,|-\frac{np}{2}+\frac{1}{2}\mathrm{tr}\left(\widehat{\boldsymbol{\Sigma}^{-1}}\,\boldsymbol{\Sigma}\right)\mathrm{tr}\left(\widehat{\boldsymbol{\Delta}^{-1}}\,\boldsymbol{\Delta}\right).
\end{aligned}
$$

Note that $\mathrm{E}_{(\boldsymbol{\Sigma},\boldsymbol{\Delta})}\left[\mathbf{X}\,\mathbf{A}\,\mathbf{X}^{\mathrm{T}}\right]=\mathrm{tr}(\boldsymbol{\Delta}\,\mathbf{A}^{T})\,\boldsymbol{\Sigma}$ where $\mathbf{A}$ is $p\times p$ [7]. □

Proof of Proposition 4. We first show that $\mathrm{E}\left[\mathrm{tr}\left(\mathbf{X}^{\mathrm{T}}\,\boldsymbol{\Sigma}^{\text{-}1}\,\mathbf{X}\,\boldsymbol{\Delta}^{\text{-}1}\right)|X_o,\theta'\right]$ $=\mathrm{tr}\left[\left(\hat{\mathbf{X}}^{T}\,\boldsymbol{\Sigma}^{\text{-}1}\,\hat{\mathbf{X}}+\mathbf{G}(\boldsymbol{\Sigma}^{\text{-}1})\right)\boldsymbol{\Delta}^{\text{-}1}\right]$.
Let $\mathbf{A}=\mathbf{X}^{\mathrm{T}}\,\boldsymbol{\Sigma}^{\text{-}1}\,\mathbf{X}$, then,

$$
\mathrm{E}\left[\mathrm{tr}\left(\mathbf{X}^{\mathrm{T}}\,\boldsymbol{\Sigma}^{\text{-}1}\,\mathbf{X}\,\boldsymbol{\Delta}^{\text{-}1}\right)|X_o,\theta'\right]=\mathrm{tr}\left[\mathrm{E}\left(\mathbf{A}\,|X_o,\theta'\right)\boldsymbol{\Delta}^{\text{-}1}\right]
$$

$$
\begin{aligned}
\text{And, }\mathrm{E}(\mathbf{A}_{jj'}\,|X_o\theta') &= \mathrm{E}\left(X_{cj}^{T}\,\boldsymbol{\Sigma}^{\text{-}1}\,X_{cj'}|X_o,\theta'\right)\\
&= \mathrm{E}\left[\sum_{k=1}^{n}\sum_{t=1}^{n}x_{tj}x_{kj'}\sigma_{tk}^{-1}|X_o,\theta'\right]\\
&= \sum_{k=1}^{n}\sum_{t=1}^{n}\hat{x}_{tj}\hat{x}_{kj'}\sigma_{tk}^{-1}+\sum_{k=1}^{n}\sum_{t=1}^{n}C_{tk}^{(jj')}\sigma_{tk}^{-1}\\
&= \hat{X}_{cj}^{T}\,\boldsymbol{\Sigma}^{\text{-}1}\,\hat{X}_{cj'}+\mathrm{tr}\left(\mathbf{C}^{(jj')}\,\boldsymbol{\Sigma}^{\text{-}1}\right)
\end{aligned}
$$

Thus, $\mathrm{E}\left(\mathbf{A}\,|X_o,\theta'\right)=\hat{\mathbf{X}}^{T}\,\boldsymbol{\Sigma}^{\text{-}1}\,\hat{\mathbf{X}}+\mathbf{G}(\boldsymbol{\Sigma}^{\text{-}1})$.
The calculation showing $\mathrm{E}\left[\mathrm{tr}\left(\mathbf{X}^{\mathrm{T}}\,\boldsymbol{\Sigma}^{\text{-}1}\,\mathbf{X}\,\boldsymbol{\Delta}^{\text{-}1}\right)|X_o,\theta'\right]=\mathrm{tr}\left[\left(\hat{\mathbf{X}}\,\boldsymbol{\Delta}^{\text{-}1}\,\hat{\mathbf{X}}^{T}+\mathbf{F}(\boldsymbol{\Delta}^{\text{-}1})\right)\boldsymbol{\Sigma}^{\text{-}1}\right]$ is the analogous to the above calculation with $\mathbf{B}=\mathbf{X}\,\boldsymbol{\Delta}^{\text{-}1}\,\mathbf{X}^{\mathrm{T}}$ and

$$
\mathrm{E}\left(\mathbf{B}_{ii'}\,|X_o,\theta'\right)=\hat{X}_{ir}\,\boldsymbol{\Delta}^{\text{-}1}\,\hat{X}_{i'r}^{T}+\mathrm{tr}\left(\mathbf{D}^{(ii')}\,\boldsymbol{\Delta}^{\text{-}1}\right).
$$

□

Proof of Theorem 2. Notice that Theorem 2 gives the conditional distributions in a two step process. The first step involves finding the distribution of a row or column conditional on the rest of the matrix. Using the notation given, this means $(\mathbf{X}_{i,r} \,|\, \mathbf{X}_{k,r}) = (\mathbf{X}_{i,m_i}, \mathbf{X}_{i,o_i} \,|X_{k,r}) \sim N(\psi, \mathbf{\Gamma})$ and $(\mathbf{X}_{c,j} \,|\, \mathbf{X}_{c,l}) = (\mathbf{X}_{m_j,j}, \mathbf{X}_{o_j,j} \,|X_{c,l}) \sim N(\eta, \mathbf{\Phi})$. These are given in Gupta and Nagar [7]. Given the conditional distributions of a row or column, the second step gives the distribution of a set of elements, $m_i$ or $m_j$ within a row or column respectively. Notice that these conditional distributions come directly from the conditional distribution formulas for partitioned multivariate normal matrices, given the partitions of $\psi$, $\eta$, $\mathbf{\Gamma}$ and $\mathbf{\Phi}$. Hence, the connection between the two steps is all that is needed. We show this for part (a) and the proof for part (b) is analogous.

If we let $W = \mathbf{X}_{i,m_i}$, $Y = \mathbf{X}_{i,o_i}$ and $Z = \mathbf{X}_{\neq i,r}$, and assume that they are centered so that $\text{vec}(W \;\; Y \;\; Z) \sim N(\mathbf{0}, \mathbf{\Omega})$, then notice that the first step gives $p\left((W,Y)|Z\right)$ and the second step gives $p\left((W|Z)|(Y|Z)\right)$. We show that this gives the desired conditional distribution.

$$\begin{aligned} p\left((W|Z)|(Y|Z)\right) &= \frac{p\left((W|Z),(Y|Z)\right)}{p\left(Y|Z\right)} \\ &= \frac{p\left(W,Y|Z\right)}{p(Y|Z)} \\ &= p\left(W|Y,Z\right). \end{aligned}$$

Since matrix normal random variables can be written as multivariate normal random variables, we have established that the two steps give the desired conditional distribution. □

Proof of Theorem 3. We will show that the iterates in Steps 2 and 3 of the Alternating Conditional Expectations Algorithm are the block Gauss-Seidel iterates solving the linear system giving $\mathrm{E}\left(\mathbf{X}_m \,|\, \mathbf{X}_o\right)$. For simplicity, we consider $\mathbf{X}_{2\times 2}$ with $\text{vec}(\mathbf{X}) = [x_1 \;\; x_2 \;\; x_3 \;\; x_4]^T \sim N(\mathbf{0}, \mathbf{\Omega})$ where $x_1$, $x_2$ are missing and $x_3$, $x_4$ are observed. If we partition $\mathbf{\Omega}$ according to indices of $\text{vec}(\mathbf{X})$, the $k+1$ iteration of Step 2 or 3 is then

$$\begin{aligned} x_1^{(k+1)} &= \mathbf{C}\left[x_2^{(k)} \;\; x_3 \;\; x_4\right]^T \\ &\qquad \text{where } \mathbf{C}_{1\times 3} = \mathbf{\Omega}_{x_1,\neq x_1}\, \mathbf{\Omega}^{-1}_{\neq x_1,\neq x_1} \\ x_2^{(k+1)} &= \mathbf{D}\left[x_1^{(k)} \;\; x_3 \;\; x_4\right]^T \\ &\qquad \text{where } \mathbf{D}_{1\times 3} = \mathbf{\Omega}_{x_2,\neq x_2}\, \mathbf{\Omega}^{-1}_{\neq x_2,\neq x_2} \end{aligned} \tag{5}$$

Define $\mathbf{A} = \begin{pmatrix} 1 & -\mathbf{C}_1 \\ -\mathbf{D}_1 & 1 \end{pmatrix}$ and $b = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$ where $b_1 = \mathbf{C}_{2:3} \left[x_3 \;\; x_4\right]^T$ and $b_1 = \mathbf{D}_{2:3} \left[x_3 \;\; x_4\right]^T$. Then, solving the linear system $\mathbf{A} \left[x_1 \;\; x_2\right]^T = b$ gives the conditional expectation, $\mathbf{A}^{-1} b = \mathrm{E}\left(\left[x_1 \;\; x_2\right]^T \mid \left[x_3 \;\; x_4\right]^T\right) = \boldsymbol{\Omega}_{1:2,3:4} \, \boldsymbol{\Omega}_{3:4,3:4}^{-1} \left[x_3 \;\; x_4\right]^T$. In addition, the Gauss-Seidel iterates,

$$\begin{aligned} x_1^{(k+1)} &= \left(b_1 - a_{12} x_2^{(k)}\right) / a_{11} \\ x_2^{(k+1)} &= \left(b_2 - a_{21} x_1^{(k)}\right) / a_{22} \end{aligned}$$

give back equations (5). Also note that since $\boldsymbol{\Omega}$ is positive definite, so is $\mathbf{A}$, ensuring convergence of the algorithm. Thus, finding the conditional expectation of each individual missing value in an iterative fashion converges to the conditional expectation of all missing values given the observed data. Steps 2 and 3 of the Alternating Conditional Expectation Algorithm, however, group the missing values by row or by column. Thus, these steps form the block Gauss-Seidel iterates which also converge, often with a faster rate of convergence [8]. In addition, note that the groups of missing values by row and by column form overlapping blocks, or multi-splittings. These updating schemes have been shown to often converge faster than non-overlapping schemes in linear iterative methods [5].

□

## References.

[1] M. J. Daniels and R. E. Kass. Shrinkage estimators for covariance matrices. *Biometrics*, 57(4):1173–1184, 2001.

[2] J. A. Fessier and A. O. H. III. Penalized maximum-likelihood image reconstruction using space-alternating generalized em algorithms. *IEEE Transactions on Image Processing*, 4(10):1417–1429, 1995.

[3] J. A. Fessler and A. O. H. III. Space-alternating generalized expectation-maximization algorithm. *IEEE Transactions on Signal Processing*, 42(10):2664–2677, 1994.

[4] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the lasso. *Biostatistics*, 9(3):432–441, 2007.

[5] A. Frommer and B. Phol. A comparison result for multisplittings and waveform relaxation methods. *Numerical Linear Algebra with Applications*, 2(4):335–346, 1995.

[6] P. J. Green. On use of the em for penalized likelihood estimation. *J. of the Royal Statisical Society*, 52(3):443–452, 1990.

[7] A. K. Gupta and D. K. Nagar. *Matrix variate distributions*. CRC Press, 1999.

[8] P. J. Lanzkron, D. J. Rose, and D. B. Szyld. Convergence of nested classical iterative methods for linear systems. *Numerical Mathematics*, 58:685–702, 1991.

[9] R. J. A. Little and D. B. Rubin. *Statistical analysis with missing data*. Wiley-Interscience, 2002.

[10] X.-L. Meng. On the rate of convergence of the ecm algorithm. *Annals of Statistics*, 22(1):326–339, 1994.

[11] X.-L. Meng and D. Rubin. Maximum likelihood estimation via the ecm algorithm: A general framework. *Biometrika*, 80(2):267–278, 1993.

[12] T. Orchard and M. A. Woodbury. A missing information principle: theory and applications. *Proc. Sixth Berkeley Symp. on Math. Statist. and Prob.*, 1:697–715, 1972.

[13] A. R. D. Pierro. Modified expectation maximization algorithm for penalized likelihood estimation in emission tomography. *IEEE Transactions on medical imaging*, 14(1), 1995.

[14] G. O. Roberts and S. K. Sahu. Updating schemes, correlation structure, blocking and parameterization for the gibbs sampler. *J. R. Statist. Soc.*, 59(2):291–317, 1997.

[15] D. B. Rubin. Multiple imputation after 18+ years. *J. Amer. Statist. Assoc.*, 91(434):473–489, 1996.

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