Title
A Comparison of Imputative Capability on Algorithms for fitting the PARAFAC model to Biological Data

Author
Hodzic, Enio

Publication Date
2023

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
Los Angeles

A Comparison of Imputative Capability on Algorithms for fitting the PARAFAC model to Biological Data

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Science in Bioengineering

by

Enio Hodzic

2023

ABSTRACT OF THE THESIS

A Comparison of Imputative Capability
on Algorithms for fitting the PARAFAC model
to Biological Data

by

Enio Hodzic
Master of Science in Bioengineering
University of California, Los Angeles, 2023
Professor Aaron S. Meyer, Chair

Researchers often find it challenging to simplify complex data sets for downstream analysis. The combination of multiple variables can lead to complexity, and organizing such data sets into a higher dimensional structure can be more intuitive. For these data sets with an inherent multi-modal structure, a variety of dimensionality reduction techniques have allowed researchers to explore and infer biological interactions more effectively. Higher-order dimensionality reduction techniques all serve to accomplish the same purpose - to reduce the original data set and recover meaningful and interpretable patterns. The CANDECOMP/PARAFAC (CP) model, a frequent choice among researchers for its interpretability, still requires metrics for validating its performance and assuring an appropriate model complexity is selected. While a common benchmark for these methods' validation is typically the total residual error, imputation error (prediction error) can serve as a more trusted alternative. We describe an algorithm for fitting the PARAFAC model, censored alternating least squares, that innately handles missing values and compare it amongst alternating least

squares and direct optimization using simulated and real data sets with varying degrees of missing values using these performance metrics. While each method has its own benefits, censored alternating least squares appears best suited for handling missing values, commonly present in the data that researchers look to investigate.

The thesis of Enio Hodzic is approved.

Harold Pimentel

Deanna Needell

Jennifer L. Wilson

Aaron S. Meyer, Committee Chair

University of California, Los Angeles

2023

To my parents, Seid and Diana

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

I would like to thank my mentor, Dr. Aaron Meyer, for his assistance and guidance with my work. His deep understanding of the topics I researched were paramount in my learning and growth as an academic. I also thank Ph.D student Cyrillus Tan for his continual guidance and assistance in many technical and theoretical questions I had. Finally, I thank undergraduate Ethan Hung for his technical contributions.

CHAPTER 1

Introduction

Biologists frequently encounter complex and obscure data sets leading to a variety of data analysis approaches. Since biological systems often involve multiple confounding factors, researchers typically rely on dimensionality reduction for analyzing these types of data sets effectively. While for simple data sets involving two variables, conventional matrix factorization methods are sufficient, data structures involving multiple variables have the capability to be structured in a multi-modal fashion (arranged in several dimensions, such as measurement, subject, and time). While offering an intuitive way of organizing the data structure, analysis is frequently difficult when coupled with pre-existing missing values commonly seen within biological data sets. Types of missingness patterns encountered within these types of data sets can vary, from random scattering across the entirety of the data set to individual cords depending on the source of the missing values [1]. These missing cords correspond to structured missing values occurring along a specific variable, such as a missing sample along a temporal dimension.

Dimensionality reduction methods like Principal Component Analysis (PCA) and other matrix factorization approaches are powerful tools for exploring multivariate data by providing a means of capturing patterns specific to each mode [2, 3]. Multi-modal data sets, otherwise known as tensors, require an initial unfolding for matrix factorization approaches. Two axes of the tensor are collapsed and combined to matricize the data set, concealing any relationships between the two modes being collapsed. Tensor decomposition methods such as CANDECOMP/PARAFAC (CP) and Tucker improve upon matrix methods by retaining

the inherent multi-modal structure [4]. CP decomposition methods are particularly relevant due to their ease of factor interpretation, being represented by the product of mode-specific factors similar to PCA. In the case of biological data, factors can infer a variety of relationships such as mechanistic binding properties and genome expression data [5, 6]. In the presence of pre-existing missing values, tensor factors have the capability to impute these entries based on the existing data. Furthermore, these tensor factors not only reduce noise in the original data but are capable of elucidating patterns when overlapping measurements or co-dependent processes are involved – such patterns might otherwise be lost when flattening data for matrix factorization methods [4]. As a result, these decomposed factors reveal unseen relationships between modes while closely approximating the original data set.

Assessing the validity of the factorization method is a necessity to assure the factor interpretation is sound and the appropriate number of components are selected. Various methods exist for selecting component number, from residual based methods to core consistency metrics [7]. A simple approach involves a predefined error metric between the original data set and reconstructed data set using the computed factors. Model rank can be chosen through a scree test, where introducing additional components yield negligible improvement to the overall model fit. In the biological context, a low fitted error between original data sets signifies an appropriate factor representation of the biological data at hand. It may not however confirm any biological interactions between factors, diminishing the analytical value of the method. A more appropriate error metric arises from the artificial insertion of missing values, recomputing the factors, and calculating the error, akin to cross validation for tensor completion [8]. Due to the nature of tensor factorization, factors are able to be solved with the presence of missing values within the data set. The minimal differences between removed values and imputed values from the tensor factors indicates that the biological interpretation is in fact contained within the factor plots while also providing a means to choose an appropriate tensor rank.

Various algorithms exist for solving tensor factors each with varying effectiveness in

imputative accuracy. The most common approach is alternating least squares (ALS), in which each factor of the tensor is solved while the others are held constant, and subsequently alternating across all dimensions of the tensor [9]. The "alternating" name stems from solving independent least squares problems for each mode of the tensor in an iterative fashion using the factor matrices and the unfolded tensor. However, ALS can lead to sensitive solutions affected by minor perturbations [10]. As this method requires a full data set, missing values are initialized with some value impacting factor solving. An alternative method is through directly optimizing all factors simultaneously through gradient based optimization (DO). Rather than solving each factor in an alternating fashion, a parameter vector is built including the values of all factors and its gradient subsequently computed [11]. However, gradient based methods are prone to slow or non-convergence and numerical uncertainty [12, 13, 14]. Censored alternating least squares (CLS) addresses missing values using a methodology similar to that of alternating least squares. However, it tackles each least squares problem independently by considering the missing patterns within the data structure. While this method avoids the influence of missing values on factor calculation, it involves additional computations within each iteration of the algorithm.

In the present study, we compared these algorithms against one another to assess imputation accuracy and convergence time. Each tensor model was trained on both simulated data and real biological data to test imputative capability. The tensor model trained using CLS was found to yield the best overall imputation accuracy compared to DO and ALS with data sets containing low noise and few missing values. For data sets with high noise and a high fraction of missing values, the benefits of CLS diminish and is prone to sensitive factors if regularization is not performed or initialization is poor.

CHAPTER 2

Methods

2.1 PARAFAC Model

For a three-mode tensor $\mathbf{T} \in \mathbb{R}^{I \times J \times K}$, the PARAFAC model [15, 16] is defined as

$$t_{ijk} = \sum_{r=1}^{R} a_{ir} b_{jr} c_{kr} + r_{ijk} \tag{2.1}$$

where R is the total number of factors, r_{ijk} are the residuals, and a_{ir}, b_{jr}, c_{kr} are the parameters being estimated. The model is solved by finding the factor matrices **A, B, C** that minimize the following loss function:

$$L(a_{11}, a_{12}, \ldots, c_{KR}) = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} (t_{ijk} - \sum_{r=1}^{R} a_{ir} b_{jr} c_{kr})^2 \tag{2.2}$$

Note that the factor matrices, $\mathbf{A}, \mathbf{B}, \mathbf{C}$, include all of the parameters in matrix form. The PARAFAC model can be rewritten in the form of a matricised format for tensor by introducing the column-wise Khatri Rao product $\odot$ [17] as

$$\mathbf{T}^{I \times JK} = \mathbf{A}(\mathbf{C} \odot \mathbf{B})^T + \mathbf{R}^{I \times JK} \tag{2.3}$$

2.2 Simulated Data Generation

Simulated data was generated through first generating the factor values based on a predefined probability distribution to obtain known rank random data sets. Three factors were generated to obtain a final simulated data set of known rank and dimension using the following method. For each entry in the factor matrices $\mathbf{A}, \mathbf{B}, \mathbf{C}$ decomposing tensor $\mathbf{T}$ with R components of size $I \times J \times K$

$$\begin{aligned} \mathbf{A} &\in \mathbb{R}^{I \times R} \\ \mathbf{B} &\in \mathbb{R}^{J \times R} \\ \mathbf{C} &\in \mathbb{R}^{K \times R} \end{aligned} \tag{2.4}$$

draw random samples from a gamma distribution of shape k and scale θ for each entry of the factor matrices. The gamma distribution was chosen to model biological processes that involve independent events occurring at a constant rate. Noise was added through an additional random variable sampled from a normal distribution and scaled if necessary. For a single entry of $\mathbf{A}$ this is written as

$$a_{ij} = Gamma(k, \theta) + \eta\mathcal{N}(0, 1) \tag{2.5}$$

Matrices $\mathbf{B}, \mathbf{C}$ are handled identically. A reconstructed tensor can be obtained using Eq. (2.1). All simulated data was generated using a shape parameter of 1 and scale parameter of 1.

2.3 Initial Data set and Artificial Missingness

Missing values were added to the data set in two separate ways: random entry removal and random cord removal (Fig. 1). The percentage missingness value corresponds to the ratio of data entry indices selected to be removed to all possible indices that can be removed.

For random entry removal take all possible indices for the tensor $\mathbf{T}$ of size $I \times J \times K$ denoted as set S

$$S = \{(x, y, z) \mid x, y, z \in (I \times J \times K)\} \tag{2.6}$$

Let X be a random variable of a discrete uniform distribution over the set S sampled N times without replacement. With a chosen percent missingness, N is simply the the cardinality of S multiplied with the percent missingness and rounded if needed. Thus we obtained a new set S', a subset of S that contains the indices from tensor $\mathbf{T}$ that were removed.

Cords are specific pairs of indices of the tensor that include all entries along an axis. For cord removal, we picked a certain mode to remove along tensor $\mathbf{T}$. In a similar fashion to entry removal, cords were removed along the chosen mode until the percent missingness was satisfied.

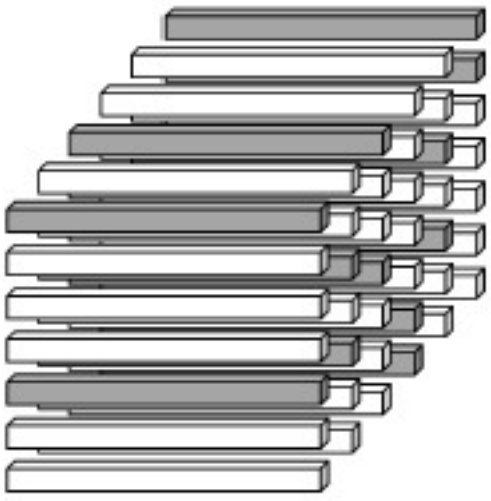

Figure 1: Visualization of cord removal from a tensor. Grey cords are artificially removed from the tensor and replaced with NaN values.

2.4 Error Metrics

In order to assess each algorithm's performance on the known data, the relative fitting error was calculated at each iteration using the reconstructed tensor and the original data. For each entry not artificially removed, the squared error was calculated with the original tensor $\mathbf{T}$ and the reconstructed tensor $\hat{\mathbf{T}}$ and relatively scaled. The mask tensor $\mathbf{M}$ serves as a weight tensor where an element that is missing is set as zero (T_{ijk} is missing, $M_{ijk} = 0$) and an element that is observed is set as one (T_{ijk} exists, $M_{ijk} = 1$).

$$v = \frac{\|\mathbf{M} * (\mathbf{T} - \hat{\mathbf{T}})\|_F^2}{\|\mathbf{M} * \mathbf{T}\|_F^2} \tag{2.7}$$

The imputation error was calculated the same way, except using the known values prior to their removal from $\mathbf{T}$ (T_{ijk} is missing, $M_{ijk} = 1$).

2.5 Factor Initialization

Providing an accurate initial guess is crucial when solving the factors using ALS, CLS, or DO. Each factor was initialized by iteratively solving the truncated SVD of the unfolded tensor with missing values initially set as zero and imputed thereafter. The factor matrix was initialized as the left singular vectors of the SVD of the flattened tensor upon convergence. In the case where the rank being solved is greater than the dimension of the flattened tensor, the remainder of the columns were set as ones.

$$\mathbf{A}(\mathbf{B} \odot \mathbf{C})^T = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \tag{2.8}$$

Alternatively, random initialization was also used where indicated. In this case, the factor matrices were initialized using random values from a standard normal distribution.

2.6 Alternating Least Squares (ALS)

The alternating least squares approach is the primary method used for fitting the PARAFAC model. Its procedure is described extensively [18, 17]. Referring to the previous definitions of the PARAFAC model (2.1, 2.3) and loss function (2.2), the loss function can be rewritten using the factor matrices as

$$L(\mathbf{A}, \mathbf{B}, \mathbf{C}) = \|\mathbf{T}^{I \times JK} - \mathbf{A}(\mathbf{C} \odot \mathbf{B})^T\|_F^2 \tag{2.9}$$

Calculating all three factor matrices that minimize L simultaneously proves rather difficult. However, if initial approximations are available for $\mathbf{B}$ and $\mathbf{C}$, then $\mathbf{A}$ can be calculated by least squares as

$$\mathbf{A}^{(1)} = \mathbf{T}^{I \times JK}((\mathbf{C}^{(0)} \odot \mathbf{B}^{(0)})^+)^T \tag{2.10}$$

$\mathbf{B}$ and $\mathbf{C}$ are done in a similar fashion for their solved interim factor matrices

$$\begin{aligned} \mathbf{B}^{(1)} &= \mathbf{T}^{J \times IK}((\mathbf{A}^{(1)} \odot \mathbf{C}^{(0)})^+)^T \\ \mathbf{C}^{(1)} &= \mathbf{T}^{K \times IJ}((\mathbf{B}^{(1)} \odot \mathbf{A}^{(1)})^+)^T \end{aligned} \tag{2.11}$$

These three alternating steps are repeated until a predefined convergence criterion is met and solved matrices are obtained. In the case of missing values encountered within the tensor, the entries were initially set as zero. After each iteration, the missing values were imputed using the factor matrices. We used Tensorly's ALS implementation and tracked metrics throughout each iteration [19], terminated either at a tolerance of 10^{-7} or at 50 iterations.

2.7 Direct Optimization (DO)

An alternative to the ALS approach is through solving all factor matrices simultaneously through a gradient based optimization approach [11]. Each of the factor matrices $(\mathbf{A}, \mathbf{B}, \mathbf{C})$ are vectorized and stacked to obtain the parameter vector $\mathbf{x}$

$$\mathbf{x} = \begin{bmatrix} a_1 \\ \vdots \\ a_{I\times R} \\ \vdots \\ b_1 \\ \vdots \\ c_{K\times R} \end{bmatrix} \tag{2.12}$$

Using the parameter vector, the gradient can be derived from the loss function (Eq. 2.2) as follows:

$$L(\mathbf{x}) = \|\mathbf{T}\|^2 - 2\langle \mathbf{T}, \hat{\mathbf{T}} \rangle + \|\hat{\mathbf{T}}\| \tag{2.13}$$

Taking the partial derivative for each entry in the factor matrices is expressed as:

$$\frac{\partial L}{\partial a_r^N} = -2\sum_{r=1}^{R}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} T_{ijk} a_{ir} b_{jr} c_{kr} + 2\sum_{r=1}^{R}\left(\prod_{\substack{m=1 \\ m\neq n}}^{N} a_r^{(m)T} a_l^{(m)}\right) a_l^{(n)} \tag{2.14}$$

Where a_r^N represents the r th component and N th mode.

Taking each partial derivative and expressing in matrix form obtains the following expression for each factor:

$$\begin{aligned}
\frac{\partial L}{\partial \mathbf{A}} &= -\mathbf{T}^{I\times JK}\mathbf{C} + \mathbf{A}(\mathbf{B}^T\mathbf{B} * \mathbf{C}^T\mathbf{C}) \\
\frac{\partial L}{\partial \mathbf{B}} &= -\mathbf{T}^{J\times IK}\mathbf{A} + \mathbf{B}(\mathbf{A}^T\mathbf{A} * \mathbf{C}^T\mathbf{C}) \\
\frac{\partial L}{\partial \mathbf{C}} &= -\mathbf{T}^{K\times JK}\mathbf{B} + \mathbf{C}(\mathbf{A}^T\mathbf{A} * \mathbf{B}^T\mathbf{B})
\end{aligned} \tag{2.15}$$

We used SciPy's minimize function using the L-BFGS-B method to solve for the tensor factors and recorded metrics throughout each iteration [20]. Missing values were omitted from the objective function. Optimization with missing values has been described and its methodology has been explained previously [21]. The algorithm was terminated either at a tolerance of 10^{-7} or at 50 iterations.

2.8 Censored Alternating Least Squares (CLS)

Censored alternating least squares is solved similarly to ALS but differs in its approach to handling the missing value problem. Consider a masking tensor $\mathbf{M}$ that encodes the missingness of $\mathbf{T}$, where an element that is missing is set as zero (T_{ijk} is missing, $M_{ijk} = 0$) and an element that is observed is set as one (T_{ijk} is present, $M_{ijk} = 1$).

All of the unique columns of the flattened masking tensor $\mathbf{M}$ (flattened along the direction depending on the factor being solved) represent all possible missingness patterns of the flattened tensor $\mathbf{T}$. We can iterate over each unique missing pattern solving portions of the tensor factor through a least squares fit (Fig. 2).

For each missing pattern, the columns of the flattened tensor $\mathbf{T}$ that correspond to that missing pattern are kept, being used to minimize the loss function for those columns of the factor matrix. Rows with missingness from $\mathbf{T}$ are dropped and likewise dropped from the the factor matrix $\mathbf{X}$. Equations for each of the factor matrices are identical to the ALS factor matrix equations (Eq. 2.10, 2.11) but with portions of the matrices selected. It is important to note that CLS is unable to be solved when an entire "slice" of the tensor is

missing, preventing the solving of some weights within the factor matrix.

To apply an L2 penalty for CLS, a regularization parameter can be added to the loss function (Eq. 2.9) to prevent any aberrant factor weights from a lack of data for fitting.

$$L(\mathbf{A}, \mathbf{B}, \mathbf{C}) = \|\mathbf{T}^{I \times JK} - \mathbf{A}(\mathbf{C} \odot \mathbf{B})^T\|_F^2 + \lambda \|(\mathbf{A}, \mathbf{B}, \mathbf{C})\|_2^2 \tag{2.16}$$

where λ penalizes the factor matrix being solved.

Metrics were recorded at each iteration of the algorithm and the algorithm was terminated either at a tolerance of 10^{-7} or at 50 iterations.

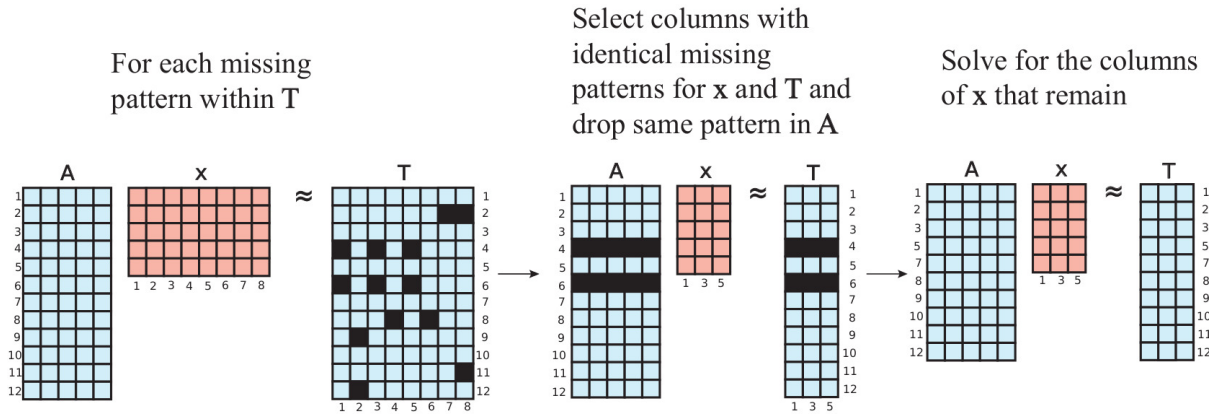

Figure 2: Visualization of the censored alternating least squares approach. Missing patterns are first observed in the flattened tensor. Columns are selected with identical missing patterns from both $\mathbf{T}$ and $\mathbf{x}$. Least squares is performed on the remaining columns and repeated for all other unique missing patterns to sequentially fill the factor matrix $\mathbf{x}$.

CHAPTER 3

Results

3.1 Simulated Data Sets

Simulated data sets are useful for controlling various features to assess their impact on each method. Thus, we first examined the performance of each of the methods on simulated data of known rank to assess performance (Fig. 3). Simulated data sets were generated using a rank of six, dimensionality $432 \times 6 \times 11$, with either entry drop, or cord drop (along the first mode) missingness patterns at 10% missingness. A rank of six was chosen to provide enough data set complexity for method comparison. A total of 50 simulated data sets were used for reducing error metric variance with the median plotted and error bars representing the interquartile range. SVD initialization was used for initialization for all three methods. The noise scaling parameter η was 0.1 indicating a relatively mild noise influence.

We first examined the entry imputation performance of CLS (Fig. 3a), ALS (Fig. 3b), and DO (Fig. 3c) across varying component numbers. Overall, all three methods were able to fit the 6 component model with similar fitted error. All methods had similar imputation error as model rank increased.

We then examined the cord imputation performance of CLS (Fig. 3d), ALS (Fig. 3e), and DO (Fig. 3f) in a similar fashion. All methods fit up to 5 components before different behaviors for each algorithm began to emerge. CLS and DO continued to properly fit for additional components before imputation error began to plateau and variance increased (Fig.

3d, 3f). ALS began to overfit and imputation accuracy worsened with more components (Fig. 3e). Overall, CLS reached the best imputation accuracy over ALS and DO for the rank six model. Fitting error behavior between the alternating algorithms was similar in that they continually decreased with increasing model rank.

Cord imputation performance at each iteration of the algorithm was examined for CLS (Fig. 3g), ALS (Fig. 3i), and DO (Fig. 3j) for the rank six model. Overall, the alternating algorithms differed in iterative behavior when compared to that of DO. Both CLS and ALS quickly converged (Fig. 3g, 3h). In contrast, DO converged to its lowest error bound at a slower rate and at a much higher iteration than that of the alternating algorithms (Fig. 3i).

Missingness percentage affects the overall performance of each algorithm in its ability to accurately impute values. Hence, we then increased the missingness percentage of the initial simulated data set to 25% (Fig. 4) while keeping all other parameters consistent with the 10% missing case. For entry missingness, similar behaviors were seen to the 10% missing case for CLS, ALS, and DO (Fig. 4a, 4b, 4c). Overall, all methods were able to fit up to around 4 or 5 components at similar error metric improvement rates. CLS saw the best imputation error improvement with increasing model rank (Fig. 4a). In contrast, ALS saw additional components lead to overfitting and imputation accuracy declining (Fig. 4b). DO remained stagnant and neither improved nor worsened (Fig. 4c).

Cord missingness method behavior began to change as more missing values were introduced to the original data set (Fig. 4d, 4e, 4f). ALS performed significantly worse compared to both CLS and DO and began overfitting at around 4 components (Fig. 4e). CLS and DO remained similar in their imputation accuracy as component size increased. However, CLS did exhibit significant variation for higher components while DO remained stable (Fig. 4d, 4f). All methods had similar fitted error behavior. Cord imputation performance per iteration with additional missing values remained relatively consistent compared to the 10% case for CLS, ALS, and DO (Fig. 4g, 4h, 4i). CLS and ALS quickly converged within early iterations while DO linearly decreased.

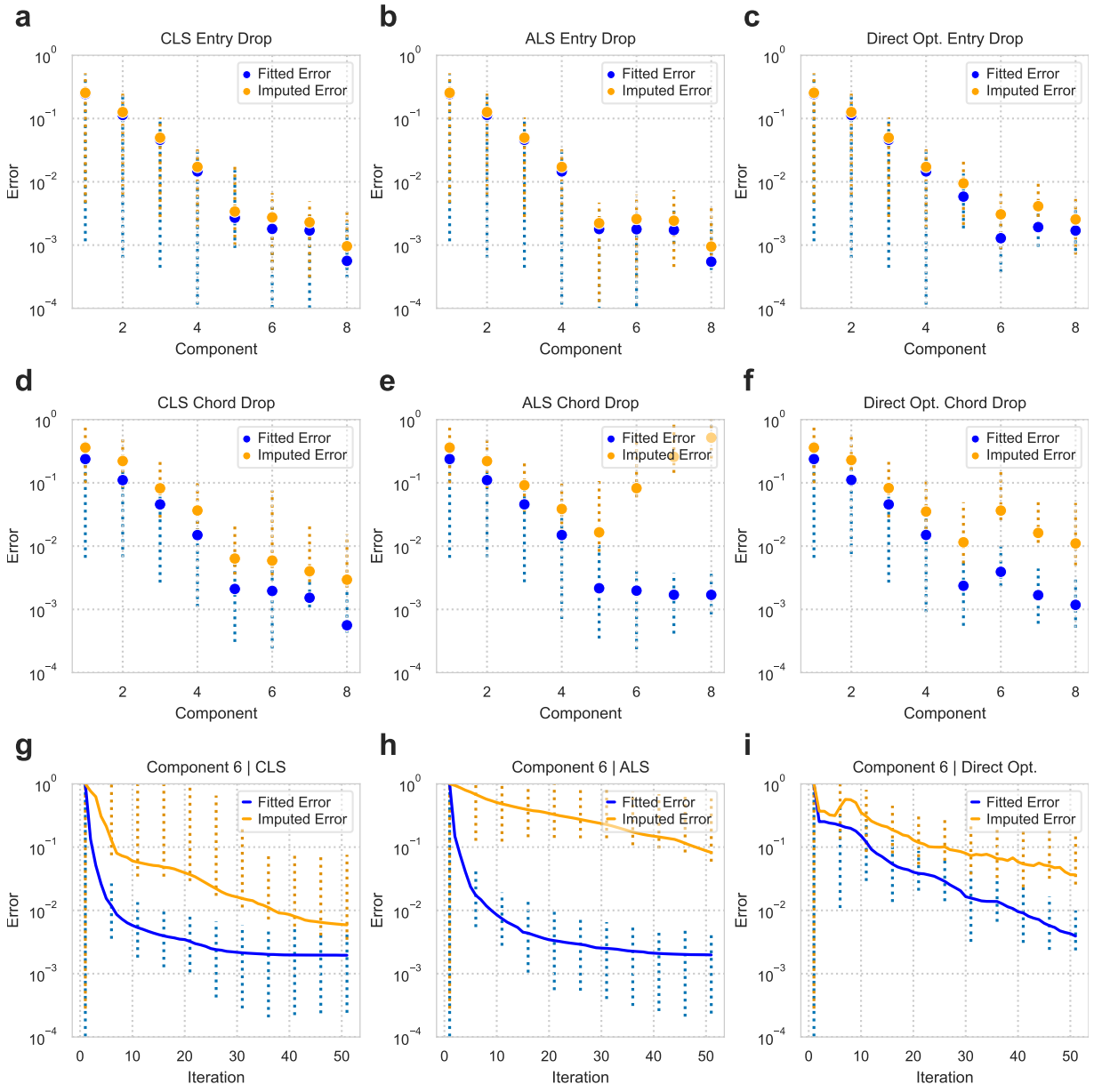

Figure 3: Algorithm comparison for simulated data with 10% missing values. a-c) CLS (a), ALS (b), DO (c) error metrics against component number for entry missingness. d-f) CLS (d), ALS (e), DO (f) error metrics against component number for cord missingness. g-i) CLS (g), ALS (h), DO (i) error metrics against iteration number for rank six model. 50 trials with IQR and median plotted.

Due to the variability of the imputation accuracy for CLS, we tested random factor initialization as well as increasing the scale of noise. When compared to SVD factor initialization, CLS achieved better cord imputation accuracy with more components as well as reduced imputation variance (Fig. S1d). Overall, CLS had the best imputation accuracy over the other methods tested when using random factor initialization. Negligible differences were seen for other methods for both entry and cord patterns (Fig. S1). With noise scaled by a factor of 10 ($\eta = 1$) indicating a severe noise influence, SVD initialization led to high cord imputation variance for CLS across trials (Fig. S2d) while initialization through random CP factors led to reduced variability (Fig. S3d). The additional noise reduced the rate of imputation improvement as model rank increased for all methods (Fig. S2, S3). No significant changes were seen for performance per iteration with using random CP factors and increased noise scaling.

3.2 Biological Data Sets

Upon the results using the simulated data sets, we applied the same analysis to a published systems serology data set [22]. SARS-CoV-2 RT-PCR negative and COVID-19 patients were sampled throughout their infection and their antibodies tested for Fc region and antigen binding. This antibody response data set is of dimensionality $432 \times 6 \times 11$ and is arranged to retain the sample, antigen, and receptor modes. A missing percentage of 10% for both entry and cord was applied.

For entry missingness, all three algorithms saw improvements in both error metrics with increasing component number (Fig. 5a, 5b, 5c). Both CLS and ALS were almost identical in their behavior and had similar performance metrics as components increased (Fig. 5a, 5b). For DO, both performance metrics were worse than that of CLS and ALS for components greater than 3 and saw little improvement from components 3 to 6 (Fig. 5c).

For cord missingness, CLS reached the lowest bound across all algorithms and components

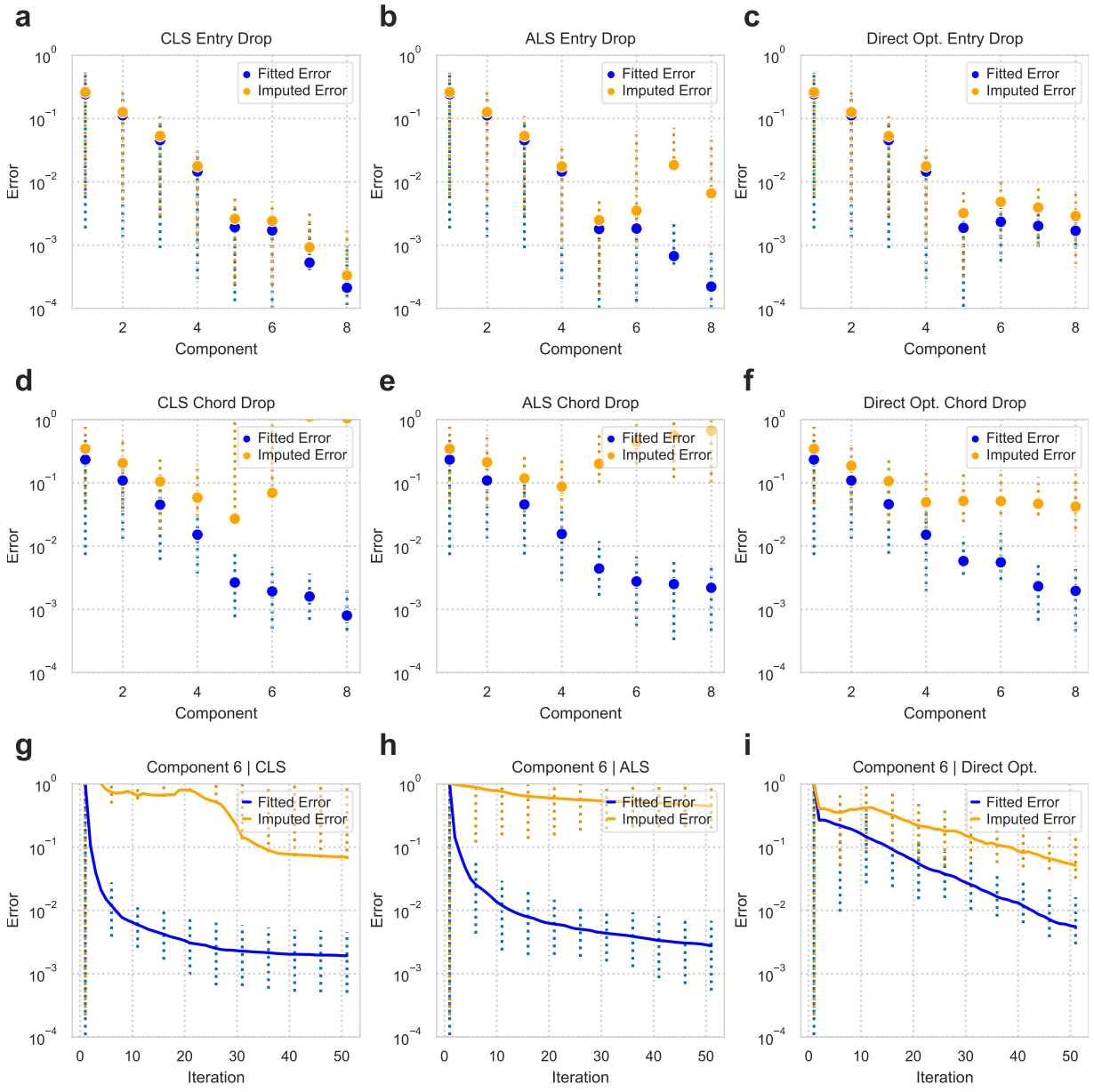

Figure 4: Algorithm comparison for simulated data with 25% missing values. a-c) CLS (a), ALS (b), DO (c) error metrics against component number for entry missingness. d-f) CLS (d), ALS (e), DO (f) error metrics against component number for cord missingness. g-i) CLS (g), ALS (h), DO (i) error metrics against iteration number for rank six model. 50 trials with IQR and median plotted.

(Fig. 5d, 5e, 5f). ALS fit up to 4 components until overfitting for higher ranks (Fig. 5e). CLS demonstrated improved performance with respect to ALS with a lower imputation error bound and decreased rate of overfitting (Fig. 5d). In contrast, DO reached its lowest imputation error at 5 components and overfit for higher rank models (Fig. 5f). Both CLS and ALS exhibited identical fitting error trends while DO did not improve from components 3 to 6, similar with the entry missing case.

Iteration plots for ALS and CLS exhibit behavior consistent with the simulated data case (Fig. 5g, 5h). For both ALS and CLS, fitting error quickly converges to the lower bound within the first few iterations. DO fitting error steadily decreases with additional iterations until convergence is reached. Imputation error behavior for both alternating algorithms is consistent with simulated data, with reasonably quick convergence for CLS and an initial drop with a gradual decline for ALS. Imputed error for DO fluctuated significantly throughout its run until stabilizing for higher iterations (Fig. 5i).

The same serological data set was similarly tested with an increased missingness percentage of 25%. As published data can vary in the degree of missing data, we sought to see how consistent error metrics were when less data was available for the model. For entry missingness, ALS and DO were only slightly impacted from the introduction of additional missing values (Fig. 6b, 6c). ALS began to overfit at an earlier component when compared to the 10% missing case, but still showed the best imputation performance. DO remained unaffected with additional missingness, with identical imputation accuracy across components. CLS followed the same trend up to 4 components, but additional components led to immediate overfitting with an imputation error over 1 (Fig. 6a). Fitting error followed the same trend as the previous cases, as both CLS and ALS improved upon DO as model rank increased.

For cord missingness, CLS and DO showed general improvement for early components while ALS immediately began to overfit (Fig. 6d, 6e, 6f). ALS saw its best imputation accuracy at 1 component and proceeded to overfit with increasing model rank, evident by the

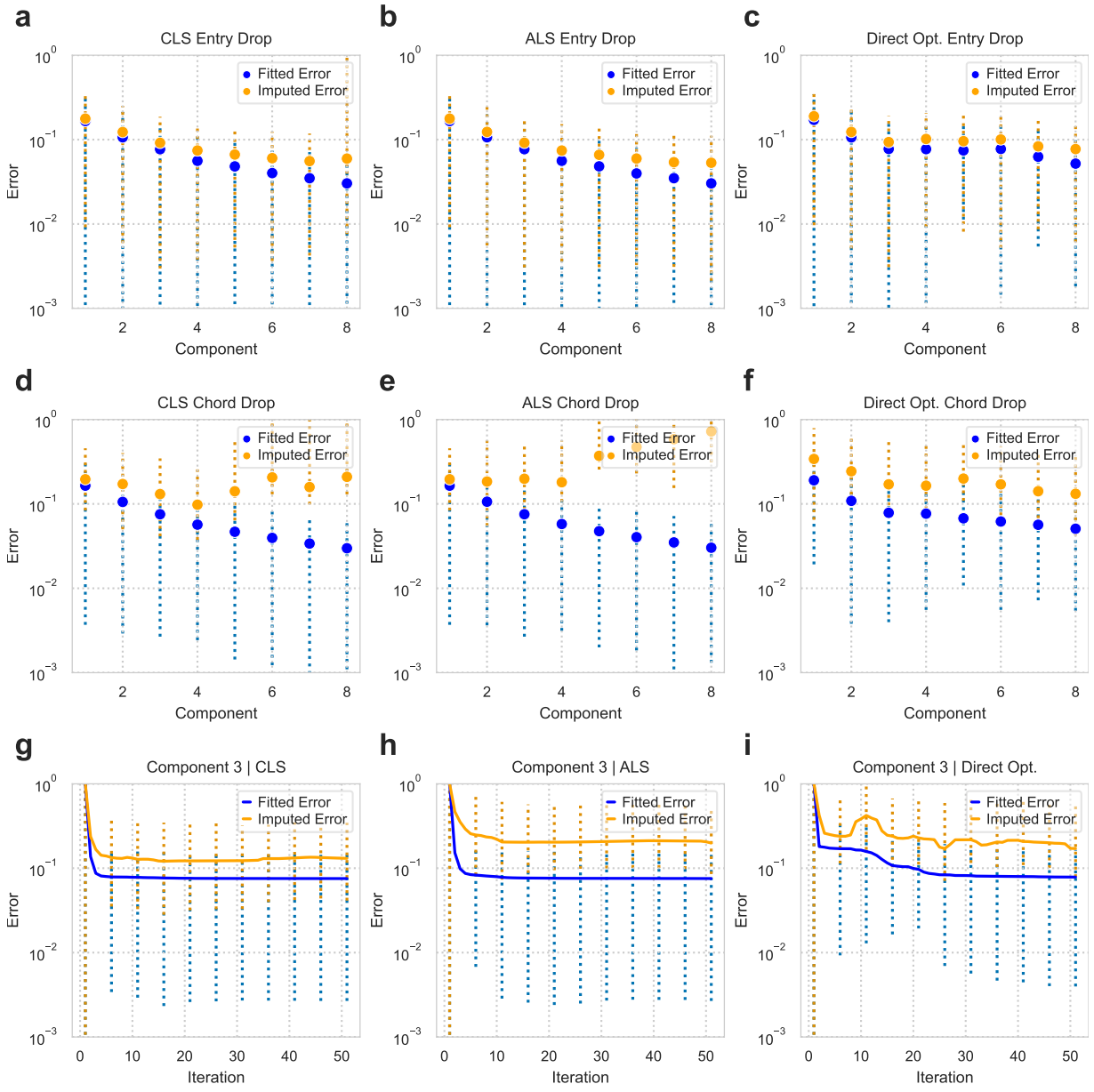

Figure 5: Algorithm comparison for serology data with 10% missing values. a-c) CLS (a), ALS (b), DO (c) error metrics against component number for entry missingness. d-f) CLS (d), ALS (e), DO (f) error metrics against component number for cord missingness. g-i) CLS (g), ALS (h), DO (i) error metrics against iteration number for rank 3 model. 50 trials with IQR and median plotted.

general upward trend of imputation error (Fig. 6e). CLS improved until 3 components and progressively worsened with increasing components with high variability in imputation error (Fig. 6d). DO followed a similar trend, with the imputation error decreasing until a plateau was reached (Fig. 6f). Overall, CLS had the best imputation accuracy at 3 components. Fitting error for both CLS and ALS remained identical to both missing percentages tested, as well as entry and cord missing patterns used.

Iteration plots remain similar to those from the 10% missingness case (Fig. 6g, 6h, 6i). Overall, all three algorithms present the same behavior as previous tests with low missing percentage on the same serological data set.

3.3 Ridge Regression

For CLS in the presence of a high percentage of missing values, factor weights were prone to high variance and resulted in aberrant imputatative accuracy (Fig. 6a). L2 regularization was used on the CLS algorithm to reduce any variance caused by the presence of missing values and improve imputation. The optimal regularization parameter value will differ between the component number choice and as such was swept to find its optimal value. (Fig. 7). For a 4 component model, regularization showed little to no improvement for entry imputation error (Fig. 7a). For both the 5 and 6 component models, imputation accuracy significantly improved when the regularization parameter reached at least 10^{-2} (Fig. 7b, 7c). When the parameter is set to 10^{-1} significant entry imputation error was reduced and the results were consistent with components that did not require regularization.

For a full analysis of the regularized algorithm performance, error metrics were plotted for both entry and cord missingness in a similar fashion using an L2 regularization parameter of 10^{-1}, the value where entry saw the earliest improvement (Fig. 8). CLS performed as expected with entry imputation error gradually decreasing with increasing component, extending the trend seen in the non-regularized case (Fig. 8a). Imputation error stopped

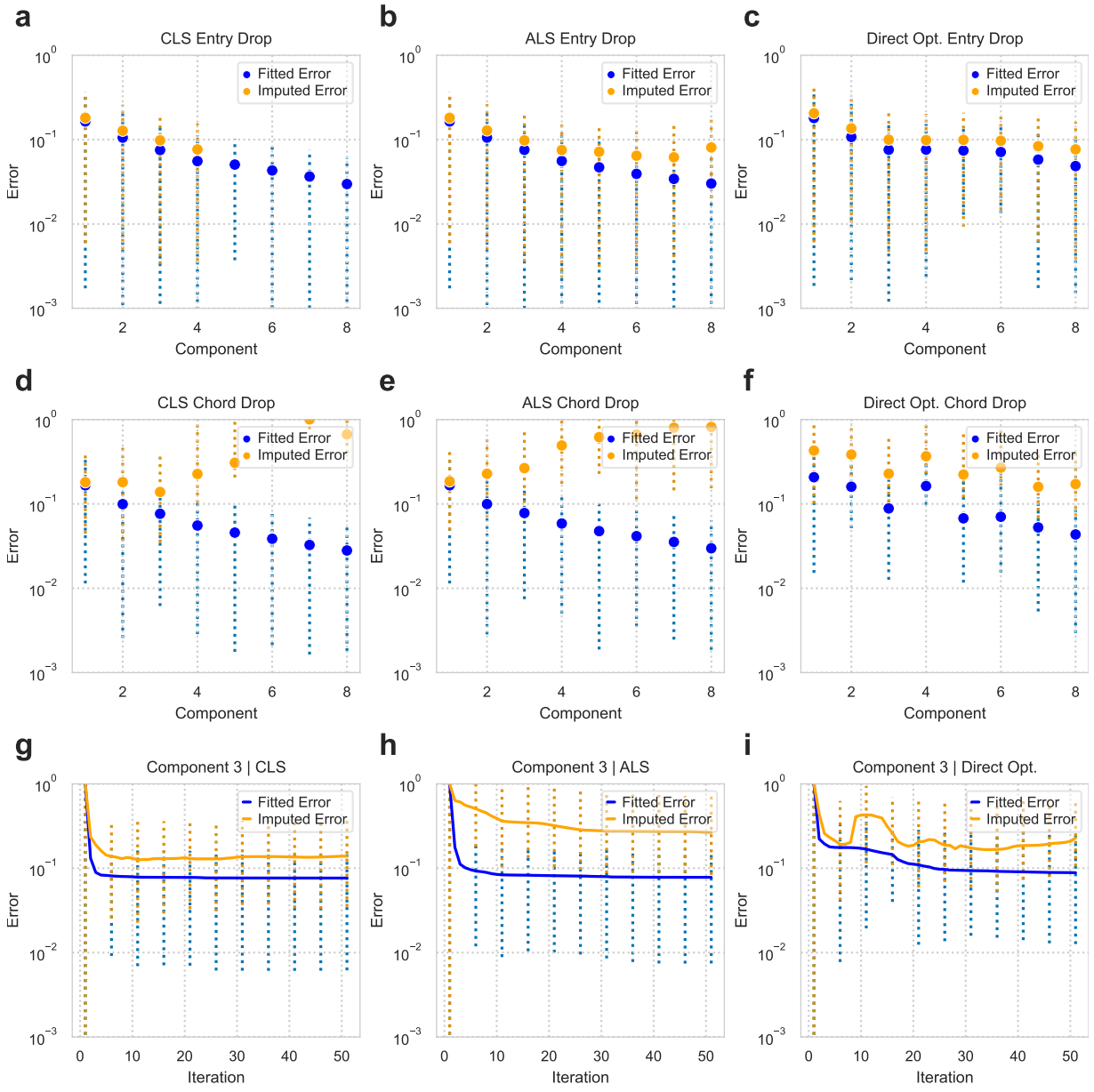

Figure 6: Algorithm comparison for serology data with 25% missing values. a-c) CLS (a), ALS (b), DO (c) error metrics against component number for entry missingness. d-f) CLS (d), ALS (e), DO (f) error metrics against component number for cord missingness. g-i) CLS (g), ALS (h), DO (i) error metrics against iteration number for rank 3 model. 50 trials with IQR and median plotted.

improving around 6 components, consistent with ALS and CLS for both missing cases. Fitted error is otherwise unaffected from the regularization parameter chosen and remains identical with the previous cases.

The general trend between cord missingness imputation between the original and regularized seems to be present for CLS (Fig. 8b). Fitting error remained unaffected from the regularization parameter chosen. Both error metrics quickly converged to their lower bound within the first few iterations as was seen in the non-regularized case (Fig. 8c).

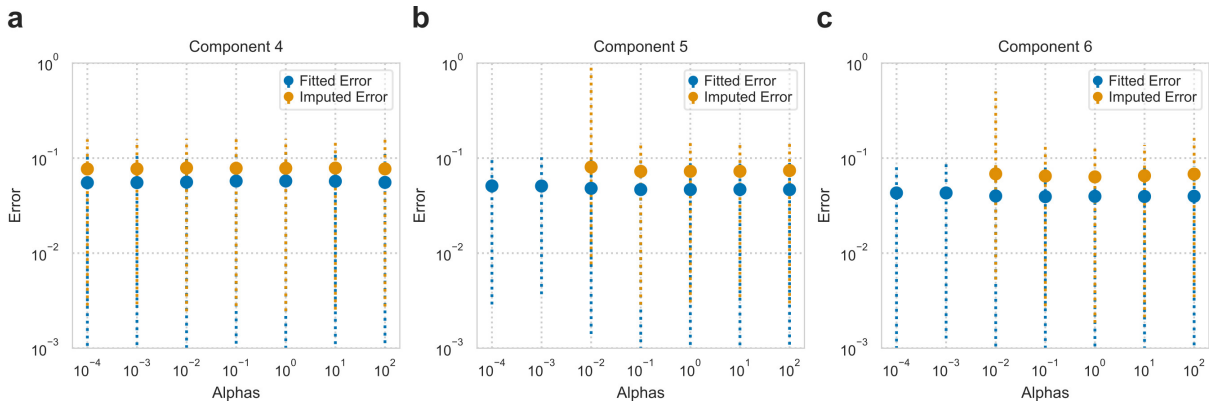

Figure 7: L2 regularization for serology data with 25% entry missing values. a) 4 component model with error metrics plotted against regularization parameter. b) 5 component model with error metrics plotted against regularization parameter. c) 6 component model with error metrics plotted against regularization parameter. 50 trials with IQR and median plotted.

3.4 Run Time Analysis

Each algorithm would reach its convergence on a different time scale, due to the nature of their implementation for solving the factor matrices. CLS features additional iterations within each factor solving, with the computations required depending on the amount of unique missing patterns present. As such, we wanted to test at what time point following

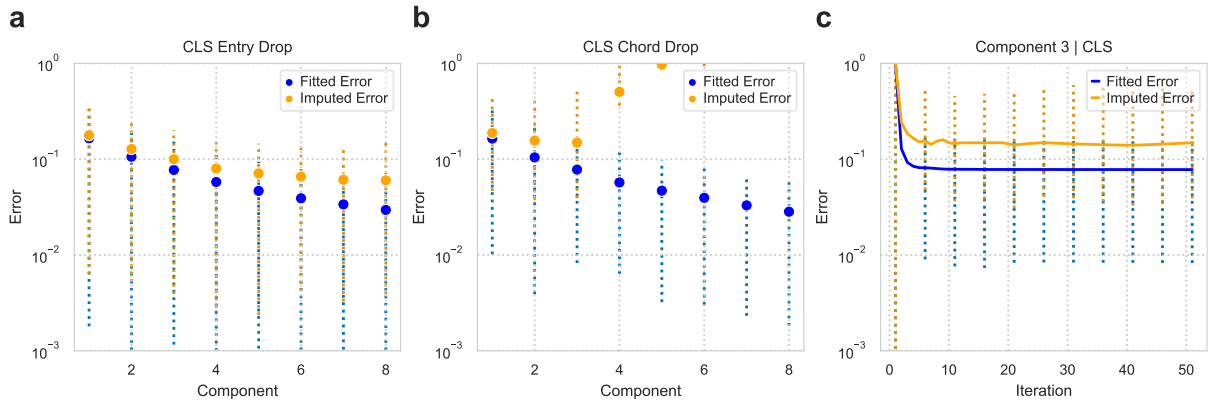

Figure 8: CLS with L2 for serology data with 25% missing values a) Error metrics against component number for entry missingness. b) Error metrics against component number for cord missingness. c) 3 component model error metrics against iteration number. L2 Regularization parameter of 0.1.

initialization did each method reach some predefined imputation error bound (Fig. 9). Models with 3 components for the 3D serological data were generated with 10% cord missingness and repeated 100 times to generate a distribution of time points when imputation error reached at most 0.25 (Fig. 9a). For runs that reached the error bound, ALS was typically the fastest, followed by CLS, and then DO. Since some runs would never reach the error bound due to the variation in left out data, the count of runs for each method was plotted (Fig. 9b). CLS had significantly fewer runs that did not reach the error bound compared to ALS and DO.

3.5 Higher Dimensional Data Set

As the serological data set was structured in a 3-way fashion, we sought to look into whether CLS improvements were more pronounced when a data set is of higher dimensionality. Assuming the long dimension had the capability to be further reduced into two separate modes, a 4D data set could be constructed. For all methods, they can be adapted to account for

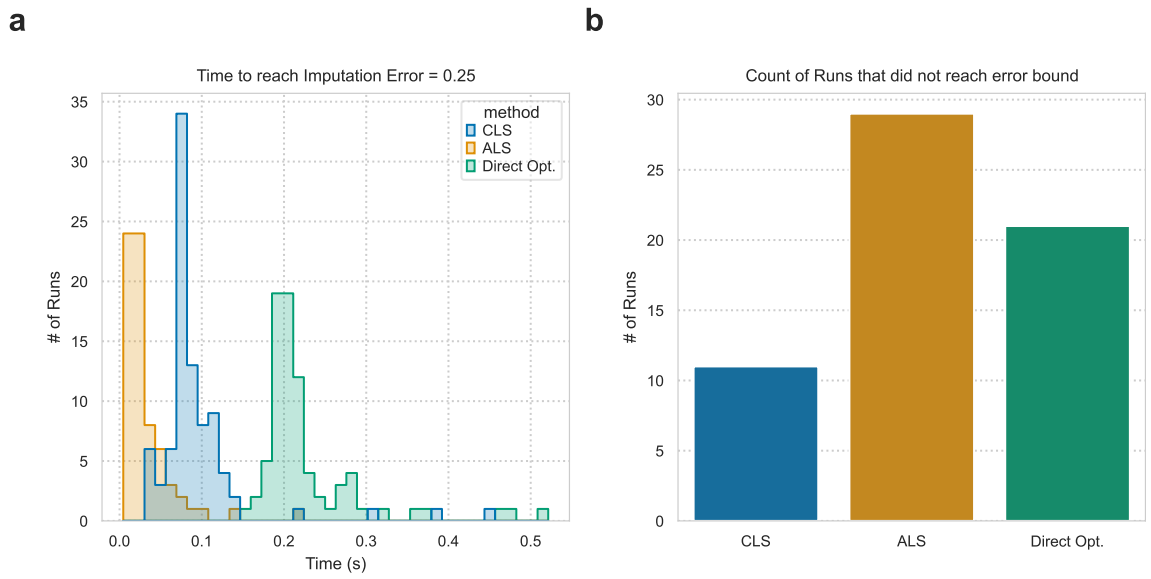

Figure 9: Rank 3 model algorithm run time analysis with 10% cord missingness. a) Distribution of algorithms run time upon reaching an imputation error of 0.25. b) Count of trials that did not reach imputation error of 0.25. CLS - *Blue*, ALS - *Orange*, DO - *Green*

higher mode models at the cost of computational efficiency. We then tested these methods on a simulated data set that was structured into a 4D fashion of rank 6 (Fig. 10). This data set was of dimensionality $24 \times 6 \times 11 \times 18$ and had 25% missingness. In this case, the mode with length 432 was reduced to two modes with lengths 24 and 18. All other parameters remained constant with previous sections.

For entry missingness, CLS and ALS both had the best overall imputation accuracy over DO for high model ranks (Fig. 10a, 10b, 10c). Both alternating algorithms behaved similarly with no significant differences in performance. DO reached component 3 before minor imputation and fitting error improvements were made with increasing component number. Both ALS and CLS had similar fitting error results until overfitting occurred, where fitting error for ALS was worse than that of CLS.

For cord missingness, CLS demonstrated the best imputation accuracy compared to ALS and DO (Fig. 10d, 10e, 10f). ALS and CLS performed identically until 6 components,

where ALS began to overfit and imputation error sharply increased. DO, in a similar fashion to entry missingness, reached component 4 before no further improvement was seen. CLS reached the best imputation accuracy at the intended 6 components and did not overfit for additional components. Fitting error trends remain identical across all methods between entry missingness and cord missingness.

Iteration plots remain similar with previous results for all methods (Fig. 10g, 10g, 10i). Both CLS and ALS have an initial sudden decrease in both fitting error and imputed error for early iterations, followed by a gradual decline for further iterations. For DO, fitted and imputed error both remain constant for early iterations, followed by a subsequent linear decrease not appearing to reach any lower bound.

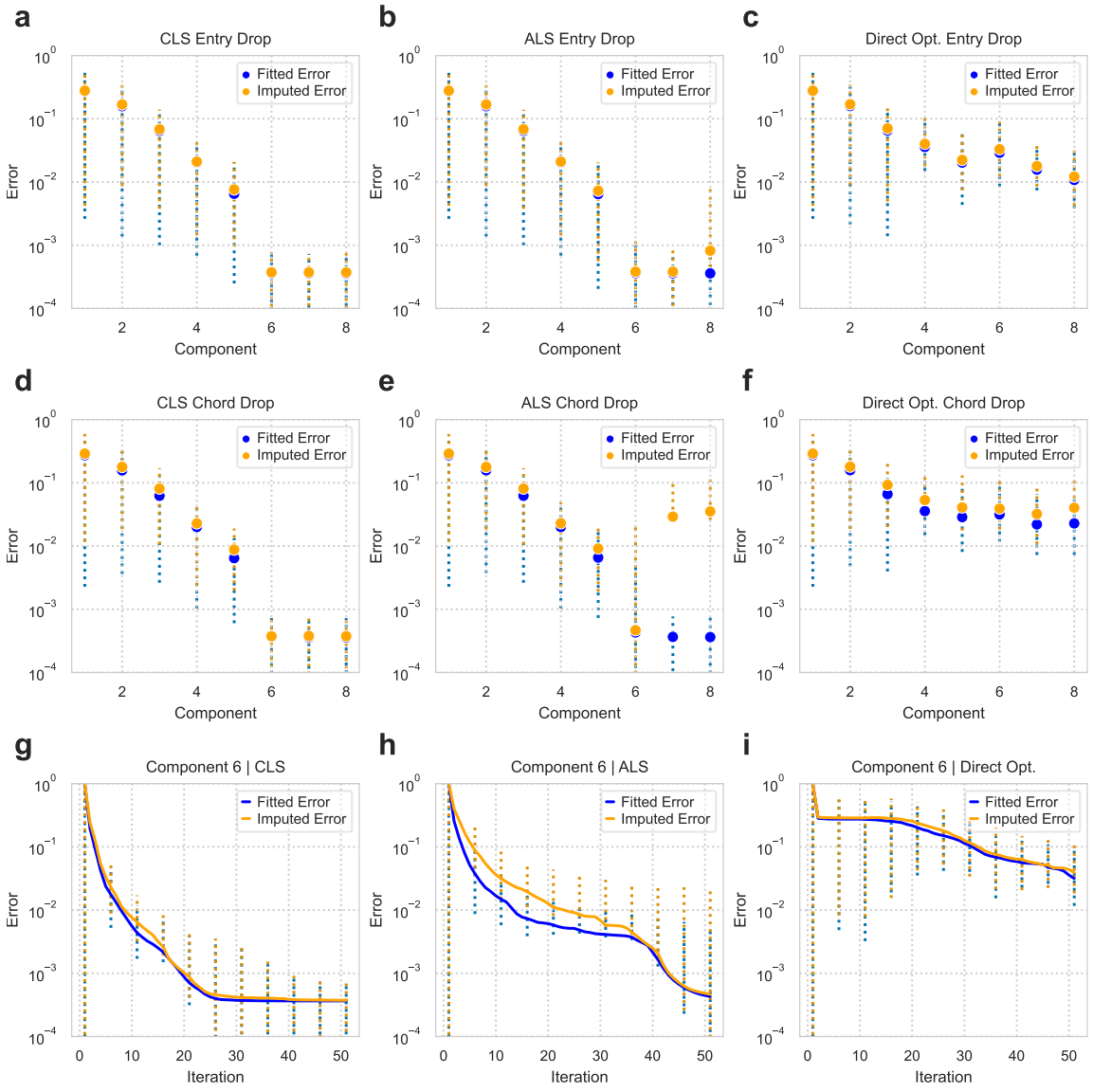

Figure 10: Algorithm comparison for 4D simulated data with 25% missing values. a-c) CLS (a), ALS (b), DO (c) error metrics against component number for entry missingness. d-f) CLS (d), ALS (e), DO (f) error metrics against component number for cord missingness. g-i) CLS (g), ALS (h), DO (i) error metrics against iteration number for rank 6 model. 50 trials with IQR and median plotted.

CHAPTER 4

Discussion

In this study, we show that CLS can improve imputation accuracy on both simulated data and a serological data set when compared to algorithms that are typically used for CP tensor decomposition. CLS specifically improved imputation accuracy over ALS and DO for cord missingness using simulated data, in both 10% missing cases (Fig. 3) and 25% missing cases (Fig. 4). On the serology data set, both CLS and ALS improved entry imputation error while CLS had the best improvement for cord missingness (Fig. 5, 6). CLS remains sensitive to increased missingness, where regularization can remove any aberrant behavior when model variance is increased (Fig. 7). We show that CLS with the L2 penalty fixes entry imputation performance (Fig. 8). We also show that CLS maintains adequate run time with the additional iterations associated with the method (Fig. 9). With the 4D simulated data set, we show that the improvements made by CLS are more pronounced (Fig. 10).

While both CLS and ALS operate in an alternating fashion, their results begin to differ for higher rank models. Among the 3D simulated data sets, ALS proceeded to overfit regardless of the amount of values left out of the original data set. CLS began to overfit once a threshold of percent missing was achieved, only evident when cords were removed. Many factors can play a role as to why both alternating algorithms differ in their imputative capability. Appropriate factor initialization is known to affect convergence behavior and can assure a global minimum is reached [18]. With the presence of missing values, ALS utilizes the reconstructed tensor throughout each iteration in its factor solving. Hence, a poor initial estimation may play a role. While factors were initialized using the SVD, similar

imputation behavior was seen among methods when random initialization was used instead, sometimes improving upon SVD imputation initialization (Fig. S1). SVD imputation under these conditions may lead to worse outcomes. CLS tackles the missing value problem in a different manner, through the iteration of missing patterns seen within the data set. Each factor weight is only solved using data that is originally present, avoiding the issue of accurate initial imputation. When high missingness is present, CLS only utilizes few values in the solving of a specific tensor factor weight. This may affect the resulting high variance present in high component models. Factor inconsistency between runs may lead to this behavior, where specific removed indices cause disparities between tensor factors across repeats. DO plateaus in imputation accuracy with increasing model rank and worsens with increased missingness. While no improvement is seen, imputation variance does not significantly increase as shown with CLS. A likely reason is the time for convergence for high rank models using DO. Additional iterations seem necessary for DO specifically for the simulated data sets. Convergence was reached for the serology data set and additional iterations were not shown to always yield imputation error improvement. Nonetheless, the extra iterations required would result in a significantly longer amount of time for factor solving.

Simulated data allows the adjustments of features for assessment of each method. However, the CP model assumes linearity between modes of which the simulated data was built upon. Real data is assumed to have this relationship but may not always be the case. Regardless, some general trends could be seen that were shared between both the simulated data and serology data set. In the case of low percent missingness, CLS and ALS both improve upon DO for entry missingness. In the simulated case, ALS overfit prior to CLS while the opposite is true for the serology data set. This trend is even more clear when additional values are removed where CLS overfitting is sudden and imputation accuracy is extremely poor. We addressed this through regularizing CLS with a parameter value that removed the aberrant behavior. Following regularization, CLS behavior did revert back to what was seen with the simulated data. One reason may be through inherent noise within the serology

data, with factor weights deviating far from the ideal value when coupled with both noise and missingness. Why this occurs for only CLS and not ALS leaves room for additional investigation. ALS iteratively updates the missing values within the tensor which may lead to some resilience against the confounding effects of noise and missing data. CLS ignores these missing positions leading to poor estimation of individual factor weights when minimal data entries are included in that weight's least squares solution. L2 regularization seems to prevent this extreme weight estimation, especially for high rank models. Cord missingness performance shows similar shared behavior between the simulated data and serology data. CLS performs best for both low missing cases across all model ranks, while ALS experiences overfitting. With increased missingness, imputation error variance increases are consistent among what we saw when using simulated data. A noise-missingness confounding effect may play a role as well. DO retains the same general trend, where imputation accuracy is translated up for increased missingness. Gradient based algorithms may be more resilient to noisy data when coupled with increased missingness. Gradient descent for noisy tensor completion has been investigated, but only for symmetric tensors [23].

A drawback with CLS is due to its time complexity, where additional sub-iterations are required depending on the missing patterns present. At worst, each column is individually solved prolonging the time until convergence is reached. We showed that even with this sub-iteration step, CLS performs faster than DO and still remains a viable method choice. Even in the case of entry missingness leading to significantly many more missing patterns, CLS performs faster than DO (not shown). The iteration behavior seems to confirm this, where CLS and ALS reach convergence in fewer iterations. While only fractional seconds of improvement was observed, with increasing data set size, model rank, and data set dimensionality can lead to this effect compounding and yielding more drastic differences between the methods.

The improvements seen using CLS are more pronounced when we tested the methods using 4D simulated data. Most of the observations discussed remain contained within the

4D case. ALS overfits at high components while CLS remains accurate. DO does worse than both CLS and ALS across both entry and cord missing patterns. It is possible that least squares solving favors optimization for higher modal models. The 4D simulated data dimensionality may play a role in the improvements seen by CLS and ALS as the alternating method targets individual modes. It may be possible that mode dimension size discrepancies lead to highly over-determined least squares problems affecting the overall model fit. Gradient based approaches may be a better choice in these scenarios. As the serology data is long and thin, CLS and ALS may have been affected solving the 2nd mode where only a few factor weights were being computed. Previous work has been done on optimizing these over-determined systems for alternating methods and have been shown to be less sensitive to initial starting points [24].

Picking an algorithm for factor solving may not be clear in the absolute sense, but entirely depends on the structure of the data set being modeled. Although the serology data set seems to uphold the linearity assumption, it is crucial to understand the sources of other forms of biological data sets in order to ensure they also adhere to this assumption. When random entry missing values are encountered, some flexibility is allowed in method choice. For a small percent of missing values, any method is a viable choice in terms of imputative performance. However, ALS does converge on a significantly quicker time scale than both CLS and DO and may be the best option when data sets are large. For a high percentage of missing values, either CLS or ALS should be employed. However, we did see that CLS required regularization for the serology data to maintain accurate imputation performance. Method choice for data sets with structured cord missing values are partially dependent on the noise present and the percent removed. Under a low percentage of missing values, CLS shows the best performance at imputing missing values on both the simulated and serology data. For a high percentage of missing values, CLS sometimes requires either initialization through random CP factors or regularization to avoid the risk of inconsistent factor solving. Regardless, it consistently proves to be the better option in these circumstances. For data

sets with a high degree of noise, CLS cord imputation benefits are not as pronounced but still prevalent when random CP factors are used for initialization. Random entry missing values coupled with high noise favors no method, and as such, any choice is suitable. The importance of an accurate initialization step is demonstrated here, where CLS performance is dependent on accurate initial SVD imputation influenced from missing values and noise.

When using CP decomposition, interpretability is a driving factor for its choice in data analysis. The key to effective tensor imputation is through maximizing the imputative accuracy with the minimum possible rank. It is desirable for few components to effectively capture the relevant patterns describing the data set. CLS accomplishes this key aspect of CP factor analysis under a variety of different situations that may be encountered in biological data analysis. We present various situations that may be encountered for tensor imputation and offer guidelines for method choice.

CHAPTER 5

Supplement

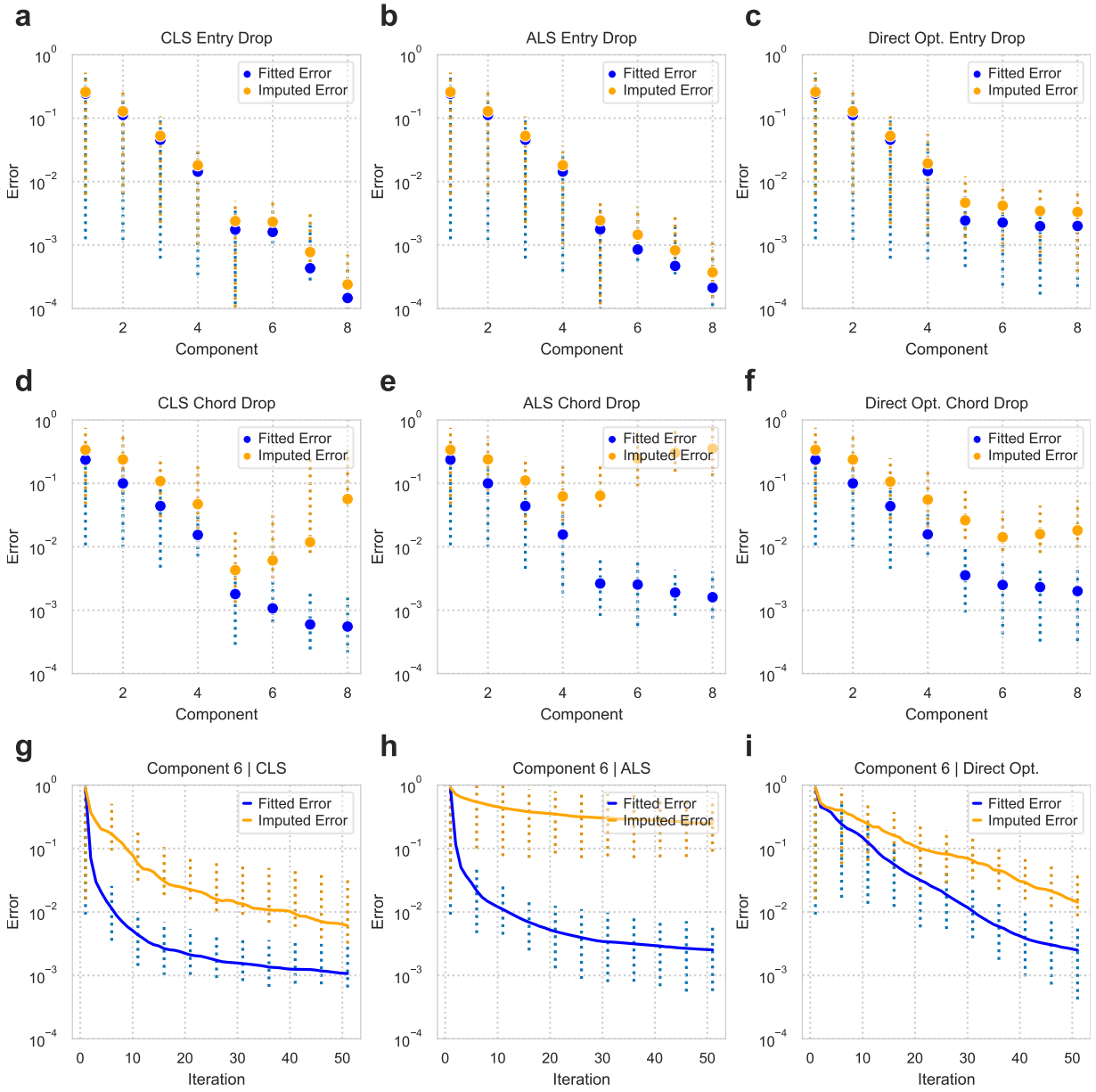

Figure S1: Algorithm comparison for simulated data with 25% missing values and initialized using random CP factors. a-c) CLS (a), ALS (b), DO (c) error metrics against component number for entry missingness. d-f) CLS (d), ALS (e), DO (f) error metrics against component number for cord missingness. g-i) CLS (g), ALS (h), DO (i) error metrics against iteration number for rank six model. 50 trials with IQR and median plotted. Initialized using random CP factors.

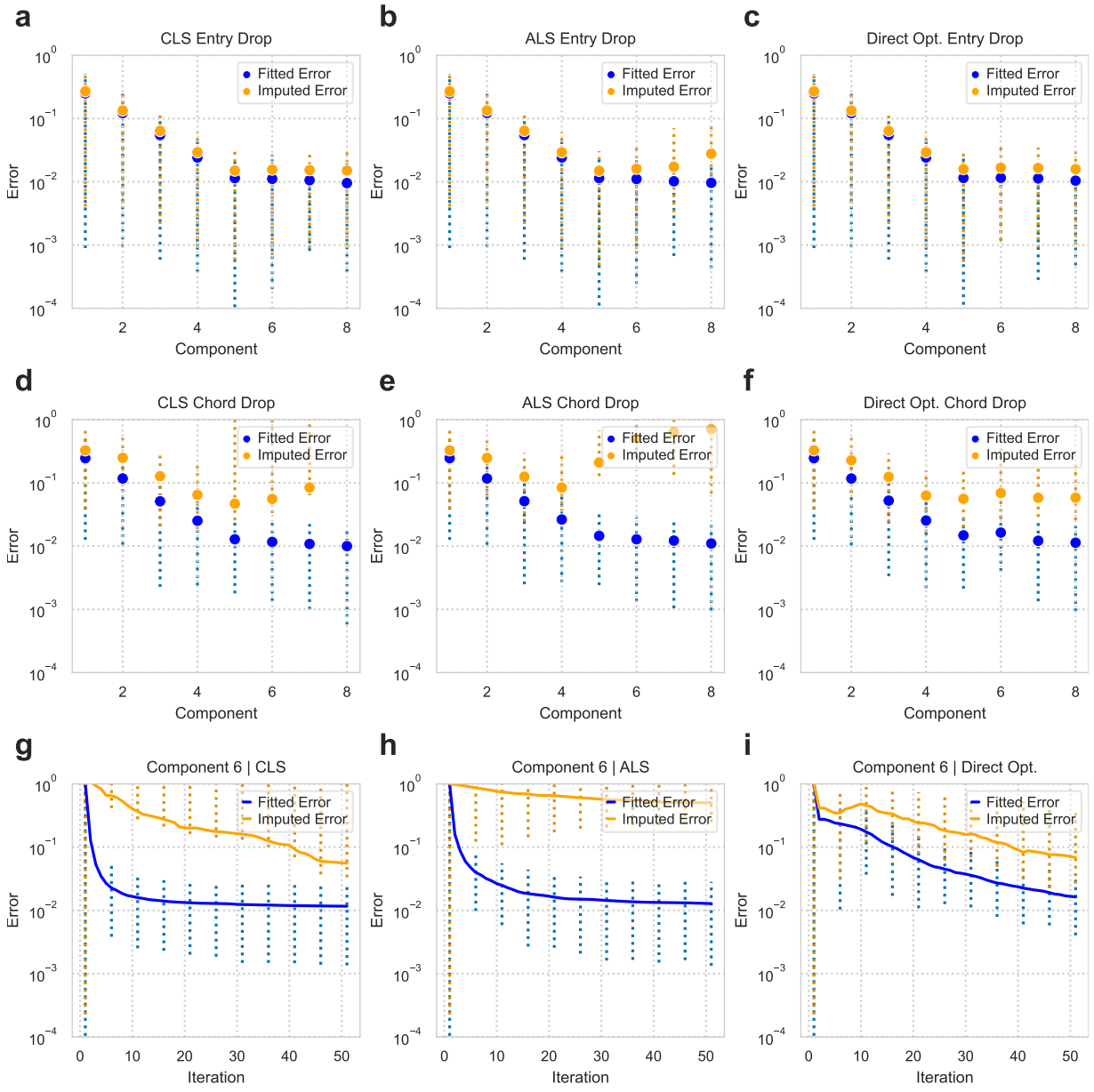

Figure S2: Algorithm comparison for noisy simulated data with 25% missing values. a-c) CLS (a), ALS (b), DO (c) error metrics against component number for entry missingness. d-f) CLS (d), ALS (e), DO (f) error metrics against component number for cord missingness. g-i) CLS (g), ALS (h), DO (i) error metrics against iteration number for rank three model. 50 trials with IQR and median plotted. Initialized using SVD. $\eta = 1$

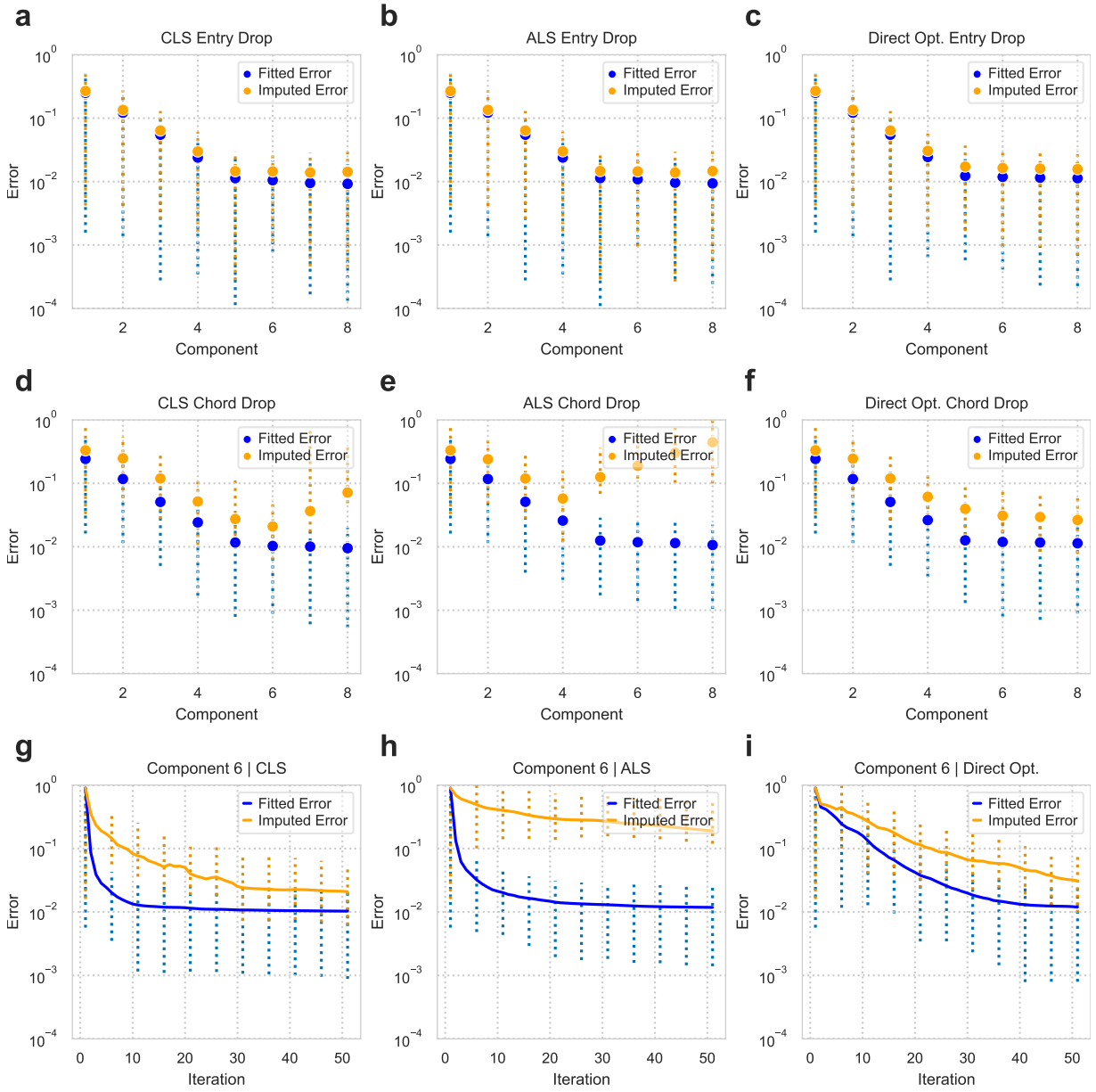

Figure S3: Algorithm comparison for noisy simulated data with 25% missing values and initialized using random CP factors. a-c) CLS (a), ALS (b), DO (c) error metrics against component number for entry missingness. d-f) CLS (d), ALS (e), DO (f) error metrics against component number for cord missingness. g-i) CLS (g), ALS (h), DO (i) error metrics against iteration number for rank three model. 50 trials with IQR and median plotted. Initialized using random CP factors. $\eta = 1$

REFERENCES

[1] Giorgio Tomasi and Rasmus Bro. PARAFAC and missing values. *Chemometrics and Intelligent Laboratory Systems*, 75(2):163–180, February 2005.

[2] Orly Alter, Patrick O. Brown, and David Botstein. Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences*, 97(18):10101–10106, August 2000. Publisher: Proceedings of the National Academy of Sciences.

[3] Neal S. Holter, Madhusmita Mitra, Amos Maritan, Marek Cieplak, Jayanth R. Banavar, and Nina V. Fedoroff. Fundamental patterns underlying gene expression profiles: Simplicity from complexity. *Proceedings of the National Academy of Sciences*, 97(15):8409–8414, July 2000.

[4] Farzane Yahyanejad, Réka Albert, and Bhaskar DasGupta. A survey of some tensor analysis techniques for biological systems. *Quantitative Biology*, 7(4):266–277, December 2019. Number: 4 Publisher: Higher Education Press.

[5] Zhixin Cyrillus Tan, Madeleine C Murphy, Hakan S Alpay, Scott D Taylor, and Aaron S Meyer. Tensor-structured decomposition improves systems serology analysis. *Molecular Systems Biology*, 17(9):e10243, September 2021.

[6] Larsson Omberg, Gene H. Golub, and Orly Alter. A tensor higher-order singular value decomposition for integrative analysis of DNA microarray data from different studies. *Proceedings of the National Academy of Sciences of the United States of America*, 104(47):18371–18376, November 2007.

[7] Rasmus Bro and Henk A. L. Kiers. A new efficient method for determining the number of components in PARAFAC models. *Journal of Chemometrics*, 17(5):274–286, 2003.

[8] Woody Austin, Tamara Kolda, and Todd Plantenga. Tensor Rank Prediction via Cross Validation, August 2014.

[9] Tamara G. Kolda and Brett W. Bader. Tensor Decompositions and Applications. *SIAM Review*, 51(3):455–500, August 2009. Publisher: Society for Industrial and Applied Mathematics.

[10] Lars Elden. Perturbation Theory for the Least Squares Problem with Linear Equality Constraints. *SIAM Journal on Numerical Analysis*, 17(3):338–350, 1980. Publisher: Society for Industrial and Applied Mathematics.

[11] Evrim Acar, Daniel M. Dunlavy, and Tamara G. Kolda. A scalable optimization approach for fitting canonical tensor decompositions. *Journal of Chemometrics*, 25(2):67–86, February 2011.

[12] Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping From Saddle Points — Online Stochastic Gradient for Tensor Decomposition. In *Proceedings of The 28th Conference on Learning Theory*, pages 797–842. PMLR, June 2015. ISSN: 1938-7228.

[13] Animashree Anandkumar and Rong Ge. Efficient approaches for escaping higher order saddle points in non-convex optimization. In *Conference on Learning Theory*, pages 81–102. PMLR, June 2016. ISSN: 1938-7228.

[14] Takanori Maehara, Kohei Hayashi, and Ken-ichi Kawarabayashi. Expected Tensor Decomposition with Stochastic Gradient Descent. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1), February 2016. Number: 1.

[15] R. Harshman. Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multi-model factor analysis. 1970.

[16] J. Douglas Carroll and Jih-Jie Chang. Analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart-Young" decomposition. *Psychometrika*, 35(3):283–319, September 1970.

[17] Age Smilde, Rasmus Bro, and Paul Geladi. Algorithms. In *Multi-Way Analysis with Applications in the Chemical Sciences*, pages 111–144. John Wiley & Sons, Ltd, 2004.

[18] Rasmus Bro. PARAFAC. Tutorial and applications. *Chemometrics and Intelligent Laboratory Systems*, 38(2):149–171, October 1997.

[19] Jean Kossaifi, Yannis Panagakis, Anima Anandkumar, and Maja Pantic. TensorLy: Tensor Learning in Python. *Journal of Machine Learning Research*, 20(26):1–6, 2019.

[20] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

[21] Evrim Acar, Daniel M. Dunlavy, Tamara G. Kolda, and Morten Mørup. Scalable tensor factorizations for incomplete data. *Chemometrics and Intelligent Laboratory Systems*, 106(1):41–56, March 2011.

[22] Tomer Zohar, Carolin Loos, Stephanie Fischinger, Caroline Atyeo, Chuangqi Wang, Matthew D. Slein, John Burke, Jingyou Yu, Jared Feldman, Blake Marie Hauser, Tim Caradonna, Aaron G. Schmidt, Yongfei Cai, Hendrik Streeck, Edward T. Ryan, Dan H.

Barouch, Richelle C. Charles, Douglas A. Lauffenburger, and Galit Alter. Compromised Humoral Functional Evolution Tracks with SARS-CoV-2 Mortality. *Cell*, 183(6):1508–1519.e12, December 2020.

[23] Changxiao Cai, Gen Li, H. Vincent Poor, and Yuxin Chen. Nonconvex Low-Rank Tensor Completion from Noisy Data. *Operations Research*, 70(2):1219–1237, March 2022. Publisher: INFORMS.

[24] Casey Battaglino, Grey Ballard, and Tamara G. Kolda. A Practical Randomized CP Tensor Decomposition. *SIAM Journal on Matrix Analysis and Applications*, 39(2):876–901, January 2018.