A Low Rank Model for Phenotype Imputation in Autism Spectrum Disorder

Kelley M. Paskov M.S.[1], **Dennis P. Wall, Ph.D**[1]
[1]**Stanford University, Stanford, California**

Abstract

Autism Spectrum Disorder is a highly heterogeneous condition currently diagnosed using behavioral symptoms. A better understanding of the phenotypic subtypes of autism is a necessary component of the larger goal of mapping autism genotype to phenotype. However, as with most clinical records describing human disease, the phenotypic data available for autism contains varying levels of noise and incompleteness that complicate analysis. Here we analyze behavioral data from 16,291 subjects using 250 items from three gold standard diagnostic instruments. We apply a low-rank model to impute missing entries and entire missing instruments with high fidelity, showing that we can complete clinical records for all subjects. Finally, we analyze the low-rank representation of our subjects to identify plausible subtypes of autism, setting the stage for genome-to-phenome prediction experiments. These procedures can be adapted and used with other similarly structured clinical records to enable a more complete mapping between genome and phenome.

Introduction

Autism spectrum disorder (ASD) is one of the most common developmental pediatric conditions impacting 1 in 68 children[1]. Twin studies have been used to demonstrate a strong genetic component with concordance between monozygotic twins ranging from 37-95%[2–5]. For this reason, autism has been a major focus of the field of translational genomics, with now nearly 21,000 fully sequenced whole genomes from various independent collaborative efforts[6–11]. These efforts have advanced our understanding of the genetic contribution to the autism phenotype and have helped to build plausible genetic models for autism[12–15], but the specific genetic markers responsible for varying forms of autism remain unknown. At least two factors contribute to this. First, the sample size may still need to be expanded since the most likely genetic model involves combinations of common variants rather than single highly penetrant loss of function rare variants. The second arises from the all too common problem of inadequate phenotyping. In this study, we focus on the second issue.

With the attention paid to developing large research cohorts for autism sequencing, there has been a companion focus on the phenotypic characterization of research subjects. While many behavioral instruments have been developed for autism, three of the most commonly used are the Autism Diagnostic Interview-Revised (ADI-R)[17], the Autism Diagnostic Observation Schedule (ADOS)[16], and the Social Responsiveness Scale (SRS)[18]. ADI-R and SRS both ask the primary caregiver to report on a range of behaviors. The ADI-R is a clinically administered questionnaire, while the SRS takes less time to administer and is used primarily in research. ADOS is a structured exam that measures behavior during a staged clinical observation. Combined, these tests evaluate 250 behaviors that follow an ordinal level of severity from unimpaired to highly impaired. For example, one question focuses on eye contact and determines whether the child never, sometimes, often, or always makes good eye contact within an appropriate social setting.

Although these data provide highly granular information about a subject's phenotype, missing entries complicate the application of methods to build discrete phenotypic clusters. Missing entries may arise from two situations. First, entries may be missing within an instrument (entry-level incompleteness), particularly from ADI-R and ADOS since only a subset of questions are used to form the diagnosis - unused entries are often sporadically unrecorded. Second, many subjects may be missing one or more instruments entirely (instrument-level incompleteness), since every instrument has not been administered to every individual, particularly across different studies. Our goal in this study was to simultaneously rescue entry-level and instrument-level incompleteness in order to build phenotypic clusters and ultimately map genotype to phenotype.

Generalized low rank models (GLRM)[32] provide a framework for handling structured data (i.e. ordinal, boolean) with missing entries in a graceful way. Similar to principal component analysis (PCA), the goal is to find a low-rank subspace that models the existing entries as accurately as possible. If A is an $m \times n$ matrix of data, then a GLRM is a

problem of the form

$$\underset{X \in \mathbb{R}^{m \times k}, Y \in \mathbb{R}^{k \times n}}{\text{minimize}} \sum_{i,j:\, A_{ij} \text{ is present}} l(A_{ij}, [XY]_{ij}) + r_X(X) + r_Y(Y) \tag{1}$$

where $l(a, u)$ is our loss function, which operates only on the known entries of A and r_X and r_Y are regularizers on the matrix factors X and Y respectively. This is an extremely flexible model that generalizes many known algorithms. If $l(u, a) = (u - a)^2$ and $r_X(X) = r_Y(Y) = 0$ then we have PCA. Setting $r_X(X) = \|X\|_1$ and $r_Y(Y) = \|Y\|_1$ gives us sparse PCA[19]. Changing the loss function $l(u, a) = \log(1 + \exp(-au))$ gives us logistic PCA[20]. Even the k-means clustering problem can be posed in this framework. By tailoring the loss and regularizers to our problem, we can impute entry-level and instrument-level missing values while simultaneously discovering a low-rank representation for our data. This low-rank representation can then be used to cluster individuals by phenotype and to understand correlations between items both within and across instruments.

There are many other approaches to handling missing data. One of the most common is case-wise deletion, where subjects with missing entries are removed from analysis. Such an approach is not practical for most phenotypic datasets where almost all individuals are missing at least a few entries. Another simple approach is mean or median imputation, which replaces missing values with the average or median of each item. However, this distorts item variance and makes subjects with many missing entries appear very similar which can confound downstream clustering analysis. The k-nearest neighbor algorithm can be adapted to perform imputation (KNN impute)[21]. However, when working with ordinal data, defining a distance metric to identify nearby neighbors can be challenging. For example, quantifying the difference between "sometimes" using appropriate eye contact and "never" using appropriate eye contact is quite difficult. Multiple imputation by chained equations (MICE)[22] is another effective imputation technique that iteratively imputes missing values by regressing on other items in the dataset. MICE produces multiple completed datasets, creating a distribution of imputed values for each missing entry. MICE has been used successfully on behavioral data in the fields of psychology and epidemiology[23–26]. While MICE is an effective imputation technique, unlike GLRM it does not produce a low-rank representation for the data.

GLRM leverages the correlation structure between items to impute values. It can be trained for large datasets using an alternating minimization approach. Furthermore, the algorithm is parallelizable across both subjects and items. In this study, we use the GLRM framework to model our dataset of 250 items across 16,291 subjects. We focus on three tasks: 1) imputing entry-level missing data, 2) imputing instrument-level missing data, 3) identifying phenotypic clusters using the low-rank representation produced by the GLRM.

Methods

Datasets

Data were aggregated from six sources: Autism Genetic Resource Exchange (AGRE)[28], Autism Consortium (AC), National Database for Autism Research (NDAR)[29], Simons Simplex Collection (SSC)[30], Simons Variation in Individuals Project (SVIP)[31] and a dataset of ADI-R responses by neurotypical children collected by Cognoa (COG). Individuals were included in our analysis if at least one diagnostic instrument had been administered. In total, the aggregated dataset contains item-level phenotypic data for 16,291 subjects.

Preprocessing

We analyzed data from three diagnostic instruments: ADI-R, ADOS, and SRS. All three instruments consist of a series of behavioral items, divided into three major categories identified by DSM-IV: communication, social interaction, and restricted repetitive behavior. The responses to each item lie on an ordinal scale.

ADI-R consists of 93 items: 2 free-response items, 14 age of onset items, and 77 ordinal scale items whose responses

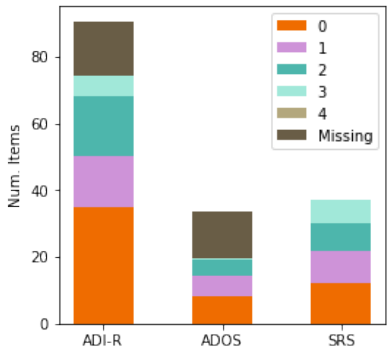
range from 0 (typical behavior) to 4 (severely atypical). We discard the free-response and age of onset items in order to focus on behavior. Of these 77 items, 62 ask for two responses: the current behavior and the lifetime behavior. This results in a total of 139 ADI-R items.

ADOS is administered as four different modules, with each module being appropriate for a different age range of children. We manually aggregated items across the four modules, combining items that were identical across multiple modules. A mapping between module-level items and the aggregated items used for analysis is available in the supplementary materials. The resulting aggregated ADOS instrument includes 46 items. Responses range from 0 (typical behavior) to 3 (severely atypical).

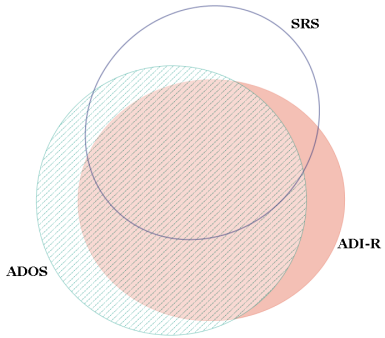
SRS consists of 65 items and is administered to the child's parent, teacher, or primary caregiver. Responses range from 0 (typical behavior) to 3 (severely atypical).

Data were aggregated and validated using a JSON schema. A JSON schema is a way to define the structure of a dataset in a clear, human-readable way that can also be programatically validated. The schema places constraints on the allowable responses to each item. Any disallowed entries found in the raw data were either manually resolved or removed. Many of the items from both instruments include an N/A option. All such entries were marked as missing.

Figure 1 shows the entry-level and instrument-level incompleteness present in our dataset, broken down by instrument. Note that instrument-level incompleteness is the larger contributor of missing entries in with 74.4% of missing entries being the result of missing instruments and the other 25.6% the result of entry-level incompleteness.



(a) Entry-level incompleteness. Distribution of responses for each instrument. Bar height indicates number of items per instrument.



(b) Instrument-level incompleteness. Venn diagram showing the overlap between subjects having data for each instrument.

Figure 1: The distribution of missing data.

Finally, we note that our dataset is highly imbalanced with respect to clinical diagnosis: 64.3% of our subjects are diagnosed with autism, 5.2% with PDD-NOS, 2.2% with Asperger, and 11.7% are clinically determined to not have autism. The remainder have a missing diagnosis.

Model

To impute missing data, we trained a low-rank model using a multi-dimensional ordinal loss. The multi-dimensional ordinal loss is a generalization of logistic PCA. In logistic PCA, we use the logistic regression loss function rather than the standard quadratic loss in order to fit a low-rank model to binary data. The rows of X represents the low dimensional representation of each subject and the columns of Y represent a separating hyperplane for each feature. Since our data are ordinal, not binary, we can extend this loss function by embedding each item in a $d-1$ dimensional space

where d is the number of possible responses for that item. Now, rather than learning a single separating hyperplane for the feature, we are learning $d-1$ separating hyperplanes, each representing the division between one ordinal response and the next.

We regularized our model by constraining the rows of X to be non-negative and to sum to 1 and adding a small amount of ℓ_2-regularization on Y. This regularization makes our model a form of fuzzy clustering. In fuzzy clustering, subjects are allowed to partially belong to multiple clusters. Here, the rows of Y represent k cluster centroids and the rows of X indicate partial cluster membership for each subject. Finally, we added an offset term to the model, so that our separating hyperplanes are not constrained to go through the origin. This is similar to the column centering that is typically done before running PCA.

We use the GLRM software package[32] to train the model via alternating minimization. We impute data for all three instruments simultaneously, using the entire dataset, even subjects who are missing one or more instruments. Items from one instrument may be used to impute items from the others. This scheme allows us to use the same model both to impute missing entries within instruments (entry-level incompleteness) and to impute entire missing instruments (instrument-level incompleteness).

Baselines

We compare the performance of our model to several baseline imputation techniques: median imputation, k-nearest neighbor imputation with $k = 3$ and MICE. We use the fancyimpute software package[33] to fit these models. For MICE, we use 75 imputations and initialize the procedure using median imputation. We use Bayesian ridge regression for the MICE predictor function and impute using the posterior predictive distribution.

Assessing Performance

We first consider the imputation accuracy for entry-level imputation - how well the model is able to impute missing entries within an instrument. To do this, we split our data entry-wise between training (90%) and testing (10%). When training our model we mask the testing entries, and then we compare the values the model imputes to the true values for these masked entries in order to evaluate performance. We use 5-fold cross-validation to select the parameter k, the dimension of the low-rank space.

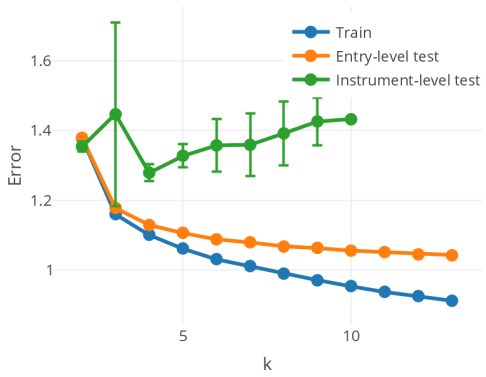
Next we consider the imputation accuracy for instrument-level imputation - how well the model is able to impute items when the entire instrument is missing. For this task, we split our data by subject into training (85%) and testing (15%). Rather than masking data entry-wise as in the first evaluation, we now mask entire instruments for our testing individuals. One-third of the testing individuals have their ADI-R items masked, one-third have their ADOS items masked, and the last third have their SRS items masked. For testing, we only select individuals with data for all three instruments in order to ensure that after masking there will still be data available for training. We again use 5-fold cross-validation to select the parameter k, the dimension of the low-rank space.

Since we are imputing an ordinal response, measuring accuracy is challenging. We present confusion matrices and use a linearly weighted Cohen's kappa to summarize overall performance. Cohen's kappa measures inter-rater agreement, taking into account the possibility that agreement may occur purely by chance. We use Cohen's kappa to compare imputed values to actual values for each model. A Cohen's kappa of 1 indicates complete agreement between imputed and actual values while a Cohen's kappa of 0 indicates complete disagreement.

The JSON schema used to aggregate the data along with the code for training the GLRM and other baseline models and evaluating performance is available at `https://github.com/walllab/PhenotypeGLRM`.

Results

Imputation Performance



(a) The effect of k on imputation error for entry-level and instrument-level imputation.

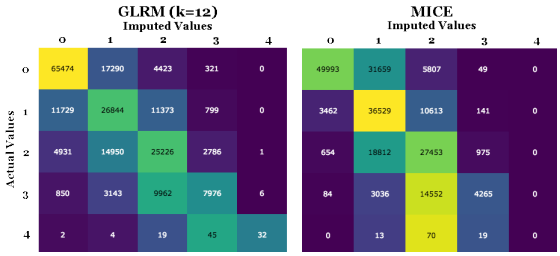
Model	Entry-level	Instrument-level
Median impute	0.302	0.323
KNN impute	0.506	0.254
MICE	0.532	0.351
GLRM ($k = 4$)	0.502	**0.445**
GLRM ($k = 12$)	**0.548**	0.427

(b) Linearly-weighted Cohen's kappa imputation performance on entry-level and instrument-level test sets.

Figure 2: Imputation performance.

We start by selecting k, the size of our low-rank model via 5-fold cross-validation. Figure 2a shows training and cross-validation error for a range of ks. A smaller low-rank space of $k = 4$ is most effective when imputing missing instruments while a larger low-rank space of $k = 12$ is most effective when imputing missing entries.

Table 2b compares our model to several baseline models: median imputation, KNN impute, and MICE. Models are compared using linearly weighted Cohen's kappa. MICE is the highest performing baseline model, so we compare the confusion matrices between our model and the MICE model in Figure 3. MICE struggles with extreme values, never imputing a value of 4 and having trouble distinguishing between 0 and 1. The GLRM is able to model the extremes more effectively. This becomes even more pronounced when imputing missing instruments. MICE imputes most values as being 1 or 2, while GLRM is able to make correct predictions across the entire ordinal range.



(a) Entry-level confusion

(b) Instrument-level confusion

Figure 3: Confusion matrices for GLRM and MICE imputation. Entries are colored by fraction of actual values.

Clustering

Due to the regularization we placed on X when building our GLRM, we can interpret our model as a form of fuzzy clustering. The rows of X represent partial cluster membership for each subject and the rows of Y represent cluster centroids. We can explore the clusters discovered during the imputation process to see if these clusters correspond to distinct autism phenotypes. For this analysis, we focus on a GLRM with $k = 4$. In cross-validation, this level of k performed best at instrument-level imputation, and since this is the largest source of incompleteness in our dataset, we expect this model to capture the structure of our data best. Because our model includes an offset term, a $k = 4$ GLRM

corresponds to three clusters (the fourth dimension of the low-rank space is used to represent the offset).

We can visualize our three clusters as the vertices of a triangle, and plot each of our subjects as a point within this triangle. Figure 4 shows a density histogram broken down by clinical diagnosis. Nearly all of our control subjects lie on the vertex corresponding to cluster 1. Asperger subjects are more diffuse and have a clear tendency towards cluster 2. The PDD-NOS region partially overlaps with that of Asperger, but is predominantly in the middle of the triangle. Finally, while our autism subjects are spread diffusely through the triangle, there are two modes: one tightly centered on the cluster 3 vertex, and the other with a similar distribution to Asperger. Some of our datasets do not distinguish between autism, PDD-NOS, and Asperger in the clinical diagnosis. Subjects from these datasets with any of the three diagnoses will be marked autism. This may explain the cluster of subjects diagnosed with autism that closely mirrors the Asperger distribution.

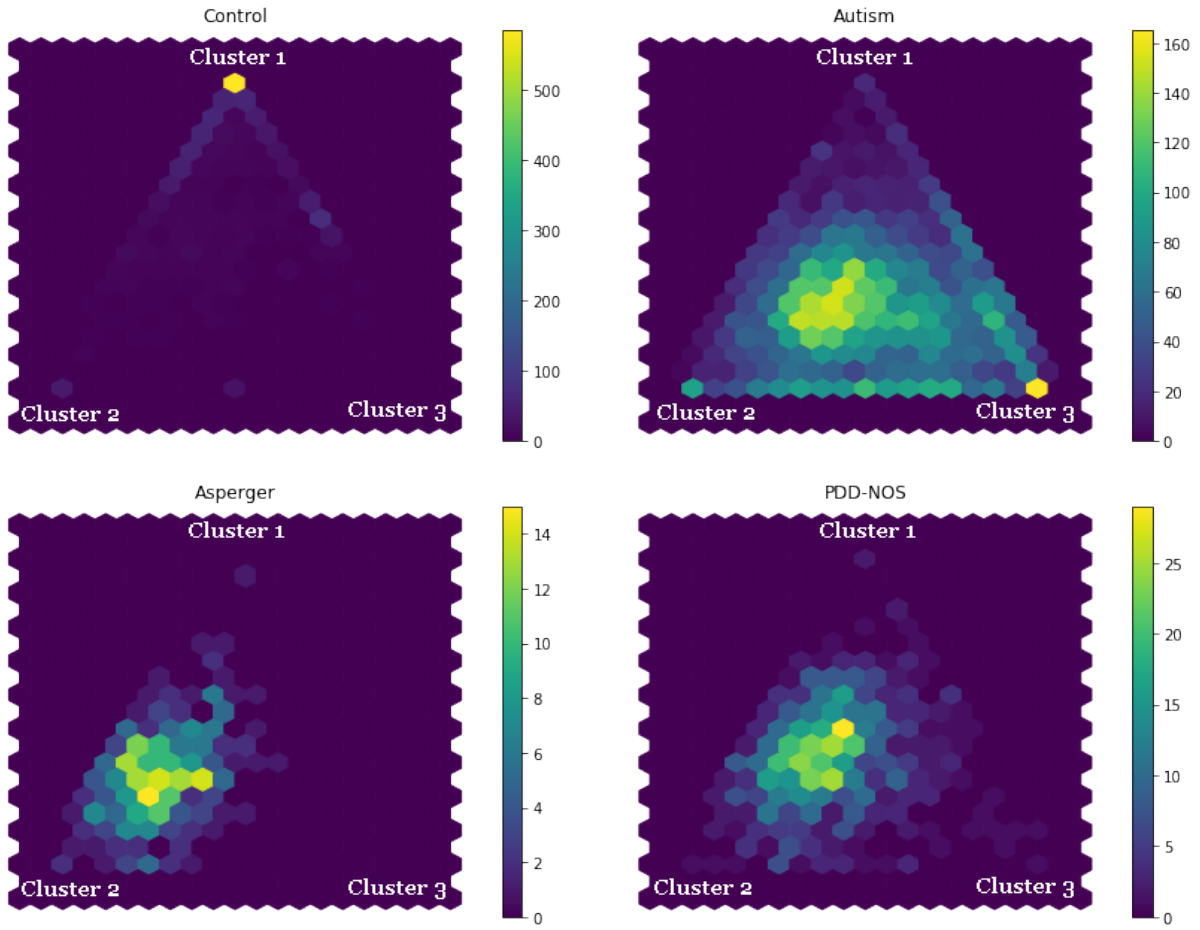


Figure 4: A histogram showing the partial cluster membership of subjects for each clinical diagnosis. The color of each tile represents the number of subjects within the tile. All points lie within a triangle whose vertices correspond to the three clusters.

We can quantitatively explore the relationship between cluster membership and clinical diagnosis by using partial membership in each cluster as a predictor of clinical diagnosis. For each cluster and each clinical diagnosis, we can set a threshold t such that subjects with partial cluster membership greater than t are predicted to have the diagnosis and subjects with partial cluster membership less than t are predicted not to have the diagnosis. We can then evaluate

the true positive and false positive rate for our prediction, generating an ROC curve by varying t. Figure 5 shows the resulting ROC curves for each cluster and each clinical diagnosis. As expected, cluster 1 partial membership is predictive of a control diagnosis. Cluster 2 partial membership is mildly predictive of an Asperger diagnosis. Partial membership in cluster 3 is predictive of an autism diagnosis. The predictive power of cluster 2 partial membership on an Asperger clinical diagnosis may be being obscured by a lack of granularity in the diagnosis for some subjects, as mentioned above. Interestingly, we see an AUC below 0.5 when trying to predict PDD-NOS or Asperger status from cluster 3 partial membership. This indicates that subjects with a diagnosis of Asperger or PDD-NOS tend to have low partial membership in cluster 3, and that this model may be able to differentiate between autism and Asperger. These results suggest that the GLRM has found a low-rank space that differentiates between control, Asperger, and autism phenotypes.

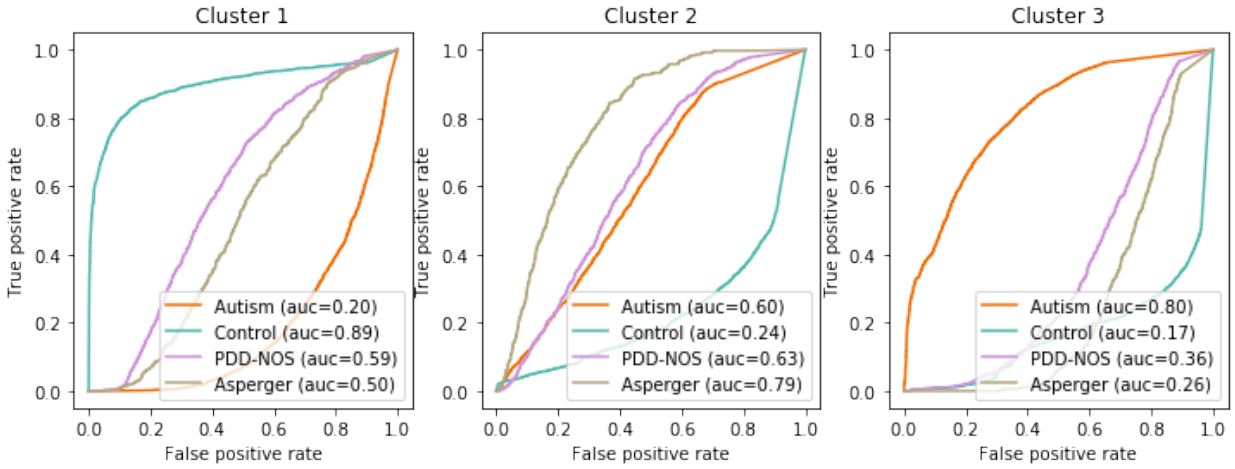


Figure 5: ROC curves showing the predictive power of partial cluster membership for each cluster when predicting each clinical diagnosis.

Discussion

GLRM outperformed other imputation methods when imputing both entry-level and instrument-level missing data. It is likely that this performance improvement is a result of using a multidimensional ordinal loss. This loss allows each ordinal value for each item to be modeled separately. This is especially useful when the distances between ordinal values are not uniform. Using our eye contact example, it may be that the difference between using appropriate eye contact always versus sometimes is much smaller than the difference between using appropriate eye contact sometimes versus never. In the extreme case, the different ordinal responses for an item may not lie on a spectrum at all but may correspond to entirely different behaviors.

We found that using a smaller-dimensional space ($k = 4$) produced better instrument-level imputation while a larger-dimensional space ($k = 12$) produced better entry-level imputation. Future work should be done to determine why this occurs. Based on the instrument-level confusion matrices for both GLRM and MICE, it is clear that both models are less likely to predict extreme values when entire instruments are missing. This could be because subjects with missing instruments provide the models with less information, making it harder to model these subjects with confidence, and causing the models to default to median imputation. We can see some evidence of this hypothesis in our cross-validation curves. Training and entry-level testing error have low variance at each value of k, unlike instrument-level testing error which varies widely from fold to fold.

GLRM has several limitations when applied to phenotypic data. First, the data may not lie in a low-rank space. The model assumes that there are a small number of prototypical individuals and that every subject in our dataset is a weighted combination of these individuals. This may not be the case for all phenotypic datasets. In particular, phe-

notypic datasets that are not disease specific may contain individuals with a large number of diverse phenotypes, and a low-rank model may not be appropriate. Another limitation of our approach lies in the non-convexity of our objective. Alternating minimization will find a local minimum solution to the objective, but does not guarantee a global minimum, making its performance dependent on the initial guess for X and Y. As with k-means, it is wise to run the algorithm several times, intitialized with different random guesses and to take the best solution to ensure a quality fit.

Preliminary analysis of the low-rank clusters produced by our model shows that this approach holds promise for differentiating subtypes of autism. We see that subjects with control, Asperger, and autism clinical diagnoses cluster into different areas of the low-rank space. The distinction between Asperger, PDD-NOS, and autism was discarded in DSM-V, however our phenotypic data suggests that there may be a phenotypic subtype of autism with Asperger-like behavior. It is notable that we see a cluster associated with Asperger, a more behaviorally homogeneous diagnosis, but not PDD-NOS which is known to be behaviorally heterogeneous[34].

Conclusion

Maximizing the utility of archived medical record data is essential for advancing the search for genetic markers of diagnostic value. Here we showed that GLRMs can impute missing entries and entire missing instruments with high fidelity. This imputation in turn enables the construction of a complete matrix of data on which we can run a variety of analyses.

Furthermore, the GLRM produces a low-rank representation of our data, which itself can be used for subsequent analysis. By adding the appropriate regularizers to our model, we can interpret this low-rank representation as a fuzzy clustering, where subjects are allowed partial cluster membership. We explored the clusters produced by our model and found that partial cluster membership was predictive of control, Asperger, and autism clinical diagnoses. There is more work to be done validating these phenotypic clusters.

Finally, we created a JSON-schema which programmatically defines the structure of the ADI-R, ADOS, and SRS instruments. This schema can be used to aggregate autism phenotype data across multiple studies. It can also be used to identify invalid entries and to correct data-entry errors.

These procedures hold promise for maximizing the value of archived clinical records from the autism population on whom we also have fully sequenced genomes. They can be adapted and used with other similarly structured clinical records to boost the value of the phenotype and enable a more complete mapping between genome and phenome.

Acknowledgements

This work was funded in part by funds to DPW from NIH (1R01EB025025-01 and 1R21HD091500-01), The Hartwell Foundation, Bill and Melinda Gates Foundation, Coulter Foundation, program grants from Stanfords Precision Health and Integrated Diagnostics Center, Bio-X Center, Predictives and Diagnostics Accelerator (SPADA) Spectrum, and Child Health Research Institute, and by the Biomedical Data Science Graduate Training grant T32 LM012409. We would like to acknowledge the Simons Foundation, the Autism Consortium, the National Database of Autism Research, the Autism Genetic Resource Exchange, and Cognoa for collecting and making available the phenotypic data used. And a special thank you to my parents Michael and Michelle and my parents-in-law Dora and Spassimir for chasing after my beautiful daughter while DPW and I wrote this paper.

References

1. Christensen DL, Baio J, Braun KV, et al. Prevalence and Characteristics of Autism Spectrum Disorder Among Children Aged 8 Years Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2012. MMWR Surveill Summ 2016;65(No. SS-3)(No. SS-3):123. DOI: http://dx.doi.org/10.15585/mmwr.ss6503a1
2. Robinson E.B., Koenen K.C., McCormick M.C., Munir K., Hallett V., Happ F., Plomin R., Ronald A. A multi-

variate twin study of autistic traits in 12-year-olds: testing the fractionable autism triad hypothesis. Behav. Genet. 2012;42:245255.

3. Sandin S., Lichtenstein P., Kuja-Halkola R., Larsson H., Hultman C.M., Reichenberg A. The familial risk of autism. JAMA. 2014;311:17701777.
4. Colvert E., Tick B., McEwen F., Stewart C., Curran S.R., Woodhouse E., Gillan N., Hallett V., Lietz S., Garnett T. Heritability of Autism Spectrum Disorder in a UK Population-Based Twin Sample. JAMA Psychiatry. 2015;72:415423.
5. Tick B., Bolton P., Happ F., Rutter M., Rijsdijk F. Heritability of autism spectrum disorders: a meta-analysis of twin studies. J. Child Psychol. Psychiatry. 2016;57:585595.
6. Yuen RK, Merico D, Bookman M, Howe JL, Thiruvahindrapuram B, Patel RV, Whitney J, Deflaux N, Bingham J, Wang Z, Pellecchia G. Whole genome sequencing resource identifies 18 new candidate genes for autism spectrum disorder. Nature neuroscience. 2017 Apr 1;20(4):602-11.
7. Jiang YH, Yuen RK, Jin X, Wang M, Chen N, Wu X, Ju J, Mei J, Shi Y, He M, Wang G. Detection of clinically relevant genetic variants in autism spectrum disorder by whole-genome sequencing. The American Journal of Human Genetics. 2013 Aug 8;93(2):249-63.
8. Yuen RK, Thiruvahindrapuram B, Merico D, Walker S, Tammimies K, Hoang N, Chrysler C, Nalpathamkalam T, Pellecchia G, Liu Y, Gazzellone MJ. Whole-genome sequencing of quartet families with autism spectrum disorder. Nature medicine. 2015 Feb 1;21(2):185-91.
9. Stavropoulos DJ, Merico D, Jobling R, Bowdin S, Monfared N, Thiruvahindrapuram B, Nalpathamkalam T, Pellecchia G, Yuen RK, Szego MJ, Hayeems RZ. Whole Genome Sequencing Expands Diagnostic Utility and Improves Clinical Management in Pediatric Medicine. NPJ genomic medicine. 2016 Jan 13;1.
10. Michaelson JJ, Shi Y, Gujral M, Zheng H, Malhotra D, Jin X, Jian M, Liu G, Greer D, Bhandari A, Wu W. Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. Cell. 2012 Dec 21;151(7):1431-42.
11. Buxbaum JD, Daly MJ, Devlin B, Lehner T, Roeder K, State MW, Autism Sequencing Consortium. The autism sequencing consortium: large-scale, high-throughput sequencing in autism spectrum disorders. Neuron. 2012 Dec 20;76(6):1052-6.
12. de la Torre-Ubieta L, Won H, Stein JL, Geschwind DH. Advancing the understanding of autism disease mechanisms through genetics. Nature medicine. 2016 Apr 1;22(4):345-61.
13. Scherer SW, Dawson G. Risk factors for autism: translating genomic discoveries into diagnostics. Human genetics. 2011 Jul 1;130(1):123-48.
14. Iossifov I, ORoak BJ, Sanders SJ, Ronemus M, Krumm N, Levy D, Stessman HA, Witherspoon KT, Vives L, Patterson KE, Smith JD. The contribution of de novo coding mutations to autism spectrum disorder. Nature. 2014 Nov 13;515(7526):216-21.
15. Sanders SJ, He X, Willsey AJ, Ercan-Sencicek AG, Samocha KE, Cicek AE, Murtha MT, Bal VH, Bishop SL, Dong S, Goldberg AP. Insights into autism spectrum disorder genomic architecture and biology from 71 risk loci. Neuron. 2015 Sep 23;87(6):1215-33.
16. Lord C, Rutter M, DiLavore PC, Risi S, Gotham K, Bishop S. Autism diagnostic observation schedule: ADOS-2. Los Angeles, CA: Western Psychological Services; 2012.
17. Lord C, Rutter M, Le Couteur A. Autism Diagnostic Interview-Revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. Journal of autism and developmental disorders. 1994 Oct 1;24(5):659-85.
18. Constantino JN, Davis SA, Todd RD, Schindler MK, Gross MM, Brophy SL, Metzger LM, Shoushtari CS, Splinter R, Reich W. Validation of a brief quantitative measure of autistic traits: comparison of the social responsiveness scale with the autism diagnostic interview-revised. Journal of autism and developmental disorders. 2003 Aug 1;33(4):427-33.
19. d'Aspremont A, Ghaoui LE, Jordan MI, Lanckriet GR. A direct formulation for sparse PCA using semidefinite programming. InAdvances in neural information processing systems 2005 (pp. 41-48).

20. Schein AI, Saul LK, Ungar LH. A generalized linear model for principal component analysis of binary data. InAISTATS 2003 Jan 6 (Vol. 3, No. 9, p. 10).

21. Batista GE, Monard MC. A Study of K-Nearest Neighbour as an Imputation Method. HIS. 2002 Dec 30;87(251-260):48.

22. Raghunathan TW, Lepkowksi JM, Van Hoewyk J, Solenbeger P. A multivariate technique for multiply imputing missing values using a sequence of regression models. Survey Methodology. 2001;27:8595.

23. Stuart EA, Azur M, Frangakis C, Leaf P. Multiple imputation with large data sets: a case study of the Children's Mental Health Initiative. American journal of epidemiology. 2009 Mar 24;169(9):1133-9.

24. Graham JW. Missing data analysis: Making it work in the real world. Annual review of psychology. 2009 Jan 10;60:549-76.

25. Klebanoff MA, Cole SR. Use of multiple imputation in the epidemiologic literature. American journal of epidemiology. 2008 Jun 30;168(4):355-7.

26. Schafer JL, Graham JW. Missing data: our view of the state of the art. Psychological methods. 2002 Jun;7(2):147.

27. Shrive FM, Stuart H, Quan H, Ghali WA. Dealing with missing data in a multi-question depression scale: a comparison of imputation methods. BMC medical research methodology. 2006 Dec 13;6(1):57.

28. Geschwind DH, Sowinski J, Lord C, Iversen P, Shestack J, Jones P, Ducat L, Spence SJ, AGRE Steering Committee. The autism genetic resource exchange: a resource for the study of autism and related neuropsychiatric conditions. American journal of human genetics. 2001 Aug;69(2):463.

29. Hall D, Huerta MF, McAuliffe MJ, Farber GK. Sharing heterogeneous data: the national database for autism research. Neuroinformatics. 2012 Oct 1;10(4):331-9.

30. Fischbach GD, Lord C. The Simons Simplex Collection: a resource for identification of autism genetic risk factors. Neuron. 2010 Oct 21;68(2):192-5.

31. Simons VIP Consortium. Simons Variation in Individuals Project (Simons VIP): a genetics-first approach to studying autism spectrum and related neurodevelopmental disorders. Neuron. 2012 Mar 22;73(6):1063-7.

32. Udell M, Horn C, Zadeh R, Boyd S. Generalized low rank models. Foundations and Trends in Machine Learning. 2016 Jun 23;9(1):1-18.

33. Alex Rubinsteyn, Sergey Feldman, Tim OD́onnell, Brett Beaulieu-Jones. hammerlab/fancyimpute: Version 0.2.0 2017. doi:10.5281/zenodo.886614.

34. Walker DR, Thompson A, Zwaigenbaum L, Goldberg J, Bryson SE, Mahoney WJ, Strawbridge CP, Szatmari P. Specifying PDD-NOS: a comparison of PDD-NOS, Asperger syndrome, and autism. Journal of the American Academy of Child & Adolescent Psychiatry. 2004 Feb 29;43(2):172-80.