DATA-DRIVEN MODELING IN BIOMEDICAL APPLICATIONS: THE SEARCH FOR BIOMARKERS IN AUTISM SPECTRUM DISORDER

Daniel P. Howsmon[1,2], Uwe Kruger[3], Stepan Melnyk[4], S. Jill James[4] and Juergen Hahn*[1,2,3]

[1]Department of Chemical & Biological Engineering, Rensselaer Polytechnic Institute, Troy, NY 12180
[2]Center for Biotechnology and Interdisciplinary Studies, Rensselaer Polytechnic Institute, Troy, NY 12180
[3]Department of Biomedical Engineering, Rensselaer Polytechnic Institute, Troy, NY 12180
[4]Department of Pediatrics, University of Arkansas for Medical Sciences, Little Rock, AR 72205

Abstract

There are a number of excellent tools for multivariable and nonlinear classification and regression tasks that are commonly employed in chemical process control and related disciplines such as chemometrics. However, these tools have not made a broad impact in biomedical and human health applications, where the majority of researchers resort to traditional, linear, univariate statistics to attempt to describe very complex phenomena. In complicated diseases such as autism spectrum disorder, it is imperative that researchers look beyond these traditional techniques and embrace more appropriate statistical techniques in order to better describe and report their findings. Furthermore, the disparity in findings between research groups encourages the use of validation procedures such as cross-validation to better ensure that results generalize to new data sets. This work showcases the use of partial least squares and Fisher discriminant analysis, as well as their kernel counterparts, for biomarker discovery in autism spectrum disorder. These techniques are able to separate participants into autism spectrum disorder and neurotypical subgroups and predict disease severity better than other approaches in the scientific literature. The wealth of statistical tools available to the chemical process community can provide great impact in non-traditional areas.

Keywords

data-driven models, Fisher Discriminant Analysis, partial least squares, kernel methods, autism spectrum disorder

Introduction

Increased global competition coupled with ever-tightening process safety, product quality, and environmental pollution targets has promoted the increased complexity of chemical processes as well as the amount of frequently recorded process data. These increases in data size and complexity also increases the difficulty of effectively process monitoring by plant operators (Kruger and Xie, 2012). Multivariable statistical process monitoring tries to address this problem by uncovering the usually small number of underlying trends hidden in this complex data and present the data in ways that are easier to visualize and comprehend. Although early work focused on improvements in the chemical process industries (e.g. (Piovoso et al., 1992; Morud, 1996)), multivariable statistical process monitoring has since impacted other manufacturing sectors including pharmaceutical (Gernaey et al., 2012), food (Grassi et al., 2014), semiconductor (Cherry and Qin, 2006), automotive (Haghani et al., 2016), and mettalurgical (Zhou et al., 2016) applications.

Just as the early chemical process industry relied on linear, univariate process monitoring, medicine traditionally relies on one measurement to separate individ-

*To whom all correspondence should be addressed *hahnj@rpi.edu* uals into healthy and disease cohorts. Examples include specific genetic mutations that lead to Tay-Sachs disease (O'Brien et al., 1971) or sickle-cell disease (Rees et al., 2010) and observations of single biochemical entities such as C-peptide level to measure loss of beta cell functionality in type 1 diabetes (Palmer et al., 2004). However, researchers studying diseases with complex and/or unknown etiology such as autism spectrum disorder (ASD) can often not find one measurement to separate the ASD and neurotypical cohorts. Moreover, interactions and nonlinearities may be important to the healthy or disease state (e.g. in heart rate time series (Goldberger et al., 2002)), making a case for the introduction of nonlinear methods. Therefore, multivariate statistical process monitoring methods have potential to make sense of these multiple measurements and present underlying trends in a way that can aid doctors in the diagnosis process.

ASD encompasses a large group of early-onset neurological diseases characterized by difficulties with social communication/interaction and expression of restricted repetitive behaviors and interests (American Psychiatric Association, 2013). In addition to these defining behavioral symptoms, individuals with ASD frequently have one or more co-occurring conditions, including intellectual disability, ADHD, speech and language delays, psychiatric diagnoses, epilepsy, sleep disorders, and gastrointestinal problems (Levy et al., 2010; Perrin et al., 2012; Pulcini et al., 2015; Saunders et al., 2015). The prevalence of ASD has been increasing at an alarming rate especially when comparing with other developmental disabilities (Braun et al., 2015) and ASD is currently estimated to affect 1.5% of the population (Centers for Disease Control and Prevention, 2012). It is associated with an impaired quality of life van Heijst and Geurts (2015) and the lifetime cost of supporting an individual with ASD amounts to \$1.4 2.4MM, depending on co-existing disorders Buescher et al. (2014).

Currently, both the etiology and pathophysiology of ASD are uncertain. Since genetic contributions to ASD have recently been estimated at 37–55% (Gaugler et al., 2014), many hypotheses surrounding possible environmental explanations have been proposed, including decreased neural synapse formation (Brigandi et al., 2015), altered folate-dependent one carbon metabolism (FOCM) and transsulfuration (TS) (James et al., 2004), and altered microbiota compositions (MacFabe, 2012). All of these proposed mechanisms involve complex mechanisms that are unlikely to be described by a single variable.

Latent variable techniques enable the discovery of important multivariate interactions, leading to improved classification and regression performance. Furthermore, latent variable techniques allow assessing the importance of individual variables and are more robust to uninformative variables. One popular latent variable technique for classification problems is Fisher Discriminant Analysis (FDA), which achieves an optimal linear separability using a typically small set of latent variables that are linear combinations of the original variable set. FDA has a long history in biological classification problems and was first used by Rao in 1948 to interpret anthropological data (Rao, 1948). Extensions of FDA, such as Kernel FDA (KFDA), exist which can take nonlinear relationships into account for classification (Mika et al., 1999). Latent variable regression techniques include partial least squares (PLS) and its nonlinear counterpart kernel PLS (KPLS) (Rosipal and Trejo, 2002). Using FDA for classification and KPLS for regression allow multivariate interactions to surface, which are often hidden when only univariate analysis is considered. To guarantee a statistically independent assessment of the multivariate classification and regression models, the presented study utilizes a cross-validatory approach, where the set of samples used for model identification does not contain samples to evaluate the performance of the identified models.

The presented work makes use of these advanced modeling and statistical analysis tools to examine metabolite data of FOCM/TS pathways in neurotypical participants (NEU) and those on the autism spectrum (ASD) as well as their siblings (SIB). Using FDA, it is possible to clearly distinguish the participants on the spectrum from their neurotypical peers and KPLS unveils a strong correlation between metabolite concentrations of these pathways and autism severity as measured by the Vineland Adaptive Behavior Composite. This work not only analyzes the largest data set of its kind of these pathways in the scientific literature (Melnyk et al., 2012), but also results in the strongest evidence to date of the association of FOCM/TS dysfunction with ASD.

Description of Data

The data used in this study comes from the Arkansas Children's Hospital Research Institute's autism IMAGE study (Melnyk et al., 2012). The protocol was approved by the Institutional Review Board at the University of

Arkansas for Medical Sciences and all parents signed informed consent. FOCM/TS metabolites from 83, 47, and 76 case (ASD), sibling (SIB), and age-matched control (NEU) children, respectively, were used for classification. The metabolites under investigation are tabulated in Table 1 and additional details of these measurements and derivations are presented in (Melnyk et al., 2012). Of the 83 participants on the autism spectrum, 55 also had Vineland II Scores recorded for use in regression analysis.

Table 1. Metabolites considered for analysis

Methionine	SAM
SAH	SAM/SAH
% DNA methylation	8-OHG
Adenosine	Homocysteine
Cysteine	Glu.-Cys.
Cys.-Gly.	tGSH
fGSH	GSSG
fGSH/GSSG	tGSH/GSSG
Chlorotyrosine	Nitrotyrosine
Tyrosine	Tryptophane
fCystine	fCysteine
fCystine/fCysteine	% oxidized glutathione

Classification into ASD, SIB, and NEU cohorts

Associating dysfunction of FOCM/TS pathways with ASD requires a distinction between or separation of ASD and NEU groups based on FOCM/TS metabolites. Therefore, cross-validatory FDA was performed using measurements of the FOCM/TS metabolites listed in Table 1. A linear classifier based on these FDA scores is then used to classify ASD and NEU participants. FDA scores and estimated probability distribution functions (PDFs) are provided in Figure 1. The cross-validated misclassification rates of only 4.9% and 3.4% for the NEU and ASD samples, respectively, eliminated more complex, nonlinear KFDA analysis from consideration.

The performance of the classifier was then evaluated on the SIB cohort, a more challenging classification problem due to partially shared genetic and environmental effects with the ASD cohort. Using all measurements in Table 1, an FDA model was trained to separate the ASD and NEU cohorts. Then, the trained FDA model was used to evaluate the SIB cohort (which was not used for training). The resulting separation of ASD, NEU, and SIB presented in Figure 2 shows a slight increase in the overlap with the ASD cohort when compared with the performance of the ASD vs. NEU classification. Furthermore, the SIB PDF shows significantly more overlap with the NEU PDF than the ASD PDF. These results support the hypothesis proposed by Melnyk et al. (2012) that the siblings of the participants on the spectrum have FOCM/TS metabolite profiles that are significantly more similar to their neurotypical peers than their siblings, even though genetically they are likely closer to their siblings than participants in the neurotypical control group.

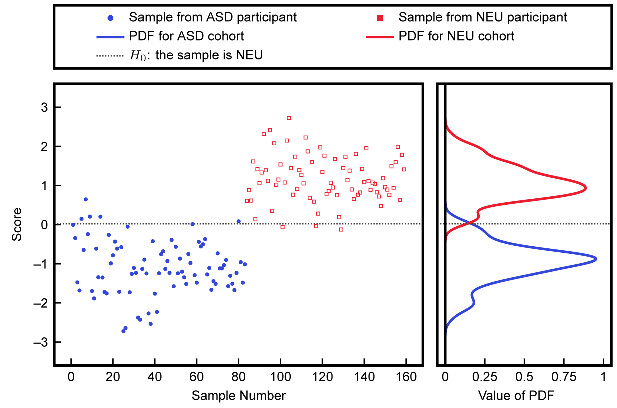

Figure 1. Classification into ASD and NEU cohorts using FDA on all FOCM/TS metabolites. The plotted scores were obtained via cross-validation and the probability distribution functions were obtained from fitting.

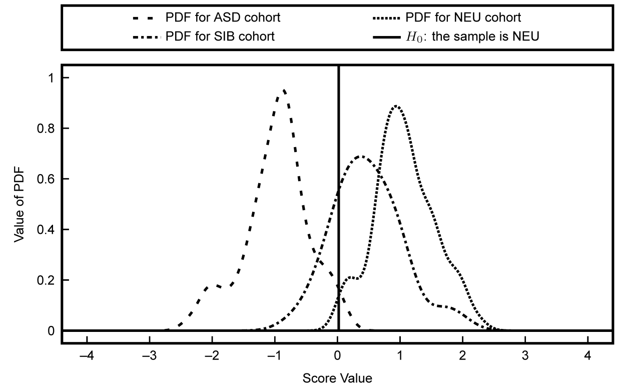

Figure 2. Classification performance on the SIB cohort. There is significantly more overlap of the SIB cohort with the NEU cohort than with the ASD cohort.

Prediction of ASD severity

In addition to separation into neurologically distinct cohorts, metabolites in the FOCM/TS pathway were investigated for predictability of autism severity. Due to the inter-dependency of pathway metabolites and possible nonlinear effects on psychological outcomes, nonlinear regression via KPLS was used to evaluate the ability of pathway metabolites to predict ASD severity (as measured by the Vineland Adaptive Behavior Composite score). All combinations of a given number of variables were evaluated for predictability. The cross-validatory R^2 of the regression was then used to determine the optimal number of variables in the regression analysis. From the results in Figure 3, the R^2 begins to decrease when more than five variables are used in the KPLS analysis. The maximum cross-validatory R^2 was 0.45, corresponding to the KPLS model with the variable combination GSSG, tGSH/GSSG, Nitrotyrosine, Tyrosine, and fCysteine used as inputs. These regression results are plotted in Figure 3. (It is important to note that a few other variable combinations provided similar results, but only the best regression model is illustrated for clarity.) This strong correlation even after cross-validation indicates the importance of FOCM/TS dysfunction in the pathophysiology of ASD.

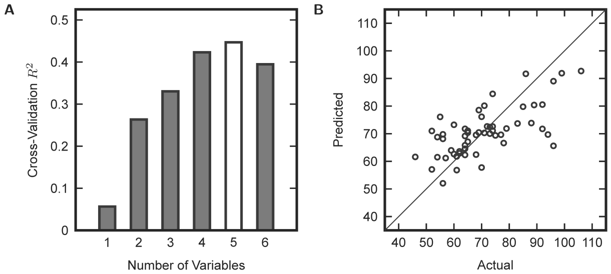

Figure 3. KPLS regression results: (a) maximum cross-validated R^2 for a given number of variables and (b) cross-validated model predictions versus actual data points for the best combination of five variables (GSSG, tGSH/GSSG, Nitrotyrosine, Tyrosine, and fCysteine).

Discussion

The multivariate statistical analysis presented herein provides unprecedented quantitative classification results for separating participants into ASD and NEU cohorts based solely on biochemical data. Existing analyses report differences in mean metabolite levels or provide qualitative illustrations of separating these two groups based on FOCM/TS metabolites (James et al., 2006, 2004; Melnyk et al., 2012). However, these strategies are not designed for classification and thus fail to successfully classify participants. Here, FDA on seven metabolites allows sufficient separation such that a linear classifier can correctly resolve 96.9% of participants. Such low misclassification rates dissuaded the use of more complex, nonlinear methods such as KFDA. Although FOCM/TS dysfunction likely does not completely detail ASD etiology, this biochemical analysis approaches the accuracy needed for a clinical diagnostic tool.

Classification performance on the SIB group fortifies the argument for FOCM/TS involvement in ASD since the large degree of shared genetic and environmental effects with the ASD population only slightly worsens the separation. The sibling recurrence rate for ASD is estimated to be 6.9–18.7% (Grønborg et al., 2013; Ozonoff et al., 2011; Constantino et al., 2010) and many siblings perform behaviorally and/or cognitively at intermediate levels between those of ASD and NEU cohorts (Constantino et al., 2010; Gizzonio et al., 2014; Ruzich et al., 2016) or express traits characteristic of ASD (Ruzich et al., 2016; **?**; **?**). Therefore, the classification performance placing the SIB group between the ASD and NEU groups, albeit much closer to the NEU group, is consistent with the broader scientific literature on psychometric analysis of siblings of people with ASD.

Nonlinear regression analysis of FOCM/TS metabolites enables prediction of key FOCM/TS metabolites that are associated with ASD severity. Based upon all variable combinations evaluated in the KPLS regression analysis, top-performing models always incorporated (1) nitrotyrosine, (2) tyrosine, (3) fGSH or tGSH/GSSG, and (4) fCysteine or fCystine/fCysteine. Interestingly, these variables are affected by high quality vitamin supplementation that also decreases ASD severity in at least a subset of cases (Frye et al., 2013; James et al., 2009; Adams et al., 2011).

Developmental pediatricians, psychologists and other professionals can effectively use the wealth of information provided by psychometric instruments such as the Vineland Adaptive Behavior Composite to diagnose and treat patients with ASD. However, these tests can rarely diagnose children under two years old since they are based solely on behavioral assessment. As it is generally acknowledged that an earlier diagnosis can lead to a more favorable outcome in the long run (Zwaigenbaum et al., 2013), the identification of biomarkers which can be used in conjunction with psychometric measurements would be of significant importance for ASD diagnosis. Furthermore, identification of these biomarkers can facilitate the understanding of these complex disorders, which offers significant potential for developing intervention strategies targeted to normalize these biomarkers in the future. However, it is important to note that these biomarkers may not simply be measurements of certain metabolites but may require nonlinear statistical analysis of the measurements, as is done in this work.

Acknowledgments

The authors thank James B. Adams for providing suggestions and comments on this work. The authors gratefully acknowledge partial financial support from the National Institutes of Health (Grant 1R01AI110642-01A1).

References

Adams, J. B., Audhya, T., McDonough-Means, S., Rubin, R. A., Quig, D., Geis, E., Gehn, E., Loresto, M., Mitchell, J., Atwood, S., Barnhouse, S., and Lee, W. (2011). Effect of a vitamin/mineral supplement on children and adults with autism. *BMC Pediatrics*, 11:111.

American Psychiatric Association (2013). *Diagnostic and Statistical Manual of Mental Disorders*. American Psychiatric Association, 5th edition.

Braun, K. V. N., Christensen, D., Doernberg, N., Schieve, L., Rice, C., Wiggins, L., Schendel, D., and Yeargin-Allsopp, M. (2015). Trends in the Prevalence of Autism Spectrum Disorder, Cerebral Palsy, Hearing Loss, Intellectual Disability, and Vision Impairment, Metropolitan Atlanta, 19912010. *PLOS ONE*, 10(4):e0124120.

Brigandi, S. A., Shao, H., Qian, S. Y., Shen, Y., Wu, B.-L., and Kang, J. X. (2015). Autistic Children Exhibit Decreased Levels of Essential Fatty Acids in Red Blood Cells. *International Journal of Molecular Sciences*, 16(5):10061–10076.

Buescher, A. V. S., Cidav, Z., Knapp, M., and Mandell, D. S. (2014). Costs of Autism Spectrum Disorders in the United Kingdom and the United States. *JAMA Pediatrics*, 168(8):721–728.

Centers for Disease Control and Prevention (2012). Prevalence of Autism Spectrum Disorders Autism and Developmental Disabilities Monitoring Network, 14 Sites, United States, 2008. *Morbidity and Mortality Weekly Report*, 61:1–19.

Cherry, G. A. and Qin, S. J. (2006). Multiblock principal component analysis based on a combined index for semiconductor fault detection and diagnosis. *IEEE Transactions on Semiconductor Manufacturing*, 19(2):159–172.

Constantino, J. N., Zhang, Y., Frazier, T., Abbacchi, A. M., and Law, P. (2010). Sibling recurrence and the genetic epidemiology of autism. *The American journal of psychiatry*, 167(11):1349–1356.

Frye, R. E., Melnyk, S., Fuchs, G., Reid, T., Jernigan, S., Pavliv, O., Hubanks, A., Gaylor, D. W., Walters, L., and James, S. J. (2013). Effectiveness of methylcobalamin and folinic acid treatment on adaptive behavior in children with autistic disorder is related to glutathione redox status. *Autism Research and Treatment*, 2013:e609705.

Gaugler, T., Klei, L., Sanders, S. J., Bodea, C. A., Goldberg, A. P., Lee, A. B., Mahajan, M., Manaa, D., Pawitan, Y., Reichert, J., Ripke, S., Sandin, S., Sklar, P., Svantesson, O., Reichenberg, A., Hultman, C. M., Devlin, B., Roeder, K., and Buxbaum, J. D. (2014). Most genetic risk for autism resides with common variation. *Nature Genetics*, 46(8):881–885.

Gernaey, K. V., Cervera-Padrell, A. E., and Woodley, J. M. (2012). A perspective on PSE in pharmaceutical process development and innovation. *Computers & Chemical Engineering*, 42:15–29.

Gizzonio, V., Avanzini, P., Fabbri-Destro, M., Campi, C., and Rizzolatti, G. (2014). Cognitive abilities in siblings of children with autism spectrum disorders. *Experimental Brain Research*, 232(7):2381–2390.

Goldberger, A. L., Amaral, L. A. N., Hausdorff, J. M., Ivanov, P. C., Peng, C.-K., and Stanley, H. E. (2002). Fractal dynamics in physiology: Alterations with disease and aging. *Proceedings of the National Academy of Sciences*, 99(suppl 1):2466–2472.

Grassi, S., Amigo, J. M., Lyndgaard, C. B., Foschino, R., and Casiraghi, E. (2014). Beer fermentation: Monitoring of process parameters by FT-NIR and multivariate data analysis. *Food Chemistry*, 155:279–286.

Grønborg, T. K., Schendel, D. E., and Parner, E. T. (2013). Recurrence of autism spectrum disorders in full- and half-siblings and trends over time: a population-based cohort study. *JAMA pediatrics*, 167(10):947–953.

Haghani, A., Jeinsch, T., Roepke, M., Ding, S. X., and Weinhold, N. (2016). Data-driven monitoring and validation of experiments on automotive engine test beds. *Control Engineering Practice*, 54:27–33.

James, S. J., Cutler, P., Melnyk, S., Jernigan, S., Janak, L., Gaylor, D. W., and Neubrander, J. A. (2004). Metabolic biomarkers of increased oxidative stress and impaired methylation capacity in children with autism. *The American Journal of Clinical Nutrition*, 80(6):1611–1617.

James, S. J., Melnyk, S., Fuchs, G., Reid, T., Jernigan, S., Pavliv, O., Hubanks, A., and Gaylor, D. W. (2009). Efficacy of methylcobalamin and folinic acid treatment on glutathione redox status in children with autism. *The American Journal of Clinical Nutrition*, 89(1):425–430.

James, S. J., Melnyk, S., Jernigan, S., Cleves, M. A., Halsted, C. H., Wong, D. H., Cutler, P., Bock, K., Boris, M., Bradstreet, J. J., Baker, S. M., and Gaylor, D. W. (2006). Metabolic endophenotype and related genotypes are associated with oxidative stress in children with autism. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 141B(8):947–956.

Kruger, U. and Xie, L. (2012). *Advances in Statistical Monitoring of Complex Multivariate Processes: With Applications in Industrial Process Control*. John Wiley & Sons.

Levy, S. E., Giarelli, E., Lee, L.-C., Schieve, L. A., Kirby, R. S., Cunniff, C., Nicholas, J., Reaven, J., and Rice, C. E. (2010). Autism spectrum disorder and co-occurring developmental, psychiatric, and medical conditions among children in multiple populations of the United States. *Journal of Developmental & Behavioral Pediatrics*, 31(4):267–275.

MacFabe, D. F. (2012). Short-chain fatty acid fermentation products of the gut microbiome: implications in autism spectrum disorders. *Microbial Ecology in Health & Disease*, 23(0).

Melnyk, S., Fuchs, G. J., Schulz, E., Lopez, M., Kahler, S. G., Fussell, J. J., Bellando, J., Pavliv, O., Rose, S., Seidel, L., Gaylor, D. W., and Jill James, S. (2012). Metabolic Imbalance Associated with Methylation Dysregulation and Oxidative Damage in Children with Autism. *Journal of Autism and Developmental Disorders*, 42(3):367–377.

Mika, S., R atsch, G., Weston, J., Sch olkopf, B., and M uller, K.-R. (1999). Fisher discriminant analysis with kernels. In *Proceedings of the Neural Networks for Signal Processing IX Workshop*, pages 41–48.

Morud, T. E. (1996). Multivariate statistical process control; example from the chemical process industry. *Journal of Chemometrics*, 10(5-6):669–675.

O'Brien, J. S., Okada, S., Fillerup, D. L., Veath, M. L., Adornato, B., Brenner, P. H., and Leroy, J. G. (1971). Tay-Sachs Disease: Prenatal Diagnosis. *Science*, 172(3978):61–64.

Ozonoff, S., Young, G. S., Carter, A., Messinger, D., Yirmiya, N., Zwaigenbaum, L., Bryson, S., Carver, L. J., Constantino, J. N., Dobkins, K., Hutman, T., Iverson, J. M., Landa, R., Rogers, S. J., Sigman, M., and Stone, W. L. (2011). Recurrence risk for autism spectrum disorders: A baby siblings research consortium study. *Pediatrics*, pages 2010–2825.

Palmer, J. P., Fleming, G. A., Greenbaum, C. J., Herold, K. C., Jansa, L. D., Kolb, H., Lachin, J. M., Polonsky, K. S., Pozzilli, P., Skyler, J. S., and Steffes, M. W. (2004). C-Peptide Is the Appropriate Outcome Measure for Type 1 Diabetes Clinical Trials to Preserve -Cell Function. *Diabetes*, 53(1):250–264.

Perrin, J. M., Coury, D. L., Hyman, S. L., Cole, L., Reynolds, A. M., and Clemons, T. (2012). Complementary and Alternative Medicine Use in a Large Pediatric Autism Sample. *Pediatrics*, 130(Supplement 2):S77–S82.

Piovoso, M. J., Kosanovich, K. A., and Yuk, J. P. (1992). Process data chemometrics. *IEEE Transactions on Instrumentation and Measurement*, 41(2):262–268.

Pulcini, C. D., Perrin, J. M., Houtrow, A. J., Sargent, J., Shui, A., and Kuhlthau, K. (2015). Examining Trends and Coexisting Conditions Among Children Qualifying for SSI Under ADHD, ASD, and ID. *Academic Pediatrics*, 15(4):439–443.

Rao, C. R. (1948). The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society. Series B (Methodological)*, 10(2):159–203.

Rees, D. C., Williams, T. N., and Gladwin, M. T. (2010). Sickle-cell disease. *The Lancet*, 376(9757):2018–2031.

Rosipal, R. and Trejo, L. J. (2002). Kernel partial least squares regression in reproducing kernel hilbert space. *J. Mach. Learn. Res.*, 2:97–123.

Ruzich, E., Allison, C., Smith, P., Watson, P., Auyeung, B., Ring, H., and Baron-Cohen, S. (2016). Subgrouping siblings of people with autism: Identifying the broader autism phenotype. *Autism Research*, 9(6):658–665.

Saunders, A., Kirk, I. J., and Waldie, K. E. (2015). Autism Spectrum Disorder and Co-Existing Conditions: A Lexical Decision Erp Study. *Clin Exp Psychol*, 1(001).

van Heijst, B. F. and Geurts, H. M. (2015). Quality of life in autism across the lifespan: A meta-analysis. *Autism*, 19(2):158–167.

Zhou, B., Ye, H., Zhang, H., and Li, M. (2016). Process monitoring of iron-making process in a blast furnace with PCA-based methods. *Control Engineering Practice*, 47:1–14.

Zwaigenbaum, L., Bryson, S., and Garon, N. (2013). Early identification of autism spectrum disorders. *Behavioural Brain Research*, 251:133–146.