REPORT

INTRODUCTION

Despite the success of machine learning in informing policies and automating decision-making, there is growing concern about the fairness (with respect to protected classes like race or gender) of the resulting policies and decisions (Miller 2015; Rudin 2013; Angwin et al. 2016; Munoz, Smith, and Patil 2016). Hence, several groups have studied how to define fairness for supervised learning (Hardt, Price, and Srebro 2016; Calders, Kamiran, and Pechenizkiy 2009; Dwork et al. 2012; Zliobaite 2015) and developed supervised learners that maintain high prediction accuracy while reducing unfairness (Berk et al. 2017; Chouldechova 2017; Hardt, Price, and Srebro 2016; Zafar et al. 2017; Olfat and Aswani 2017).

Second, unsupervised learning is often useSd as a preprocessing step for other learning methods. For instance, dimensionality reduction is sometimes performed prior to clustering, and hence fair dimensionality reduction could indirectly provide methods for fair clustering. Similarly, there are no fairness-enhancing versions of most supervised learners. Consequently, techniques for fair unsupervised learning could be combined with state-of-the-art supervised learners to develop new fair supervised learners. In fact, the past work most related to this paper concerns techniques that have been developed to generate fair data transformations that maintaining high prediction accuracy for classifiers that make predictions using the transformed data (Dwork et al. 2012; Zemel et al. 2013; Feldman et al. 2015); however, these past works are most accurately classified as supervised learning because the data transformations are computed with respect to a label used for predictions.

We briefly review this work. Dwork et al. (2012) propose a linear program that maps individuals to probability distributions over possible classifications such that similar individuals are classified similarly. Zemel et al. (2013) and Calmon et al. (2017) generate an intermediate representation for fair clustering using a non-convex formulation that is difficult to solve. Feldman et al. (2015) propose an algorithm that scales data points such that the distributions of features, conditioned on the protected attribute, are matched.

ABSTRACT

We investigate whether the standard dimensionality reduction technique of PCA inadvertently produces data representations with different fidelity for two different populations. We show on several real-world data sets, PCA has higher reconstruction error on population A than on B (for example, women versus men or lower- versus higher-educated individuals). This can happen even when the data set has a similar number of samples from A and B. This motivates our study of dimensionality reduction techniques which maintain similar fidelity for A and B. We define the notion of Fair PCA and give a polynomial-time algorithm for finding a low dimensional representation of the data which is nearly-optimal with respect to this measure. Finally, we show on real-world data sets that our algorithm can be used to efficiently generate a fair low dimensional representation of the data.

BACKGROUND

Outline and novel contributions

This paper studies fairness for principal component analysis (PCA), and we make three main contributions: First, in Section 3 we propose and motivate a novel quantitative definition of fairness for dimensionality reduction. Second, in Section 5 we develop convex optimization formulations for fair PCA and fair kernel PCA. Third, in Section 6 we demonstrate the efficacy of our semidefinite programming (SDP) formulations using several datasets, including using fair PCA as preprocessing to perform fair (with respect to age) clustering of health data that can impact health insurance rates.

Notation

Let [n] = {1,...,n}, **1**(u) be the Heaviside function, and let **e** be the vector whose entries are all 1. A positive semidefinite matrix U with dimensions q × q is denoted U ∈ Sq+ (or  when dimensions are clear). We use the notation h·,·i

to denote the inner product and I the identity matrix.

Our data consists of 2-tuples (xi,zi) for i = 1,...,n, where xi ∈ Rp are a set of features, and zi ∈ {−1,1} label a protected class. For a matrix W, the i-th row of W is denoted Wi. Let X ∈ Rn×p and Z ∈ Rn be the matrices so that Xi = (xi − x)T and Zi = zi, where. Also,

we use the notation Π : Rp → Rd to refer to a function that performs dimensionality reduction on the xi data, where d is the dimension of the dimensionality-reduced data.

Let P = {i : zi = +1} be the set of indices where the protected class is positive, and similarly let N = {i : zi = −1} be the set of indices where the protected class is negative. We use #P and #N for the cardinality of these sets. Furthermore, we define X+ to be the matrix whose rows are xTi for i ∈ P, and we similarly define to be the matrix whose rows are xTi for i ∈ N. Next, let Σ+ and  be the sample covariances matrices of X+ and X−, respectively.

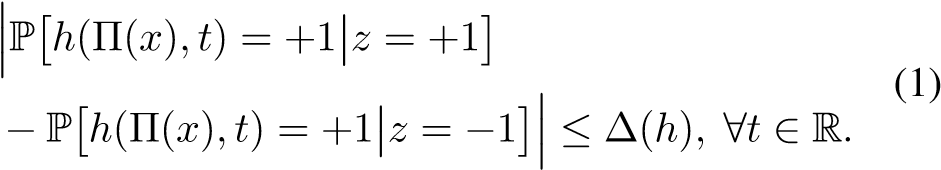
For a kernel function k : Rp×Rp → R+, let K(X,X0) = [k(Xi,Xj0)]ij be the transformed Gram matrix. Since the kernel trick involves replacing xTi xj with K(xi,xj), the benefit of the above notation is it allows us to replace X(X0)T with K(X,X0) as part of applying the kernel trick.

Fairness for dimensionality reduction

Definitions of fairness for supervised learning (Hardt, Price, and Srebro 2016; Dwork et al. 2012; Calders, Kamiran, and Pechenizkiy 2009; Zliobaite 2015; Feldman et al. 2015; Chouldechova 2017; Berk et al. 2017) specify that predictions conditioned on the protected class are roughly equivalent. However, these fairness notions cannot be used for dimensionality reduction because predictions are not made in unsupervised learning. This section discusses fairness for dimensionality reduction. We first provide and motivate a general quantitative definition of fairness, and then present several important cases of this definition.

General definition

Consider a fixed classifier h(u,t) : Rd×R → {−1,+1} that inputs features u ∈ Rd and a threshold t, and predicts the protected class z ∈ {−1,+1}. We say that a dimensionality reduction Π : Rp → Rd is ∆(h)-fair if



Moreover, let F be a family of classifiers. Then we say that a dimensionality reduction Π : Rp → Rd is ∆(F)-fair if it is ∆(h)-fair for all classifiers h ∈ F.

Our fairness definition can be interpreted via classification: Observe that the first term in the left-hand-side of (1) is the true positive rate of the classifier h in predicting the protected class using the dimensionality-reduced variable Π(x) at threshold t, and the second term is the corresponding false positive rate. Thus, ∆(h) in our definition (1) can be interpreted as bounding the accuracy of the classifier h in predicting the protected class using the dimensionality-reduced variable Π(x).

Note that eq. (1) is analogous to disparate impact for classifiers (Calders, Kamiran, and Pechenizkiy 2009; Feldman et al. 2015), where we require that treatment not vary at all between protected classes. This has often been criticized as too strict of a notion in classification, and so alternate notions of fairness have been developed, such as equalized odds and equalized opportunity (Hardt, Price, and Srebro 2016). Instead of equalizing all treatment across protected classes, these notions instead focus on equalizing error rates; for example, in the case of lending, equalized odds would require nondiscrimination among all applicants of similar FICO scores, whereas disparate impact would require nondiscrimination among all applicants. This may be preferred in cases where y and z are strongly correlated. In any case, it can easily be incorporated into our model by simply further conditioning the two terms on the left-hand-side of eq. (1) on the main label, y.

Motivation

The above is a meaningful definition of fairness for dimensionality reduction because it implies that a supervised learner using fair dimensionality-reduced data will itself be fair. This is formalized below:

Proposition 1. Suppose we have a family of classifiers F and a dimensionality reduction Π that is ∆(F)-fair. Then any classifier that is selected from F to predict a label y ∈ {−1,+1} using Π(x) as features will have disparate impact less than ∆(F).

Proposition 1 follows directly from our definition of fairness. We anticipate that in most situations the goal of the dimensionality reduction would not be to explicitly predict the protected class. Thus, our approach of bounding intentional discrimination on z represents a conservative bound on any discrimination that may incidentally arise when performing classificiation using the family F or when deriving qualitative insights form the results of unsupervised learning.

Special cases

An important special case of our definition occurs for the family Fc = {h(u,t) = **1**(u ≤ w + t) : w ∈ Rd},

where the inequality in this expression should be interpreted element-wise. In this case, our definition can be rewritten

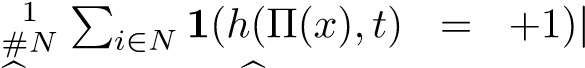
, where F is the cumulative distribution function (c.d.f.) of the random variable in the subscript. Restated, for this family our definition is equivalent to saying ∆(F) is a bound on the Kolmogorov distance between Π(x) conditioned on z = ±1 (i.e., the left-hand side of the above equation).

Other important cases are the family of linear support vector machines (SVM’s) Fv = {h(u,t) = **1**(wTu − t ≤ 0) : w ∈ Rd} and the family of kernel SVM’s Fk for a fixed kernel k. These important cases are used in Section 5 to propose formulations for fair PCA and fair kernel PCA.

Next, we briefly discuss empirical estimation of

∆(F). An empirical estimate of ∆(h) is given by



 . Similarly, we define

∆(F) = sup{∆(h) | h ∈ F}. Last, note that we can provide high probability bounds of the actual fairness level in terms of these empirical estimates:

Proposition 2. Consider a fixed family of classifiers F. If the samples (xi,zi) are i.i.d., then for any δ > 0 we have with probability at least 1 − exp(−nδ2/2) that ∆(F) ≤

∆(b F) + 8pV(F)/n + δ, where V(F) is the VC dimension of the family F.

This result follows from the triangle inequality, bounding ∆(F) with ∆(ˆ F) plus a generalization error, for which there are standard bounds via Dudley’s entropy integral (Wainwright 2017).

Remark 1. Recall that V(Fc) = d+1 (Shorack and Wellner

2009), and that V(Fv) = d + 1 (Wainwright 2017). This means  and ∆(b Fv) will be accurate when n is large relative to.

approach to designing an algorithm for fair PCA will begin by first studying the convex relaxation of a nonconvex optimization problem whose solution provides the projection defined by PCA. First, note that computation of the first d PCA components vi for i = 1,...,d can be written as the following non-convex optimization problem:

.

Now suppose we define the matrix, and note

. Thus, we can rewrite the above optimization problem as

 rank(

In the above problem, we should interpret the optimal P∗ to be the projection matrix that projects x ∈ Rp onto the d PCA components (still in the original p-dimensional space). Next, we consider a convex relaxation of (2). Since, the usual nuclear norm relaxation is equivalent to the trace

(Recht, Fazel, and Parrilo 2010). So our convex relaxation is

 trace(. (3)

Note that this base model is the same as that used by (Arora, Cotter, and Srebro 2013). The following result shows that we can recover the first d PCA components from any P∗ that solves (3).

Theorem 1. Let P∗ be an optimal solution of (3), and consider its diagonalization:, where vi is an orthonormal basis, and (without loss of generality) the λ∗i are in non-increasing order. Then the positive semidefinite  is an optimal solution to (2).

Proof. We consider two cases. First, if rank(P∗) ≤ d then λ∗i ∈ {0,1} or viTXTXvi = 0 for all i, since otherwise we could increase λ∗i if viTXTXvi > 0 (or vice versa) to improve the objective while maintaining feasibility. It follows that hXTX,P∗i = hXTX,P∗∗i. This means that P∗∗ is optimal for (3); since it is also feasible for (2), we are done. Second, if rank(P∗) > d then 0 < λ∗d < 1 since the λ∗i are ordered.

Consider.

Note that P˜ is feasible for (3), and that P∗ is a strict convex combination of P∗∗ and P˜. All points between P˜ and P∗∗ are feasible by convexity, and so the optimality of P∗ implies that P∗∗ and P˜ must also be optimal for (3) by linearity of the objective (i.e., at least one must have objective value no less than that of P∗, but if one had a strictly better objective value than the other, then no strict convex combination of the two could be optimal). The result then follows from the optimality of P∗∗ for (3) and feasibility for (2).

We conclude this section with two useful results on the spectral norm k · k2 of a symmetric matrix.

Theorem 2. Let Q be a symmetric matrix, and suppose ϕ ≥ kQk2. Then kQk2 = max{kQ + ϕIk2,k − Q + ϕIk2} − ϕ.

Proof. First diagonalize , with orthonormal basis vi and (without loss of generality) λi in non-increasing order. Then 

. But by con-

struction λi + ϕ ≥ 0 and −λi + ϕ ≥ 0 for all i = 1,...,p. Thus kQ + ϕIk2 = λ1 + ϕ and k − Q + ϕIk2 = −λp + ϕ.

RELATED WORK

Fair clustering of health data

Health insurance companies are considering the use of patterns of physical activity as measured by activity trackers in order to adjust health insurance rates of specific individuals (Sallis, Bauman, and Pratt 1998; Paluch and Tuzovic 2017). In fact, a recent clustering analysis found that different patterns of physical activity are correlated with different health outcomes (Fukuoka et al. 2018). The objective of a health insurer in clustering activity data would be to find qualitative trends in an individual’s physical activity that help categorize the risks that that customer portends. That is, individuals within these activity clusters are likely to incur similar levels of medical costs, and so it would be beneficial to engineer easy-to-spot features that can help insurers bucket customers. However, health insurance rates must satisfy a number of legal fairness considerations with respect to gender, race, and age. This means that an insurance company may be found legally liable if the patterns used to adjust rates result in an unreasonably-negative impact on individuals of a specific gender, race, or age. Thus, an insurer may be interested in a feature engineering method to bucket customers while minimizing discrimination on protected attributes. Motivated by this, we use FPCA to perform a fair clustering of physical activity. *Our goal is to find discernible qualitative trends in activity which are indicative of an individual’s activity patterns, and thus health risks, but fair with respect to age.*

We use minute-level data from the the National Health and

Nutrition Examination Survey (NHANES) from 2005–2006

(Centers for Desease Control and Prevention (CDC). National Center for Health Statistics (NCHS). 2018), on the intensity levels of the physical activity of about 6000 women, mea-

1 2 3

(Mansouri et al. 2013) (Horton and Nakai 1996) (Tsanas and Xifara 2012) 4 (Bock et al. 2004) 5 (Smith et al. 1988)

6 7 8

(Angwin et al. 2016) (Thompson et al. 2013) (Yeh and

Lien 2009) 9 (Cortez et al. 2009)

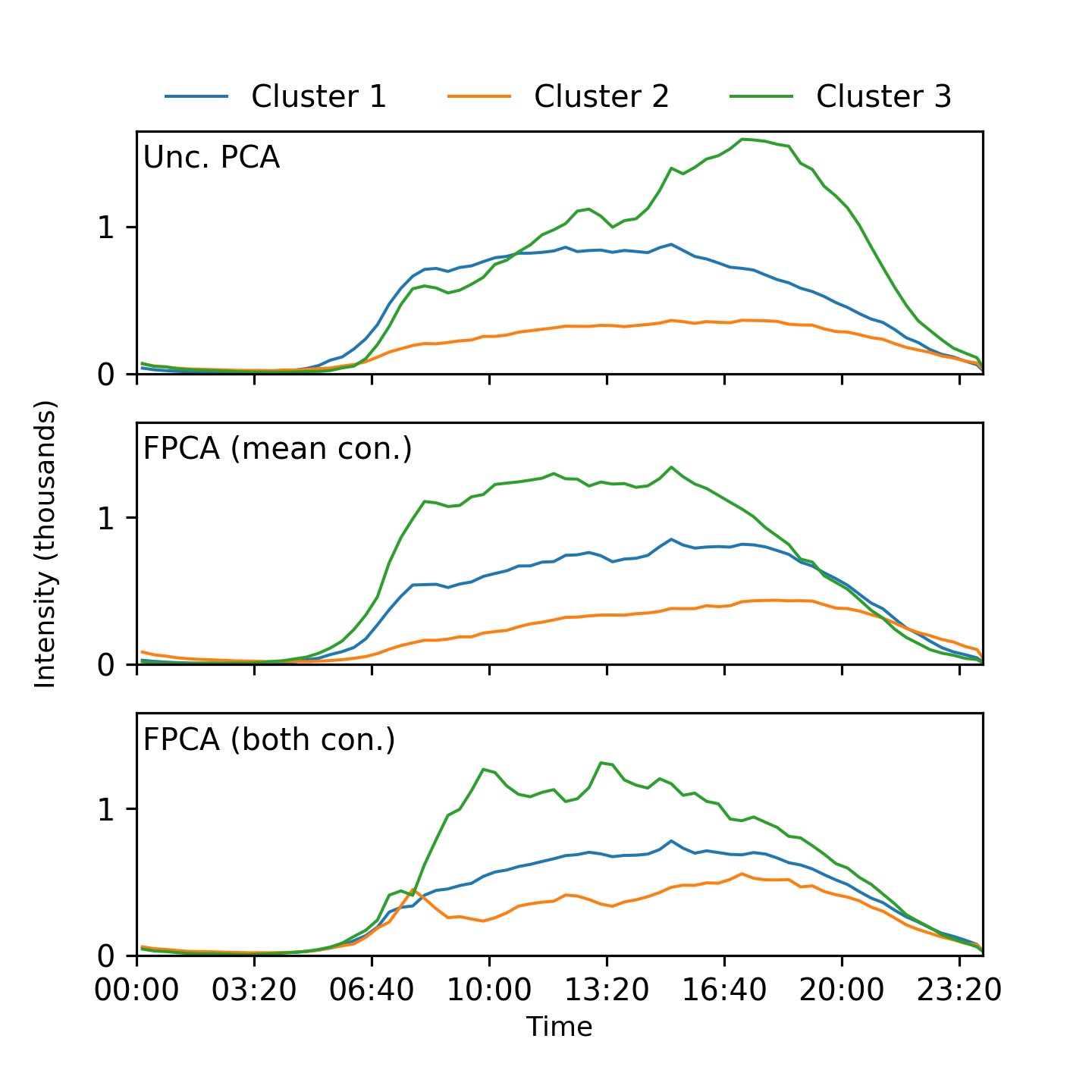


Figure 3: The mean physical activity intensities, plotted throughout a day, of the clusters generated after dimensionality reduction through PCA, FPCA with the mean constraint, and FPCA with both constraints. In each plot, each line represents the average activity level of the members of one cluster.

sured over a week via an accelerometer. In this example, we consider age to be our protected variable, specifically whether an individual is above or below 40 years of age. We exclude weekends from our analysis, and average, over weekdays, the activity data by individual into 20-minute buckets. Thus, for each participant, we have data describing her average activity throughout an average day. We exclude individuals under 12 years of age, and those who display more than 16 hours of zero activity after averaging. The top 1% most active participants, and corrupted data, were also excluded. Finally, data points corrupted or inexact due to accelerometer malfunctioning were excluded. This preprocessing mirrors that of Fukuoka et al. and reflects practical concerns of insurers as well as the patchiness of accelerometer data.

PCA is sometimes used as a preprocessing step prior to clustering in order to expedite runtime. In this spirit, we find the top five principal components through PCA, FPCA with mean constraint, and FPCA with both constraints, with *δ* = 0 and *µ* = 0*.*1 throughout. Then we conduct *k*-means clustering (with *k* = 3) on the dimensionality-reduced data for each case. Figure 3 displays the averaged physical activity patterns for the each of the clusters in each of the cases. Furthermore, Table 2 documents the proportion of each cluster comprised of examinees over 40. We note that the clusters found under an unconstrained PCA are most distinguishable after 3:00 PM, so an insurer interested in profiling an individual’s risk would largely consider their activity in the evenings. However, we may observe in Table 2 that this approach results in notable age discrimination between buckets, opening the Table 2: The proportion of each cluster that are over 40 years of age. 36.05% of all respondents are over 40. The final row displays the standard deviation of the numbers in the first three. The most fair solution would be the same age composition in all clusters, so this is a reasonable fairness metric.

|  |  |  |  |
| --- | --- | --- | --- |
|  | UNC. | MEAN | BOTH |
| Cluster 1 | 43.18% | 33.54% | 35.61% |
| Cluster 2 | 32.94% | 38.64% | 36.11% |
| Cluster 3 | 8.71% | 33.32% | 37.28% |
| Std. Dev | 14.87% | 2.46% | 1.79% |

insurer to the risk of illegal price discrimination. On the hand, the second and third plots in Figure 3 and columns in Table 2 suggest that clustering customers based on their activity during the workday, between 8:00 AM and 5:00 PM, would be less prone to discrimination.

REFERENCE

Lichman, M. 2013. UCI machine learning repository.

Mansouri, K.; Ringsted, T.; Ballabio, D.; Todeschini, R.; and Consonni, V. 2013. Quantitative structure–activity relationship models for ready biodegradability of chemicals. Journal of chemical information and modeling 53(4):867–878.

Massart, P. 2007. Concentration inequalities and model selection, volume 6. Springer.

Miller, C. C. 2015. Can an algorithm hire better than a human. The New York Times 25.

Munoz, C.; Smith, M.; and Patil, D. 2016. Big data: A report on algorithmic systems, opportunity, and civil rights. Executive Office of the President. The White House.

Olfat, M., and Aswani, A. 2017. Spectral algorithms for computing fair support vector machines. arXiv preprint arXiv:1710.05895.

Paluch, S., and Tuzovic, S. 2017. Leveraging pushed selftracking in the health insurance industry: How do individuals perceive smart wearables offered by insurance organization?

Recht, B.; Fazel, M.; and Parrilo, P. A. 2010. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. SIAM review 52(3):471–501.

Rudin, C. 2013. Predictive policing using machine learning to detect patterns of crime. Wired Magazine, August.

Sallis, J.; Bauman, A.; and Pratt, M. 1998. Environmental and policy interventions to promote physical activity a. American journal of preventive medicine 15(4):379–397.

Shorack, G. R., and Wellner, J. A. 2009. Empirical processes with applications to statistics, volume 59. SIAM.

Smith, J. W.; Everhart, J.; Dickson, W.; Knowler, W.; and

Johannes, R. 1988. Using the adap learning algorithm to forecast the onset of diabetes mellitus. In Proceedings of the Annual Symposium on Computer Application in Medical Care, 261. American Medical Informatics Association.

Thompson, J. J.; Blair, M. R.; Chen, L.; and Henrey, A. J. 2013. Video game telemetry as a critical tool in the study of complex skill learning. PloS one 8(9):e75129.

Tsanas, A., and Xifara, A. 2012. Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools. Energy and Buildings 49:560–567.

Wainwright, M. 2017. High-dimensional statistics: A nonasymptotic viewpoint.

Yeh, I.-C., and Lien, C.-h. 2009. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Systems with Applications 36(2):2473–2480.

Zafar, M. B.; Valera, I.; Rodriguez, M. G.; and Gummadi, K. P. 2017. Fairness constraints: Mechanisms for fair classification. In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics.

Zemel, R.; Wu, Y.; Swersky, K.; Pitassi, T.; and Dwork, C.

2013. Learning fair representations. In Proceedings of the 30th International Conference on Machine Learning (ICML13), 325–333.