Convex Formulations for Fair Principal Component Analysis

Matt Olfat,[1] **Anil Aswani**[1]

[1]UC Berkeley
Berkeley, CA 94720
molfat@berkeley.edu, aaswani@berkeley.edu

Abstract

Though there is a growing literature on fairness for supervised learning, incorporating fairness into unsupervised learning has been less well-studied. This paper studies fairness in the context of principal component analysis (PCA). We first define fairness for dimensionality reduction, and our definition can be interpreted as saying a reduction is fair if information about a protected class (e.g., race or gender) cannot be inferred from the dimensionality-reduced data points. Next, we develop convex optimization formulations that can improve the fairness (with respect to our definition) of PCA and kernel PCA. These formulations are semidefinite programs, and we demonstrate their effectiveness using several datasets. We conclude by showing how our approach can be used to perform a fair (with respect to age) clustering of health data that may be used to set health insurance rates.

1 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).

However, fairness in the context of unsupervised learning has not been well-studied to date. One reason is that fairness is easier to define in the supervised setting, where positive predictions can often be mapped to positive decisions (e.g., an individual who is predicted to not default on a loan maps to the individual being offered a loan). Such notions of fairness cannot be used for unsupervised learning, which does not involve making predictions. A second reason is that it is not obvious why fairness is an issue of relevance to unsupervised learning, since predictions are not made.

1.1 Relevance of fairness to unsupervised learning

Fairness is important to unsupervised learning: First, unsupervised learning is often used to generate qualitative insights from data. Examples include visualizing high-dimensional data through dimensionality-reduction and clustering data to identify common trends or behaviors. If such qualitative insights are used to generate policies, then there is an opportunity to introduce unfairness in the resulting policies if the results of the unsupervised learning are unequal for different protected classes (e.g., race or gender). We present such an example in Section 6 using individual health data.

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; however, this approach makes the restrictive assumption that predictions are monotonic with respect to each dimension. Chierichetti et al. (2017) directly perform fair clustering by approximating an NP-hard preprocessing step; however, this approach only applies to specific clustering techniques whereas the approach we develop can be used with arbitrary

clustering techniques. Finally, a series of work has emerged using auto-encoders in the the context of deep classification. This area is promising, but suffers from a lack of theoretical guarantees and is further oriented almost entirely around an explicit classification task (Beutel et al. 2017; Zhang, Lemoine, and Mitchell 2018). In contrast, our method has applications in both supervised and unsupervised learning tasks, and well-defined convergence and optimality guarantees.

1.2 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.

2 Notation

Let $[n] = \{1, \ldots, n\}$, $\mathbf{1}(u)$ be the Heaviside function, and let $\mathbf{e}$ be the vector whose entries are all 1. A positive semidefinite matrix U with dimensions $q \times q$ is denoted $U \in \mathbb{S}_+^q$ (or $U \succeq 0$ when dimensions are clear). We use the notation $\langle \cdot, \cdot \rangle$ to denote the inner product and $\mathbb{I}$ the identity matrix.

Our data consists of 2-tuples (x_i, z_i) for $i = 1, \ldots, n$, where $x_i \in \mathbb{R}^p$ are a set of features, and $z_i \in \{-1, 1\}$ label a protected class. For a matrix W, the i-th row of W is denoted W_i. Let $X \in \mathbb{R}^{n \times p}$ and $Z \in \mathbb{R}^n$ be the matrices so that $X_i = (x_i - \overline{x})^\mathsf{T}$ and $Z_i = z_i$, where $\overline{x} = \frac{1}{n} \sum_i x_i$. Also, we use the notation $\Pi : \mathbb{R}^p \to \mathbb{R}^d$ to refer to a function that performs dimensionality reduction on the x_i data, where d is the dimension of the dimensionality-reduced data.

Let $P = \{i : z_i = +1\}$ be the set of indices where the protected class is positive, and similarly let $N = \{i : z_i = -1\}$ 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 x_i^T for $i \in P$, and we similarly define X_- to be the matrix whose rows are x_i^T for $i \in N$. Next, let $\widehat{\Sigma}_+$ and $\widehat{\Sigma}_-$ be the sample covariances matrices of X_+ and X_-, respectively.

For a kernel function $k : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}_+$, let $K(X, X') = [k(X_i, X_j')]_{ij}$ be the transformed Gram matrix. Since the *kernel trick* involves replacing $x_i^\mathsf{T} x_j$ with $K(x_i, x_j)$, the benefit of the above notation is it allows us to replace $X(X')^\mathsf{T}$ with $K(X, X')$ as part of applying the kernel trick.

3 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.

3.1 General definition

Consider a fixed classifier $h(u, t) : \mathbb{R}^d \times \mathbb{R} \to \{-1, +1\}$ that inputs features $u \in \mathbb{R}^d$ and a threshold t, and predicts the protected class $z \in \{-1, +1\}$. We say that a dimensionality reduction $\Pi : \mathbb{R}^p \to \mathbb{R}^d$ is $\Delta(h)$-fair if

$$
\begin{aligned}
&\Big| \mathbb{P}\big[h(\Pi(x), t) = +1 \big| z = +1\big] \\
&- \mathbb{P}\big[h(\Pi(x), t) = +1 \big| z = -1\big] \Big| \leq \Delta(h), \ \forall t \in \mathbb{R}.
\end{aligned}
\tag{1}
$$

Moreover, let $\mathcal{F}$ be a family of classifiers. Then we say that a dimensionality reduction $\Pi : \mathbb{R}^p \to \mathbb{R}^d$ is $\Delta(\mathcal{F})$-fair if it is $\Delta(h)$-fair for all classifiers $h \in \mathcal{F}$.

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 $\Pi(x)$ at threshold t, and the second term is the corresponding false positive rate. Thus, $\Delta(h)$ 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 $\Pi(x)$.

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.

3.2 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 $\mathcal{F}$ and a dimensionality reduction Π that is $\Delta(\mathcal{F})$-fair. Then any classifier that is selected from $\mathcal{F}$ to predict a label $y \in \{-1, +1\}$ using $\Pi(x)$ as features will have disparate impact less than $\Delta(\mathcal{F})$.*

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 classification using the family $\mathcal{F}$ or when deriving qualitative insights form the results of unsupervised learning.

3.3 Special cases

An important special case of our definition occurs for the family $\mathcal{F}_c = \{h(u,t) = \mathbf{1}(u \leq w + t) : w \in \mathbb{R}^d\}$, where the inequality in this expression should be interpreted element-wise. In this case, our definition can be rewritten as $\sup_u \left| F_{\Pi(x)|z=+1}(u) - F_{\Pi(x)|z=-1}(u) \right| \leq \Delta(\mathcal{F}_c)$, 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 $\Delta(\mathcal{F})$ is a bound on the Kolmogorov distance between $\Pi(x)$ conditioned on $z = \pm 1$ (i.e., the left-hand side of the above equation).

Other important cases are the family of linear support vector machines (SVM's) $\mathcal{F}_v = \{h(u,t) = \mathbf{1}(w^\mathsf{T} u - t \leq 0) : w \in \mathbb{R}^d\}$ and the family of kernel SVM's $\mathcal{F}_k$ 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 $\Delta(\mathcal{F})$. An empirical estimate of $\Delta(h)$ is given by $\widehat{\Delta}(h) = \sup_t \left| \frac{1}{\#P} \sum_{i \in P} \mathbf{1}(h(\Pi(x),t) = +1) - \frac{1}{\#N} \sum_{i \in N} \mathbf{1}(h(\Pi(x),t) = +1) \right|$. Similarly, we define $\widehat{\Delta}(\mathcal{F}) = \sup\{\widehat{\Delta}(h) \mid h \in \mathcal{F}\}$. 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 $\mathcal{F}$. If the samples (x_i, z_i) are i.i.d., then for any $\delta > 0$ we have with probability at least $1 - \exp(-n\delta^2/2)$ that $\Delta(\mathcal{F}) \leq \widehat{\Delta}(\mathcal{F}) + 8\sqrt{\mathcal{V}(\mathcal{F})/n} + \delta$, where $\mathcal{V}(\mathcal{F})$ is the VC dimension of the family $\mathcal{F}$.*

This result follows from the triangle inequality, bounding $\Delta(\mathcal{F})$ with $\widehat{\Delta}(\mathcal{F})$ plus a generalization error, for which there are standard bounds via Dudley's entropy integral (Wainwright 2017).

Remark 1. *Recall that $\mathcal{V}(\mathcal{F}_c) = d+1$ (Shorack and Wellner 2009), and that $\mathcal{V}(\mathcal{F}_v) = d+1$ (Wainwright 2017). This means $\widehat{\Delta}(\mathcal{F}_c)$ and $\widehat{\Delta}(\mathcal{F}_v)$ will be accurate when n is large relative to d.*

4 Projection defined by PCA

Our approach to designing an algorithm for fair PCA will begin by first studying the convex relaxation of a non-convex optimization problem whose solution provides the projection defined by PCA. First, note that computation of the first d PCA components v_i for $i = 1, \ldots, d$ can be written as the following non-convex optimization problem: $\max\{\sum_{i=1}^d v_i^\mathsf{T} X^\mathsf{T} X v_i \mid \|v_i\|_2 \leq 1, \ v_i^\mathsf{T} v_j = 0, \text{ for } i \neq j\}$. Now suppose we define the matrix $P = \sum_{i=1}^d v_i v_i^\mathsf{T}$, and note $\sum_{i=1}^d v_i^\mathsf{T} X^\mathsf{T} X v_i = \sum_{i=1}^d \langle X^\mathsf{T} X, v_i v_i^\mathsf{T} \rangle = \langle X^\mathsf{T} X, P \rangle$. Thus,

we can rewrite the above optimization problem as

$$\max\left\{ \langle X^\mathsf{T} X, P \rangle \mid \mathrm{rank}(P) \leq d, \ \mathbb{I} \succeq P \succeq 0 \right\}. \quad (2)$$

In the above problem, we should interpret the optimal P^* to be the projection matrix that projects $x \in \mathbb{R}^p$ onto the d PCA components (still in the original p-dimensional space). Next, we consider a convex relaxation of (2). Since $\mathbb{I} - P \succeq 0$, the usual nuclear norm relaxation is equivalent to the trace (Recht, Fazel, and Parrilo 2010). So our convex relaxation is

$$\max\left\{ \langle X^\mathsf{T} X, P \rangle \mid \mathrm{trace}(P) \leq d, \ \mathbb{I} \succeq P \succeq 0 \right\}. \quad (3)$$

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: $P^* = \sum_{i=1}^p \lambda_i^* v_i v_i^\mathsf{T}$, where v_i is an orthonormal basis, and (without loss of generality) the λ_i^* are in non-increasing order. Then the positive semidefinite $P^{**} \triangleq \sum_{i=1}^d v_i v_i^\mathsf{T}$ is an optimal solution to (2).*

Proof. We consider two cases. First, if $\mathrm{rank}(P^*) \leq d$ then $\lambda_i^* \in \{0,1\}$ or $v_i^\mathsf{T} X^\mathsf{T} X v_i = 0$ for all i, since otherwise we could increase λ_i^* if $v_i^\mathsf{T} X^\mathsf{T} X v_i > 0$ (or vice versa) to improve the objective while maintaining feasibility. It follows that $\langle X^\mathsf{T} X, P^* \rangle = \langle X^\mathsf{T} X, P^{**} \rangle$. This means that P^{**} is optimal for (3); since it is also feasible for (2), we are done. Second, if $\mathrm{rank}(P^*) > d$ then $0 < \lambda_d^* < 1$ since the λ_i^* are ordered. Consider $\tilde{P} \triangleq (P^* - cP^{**})/(1-c), c = \min\{\lambda_d^*, 1 - \lambda_d^*\}$. Note that $\tilde{P}$ is feasible for (3), and that P^* is a strict convex combination of P^{**} and $\tilde{P}$. All points between $\tilde{P}$ and P^{**} are feasible by convexity, and so the optimality of P^* implies that P^{**} and $\tilde{P}$ 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). $\square$

We conclude this section with two useful results on the spectral norm $\|\cdot\|_2$ of a symmetric matrix.

Theorem 2. *Let Q be a symmetric matrix, and suppose $\varphi \geq \|Q\|_2$. Then $\|Q\|_2 = \max\{\|Q + \varphi\mathbb{I}\|_2, \| - Q + \varphi\mathbb{I}\|_2\} - \varphi$.*

Proof. First diagonalize $Q = \sum_{i=1}^p \lambda_i v_i v_i^\mathsf{T}$, with orthonormal basis v_i and (without loss of generality) λ_i in non-increasing order. Then $+Q + \varphi\mathbb{I} = \sum_{i=1}^p (+\lambda_i + \varphi)v_i v_i^\mathsf{T}$, $-Q + \varphi\mathbb{I} = \sum_{i=1}^p (-\lambda_i + \varphi)v_i v_i^\mathsf{T}$. But by construction $\lambda_i + \varphi \geq 0$ and $-\lambda_i + \varphi \geq 0$ for all $i = 1, \ldots, p$. Thus $\|Q + \varphi\mathbb{I}\|_2 = \lambda_1 + \varphi$ and $\| - Q + \varphi\mathbb{I}\|_2 = -\lambda_p + \varphi$. The result follows since $\|Q\|_2 = \max\{\lambda_1, -\lambda_p\}$. $\square$

Corollary 1. *Let Q be a symmetric matrix, and suppose $\varphi \geq \|Q\|_2$. If V is such that $V^\mathsf{T} V = \mathbb{I}$, then $\|V^\mathsf{T} Q V\|_2 = \max\{\|V^\mathsf{T}(Q + \varphi\mathbb{I})V\|_2, \|V^\mathsf{T}(-Q + \varphi\mathbb{I})V\|_2\} - \varphi$.*

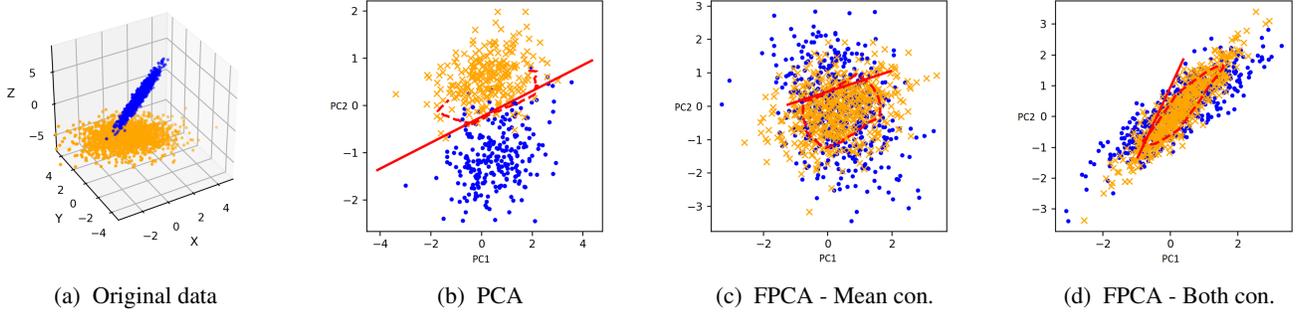| (a) Original data | (b) PCA | (c) FPCA - Mean con. | (d) FPCA - Both con. |

Figure 1: Comparison of PCA and FPCA on synthetic data. In each plot, the thick red line is the optimal linear SVM separating by color, and the dotted line is the optimal Gaussian kernel SVM.

Proof. First note that $V^\mathsf{T}(Q + \varphi\mathbb{I})V = V^\mathsf{T}QV + \varphi\mathbb{I}$ and that $V^\mathsf{T}(-Q + \varphi\mathbb{I})V = -V^\mathsf{T}QV + \varphi\mathbb{I}$. Since the spectral norm is submultiplicative, this means $\|V^\mathsf{T}QV\|_2 \le \|V^\mathsf{T}\|_2 \|Q\|_2 \|V\|_2 \le \|Q\|_2$. So $\varphi \ge \|V^\mathsf{T}QV\|_2$, and the result follows by applying Theorem 2 to $V^\mathsf{T}QV$. $\square$

Recall that using the Schur complement allows representation of $\|VRV^\mathsf{T}\|_2$ as a positive semidefinite matrix constraint when R is positive semidefinite (Boyd et al. 1994). So the above corollary is useful because it means we can represent $\|VQV^\mathsf{T}\|_2$ using positive semidefinite matrix constraints since $(Q + \varphi\mathbb{I})$ and $(-Q + \varphi\mathbb{I})$ are positive semidefinite by construction.

5 Designing formulations for fair PCA

Consider the linear dimensionality reduction $\Pi(x) = V^\mathsf{T}x$ for $V \in \mathbb{R}^{p \times d}$ such that $V^\mathsf{T}V = \mathbb{I}$. Then for linear classifier $h(u, t) = \mathbf{1}(w^\mathsf{T}u - t \le 0)$, definition (1) simplifies to $\Delta(h) = \sup_t |\mathbb{P}[w^\mathsf{T}V^\mathsf{T}x \le t|z = +1] - \mathbb{P}[w^\mathsf{T}V^\mathsf{T}x \le t|z = -1]|$. But the right-hand side is the Kolmogorov distance between $w^\mathsf{T}V^\mathsf{T}x$ conditioned on $z = \pm 1$, which is upper bounded (as can be seen trivially from its definition) by the total variation distance. Consequently, applying Pinsker's inequality (Massart 2007) gives $\Delta(h) \le \sqrt{\frac{1}{2}\mathcal{KL}(w^\mathsf{T}V^\mathsf{T}X_- \,\|\, w^\mathsf{T}V^\mathsf{T}X_+)}$, where $\mathcal{KL}(\cdot\|\cdot)$ is the Kullback-Leibler divergence, X_+ is the random variable $[x|z = +1]$, and X_- is the random variable $[x|z = -1]$. For the special case $X_+ \sim \mathcal{N}(\mu_+, \Sigma_+)$ and $X_- \sim \mathcal{N}(\mu_-, \Sigma_-)$, we have (Kullback 1997):

$$\Delta(h) \le \sqrt{\frac{1}{4}\left(\frac{s_-}{s_+} + \frac{(m_+ - m_-)^2}{s_+} + \log\frac{s_+}{s_-} - 1\right)}. \quad (4)$$

where $s_+ = w^\mathsf{T}V^\mathsf{T}\Sigma_+Vw$, $s_- = w^\mathsf{T}V^\mathsf{T}\Sigma_-Vw$, $m_+ = w^\mathsf{T}V^\mathsf{T}\mu_+$, and $m_- = w^\mathsf{T}V^\mathsf{T}\mu_-$. The key observation here is that (4) is minimized when $s_+ = s_-$ and $m_+ = m_-$, and we will use this insight to propose constraints for FPCA. If X_+ and X_- are not Gaussian, the three-point property may be used to obtain a similar bound with a couple extra terms involving the divergence between X_+ and a normal

distribution with the same mean and variance (and the analog for X_-).

We first design constraints for the non-convex formulation (2) so that $\hat{m}_+ - \hat{m}_- = w^\mathsf{T}V^\mathsf{T}f$ has small magnitude, where $f = \hat{\mu}_+ - \hat{\mu}_- = \frac{1}{\#P}\sum_{i \in P}x_i - \frac{1}{\#N}\sum_{i \in N}x_i$. Note we make the identification $P = VV^\mathsf{T}$ because of the properties of P in (2) and since $V^\mathsf{T}V = \mathbb{I}$. Observe that $w^\mathsf{T}V^\mathsf{T}f$ is small if $V^\mathsf{T}f$ is small, which can be formulated as

$$\|V^\mathsf{T}f\|^2 = \langle VV^\mathsf{T}, ff^\mathsf{T}\rangle = \langle P, ff^\mathsf{T}\rangle \le \delta^2, \quad (5)$$

where $\|\cdot\|$ is the ℓ_2-norm, and δ is a bound on the norm. This (5) is a linear constraint on P.

We next design constraints for the non-convex formulation (2) so that $\hat{s}_+ - \hat{s}_- = w^\mathsf{T}V^\mathsf{T}(\widehat{\Sigma}_+ - \widehat{\Sigma}_-)Vw$ has small magnitude. Recall the identification $P = VV^\mathsf{T}$ because of the properties of P in (2) and since $V^\mathsf{T}V = \mathbb{I}$. Next observe that $w^\mathsf{T}V^\mathsf{T}(\widehat{\Sigma}_+ - \widehat{\Sigma}_-)Vw$ is small if $V^\mathsf{T}(\widehat{\Sigma}_+ - \widehat{\Sigma}_-)V$ is small. Let $Q = \widehat{\Sigma}_+ - \widehat{\Sigma}_-$, then using Corollary 1 gives

$$\begin{aligned}
\mu + \varphi &\ge \|V^\mathsf{T}QV\|_2 + \varphi \\
&= \max\{\|V^\mathsf{T}(Q + \varphi\mathbb{I})V\|_2, \|V^\mathsf{T}(-Q + \varphi\mathbb{I})V\|_2\} \\
&= \max\{\|VV^\mathsf{T}(Q + \varphi\mathbb{I})VV^\mathsf{T}\|_2, \|VV^\mathsf{T}(-Q + \varphi\mathbb{I})VV^\mathsf{T}\|_2\} \\
&= \max\{\|P(Q + \varphi\mathbb{I})P\|_2, \|P(-Q + \varphi\mathbb{I})P\|_2\}, \quad (6)
\end{aligned}$$

where $\varphi \ge \|\widehat{\Sigma}_+ - \widehat{\Sigma}_-\|_2$, and μ is a bound on the norm. Note (6) can be rewritten as SDP constraints using a standard reformulation for the spectral norm (Boyd et al. 1994).

We design an SDP formulation for FPCA by combining the above elements. Though (2) with constraint (5) and (6) is a non-convex problem for FPCA, we showed in Theorem 1 that (3) was an exact relaxation of (2) after extracting the d largest eigenvectors of the solution of (3). Thus, we propose

666

the following SDP formulation for FPCA:

$$\max \langle X^{\mathsf{T}}X, P \rangle - \mu t \tag{7a}$$

$$\text{s.t. } \operatorname{trace}(P) \leq d, \ \mathbb{I} \succeq P \succeq 0 \tag{7b}$$

$$\langle P, ff^{\mathsf{T}} \rangle \leq \delta^2 \tag{7c}$$

$$\begin{bmatrix} t\mathbb{I} & PM_+ \\ M_+^{\mathsf{T}}P & \mathbb{I} \end{bmatrix} \succeq 0, \tag{7d}$$

$$\begin{bmatrix} t\mathbb{I} & PM_- \\ M_-^{\mathsf{T}}P & \mathbb{I} \end{bmatrix} \succeq 0 \tag{7e}$$

where $M_i M_i^{\mathsf{T}}$ is the Cholesky decomposition of $iQ + \varphi\mathbb{I}$ ($i \in \{-, +\}$), $\varphi \geq \|\widehat{\Sigma}_+ - \widehat{\Sigma}_-\|_2$, (7c) is called the *mean constraint* and denotes the use (5), and (7d) and (7e) are called the *covariance constraints* and are the SDP reformulation of (6). Our convex formulation for FPCA consists of solving (7) and then extracting the d largest eigenvectors from the optimal P^*.

We can apply the kernel trick to (7) to develop an SDP for F-KPCA. For brevity, we only note the differences with (7): Q would be replaced with $Q_k = K(X, X_+)K(X_+, X) - K(X, X_-)K(X_-, X)$ and f with $f_k = \frac{1}{\#P}K(X, X_+)\mathbf{e} - \frac{1}{\#N}K(X, X_-)\mathbf{e}$. M_+ and M_- would then be the Cholesky decompositions of the analogous matrices, and φ would also have to be set no less than $\|Q_k\|_2$. K-FPCA is then the top d eigenvectors of the optimal solution of the resulting SDP. Note that we may also replicate constraints (7c), (7d) & (7e) to handle multiple protected attributes. Note that they would need to be replicated further to account for interactions between protected attributes.

6 Experimental results

We use synthetic and real datasets from the UC Irvine Machine Learning Repository (Lichman 2013) to demonstrate the efficacy of our SDP formulations. We also show how FPCA can be used to minimize discrimination in health insurance rates (with respect to age). For any SVM run, tuning parameters were chosen using 5-fold cross-validation, and data was normalized to have unit variance in each field. Due to constraints on space, further experimental results and a overview of and comparison to the method of Calmon et al. are presented in the appendix, available at https://arxiv.org/pdf/1802.03765.pdf.

6.1 Synthetic Data

We sampled 1000 points each from X_+ and X_- distributed as different 3-dimensional multivariate Gaussians, and these points are shown in Figure 1a. Figure 1b displays the results of dimensionality reduction using the top two unconstrained principal components of X: the resulting separators for linear and Gaussian kernel SVM's are also shown. It is clear that the two sub-populations are readily distinguishable in the lower-dimensional space. Figure 1c displays the analogous information after FPCA with only the mean constraint, and Figure 1d after FPCA with both constraints. Figures 1c and 1d clearly display better mixing of the data, and the SVM's conducted afterwards are unable to separate the sub-groups as cleanly as they can in Figure 1b; furthermore, the addition
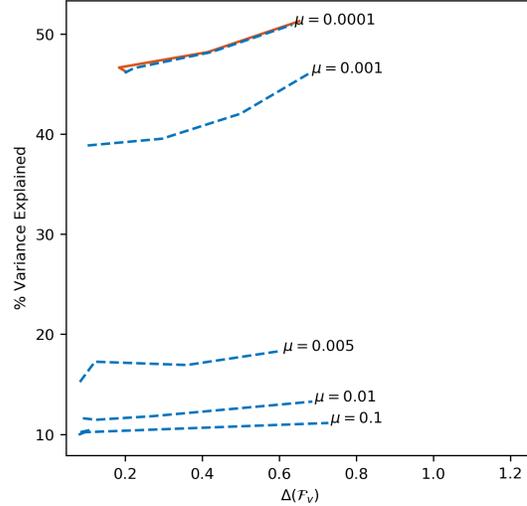

Figure 2: The sensitivity of FPCA to the δ and μ for the wine quality dataset. The full red line represents FPCA with only the mean constraint, and the dotted blue lines denote FPCA with both constraints. For each curve, $\delta \in \{0, 0.1, 0.3, 0.5\}$ was considered.

of the covariance constraints (7d) incentivizes the choosing of a dimensionality reduction that better matches the skew of the entire data set.

6.2 Real data

We next consider a selection of datasets from UC Irvine's online Machine Learning Repository (Lichman 2013). For each of the datasets, one attribute was selected as a protected class, and the remaining attributes were considered part of the feature space. After splitting each dataset into separate training (70%) and testing (30%) sets, the top five principal components were then found for the training sets of each of these datasets three times: once unconstrained, once with (7) with only the mean constraints (and excluding the covariance constraints) with $\delta = 0$, and once with (7) with both the mean and covariance constraints with $\delta = 0$ and $\mu = 0.01$; the test data was then projected onto these vectors. All data was normalized to have unit variance in each feature, which is common practice for datasets with features of incomparable units. For each instance, we estimated $\Delta(\mathcal{F})$ using the test set and for the families of linear SVM's $\mathcal{F}_v$ and Gaussian kernel SVM's $\mathcal{F}_k$. Finally, for each set of principal components V, the proportion of variance explained by the components was calculated as $\operatorname{trace}(V\widehat{\Sigma}V^{\mathsf{T}}))/\operatorname{trace}(\widehat{\Sigma})$, where $\widehat{\Sigma}$ is the centered sample covariance matrix of training set X. Table 1 displays all of these results averaged over 5 different training-testing splits.

We may observe that our additional constraints are largely helpful in ensuring fairness by all definitions. Furthermore,

Table 1: Δ-fairness for both linear and Gaussian kernel SVM for PCA and FPCA. Best results for each fairness metric are bolded.

DATA SET	UNCONSTRAINED			FPCA - MEAN CON.			FPCA - BOTH CON.		
	%VAR	LIN.	GAUS.	%VAR	LIN.	GAUS.	%VAR	LIN.	GAUS.
ADULT INCOME	11.41	0.54	0.54	9.27	0.14	0.35	5.33	**0.07**	**0.15**
BIODEG [1]	31.16	0.2	0.35	30.46	0.14	0.29	21.45	**0.10**	**0.28**
E. COLI [2]	65.01	0.65	0.80	54.31	0.46	0.59	53.75	**0.24**	**0.54**
ENERGY [3]	84.08	0.10	0.20	66.48	**0.07**	0.20	62.11	**0.07**	**0.16**
GERMAN CREDIT	11.19	0.21	0.31	10.91	0.14	0.33	8.84	**0.11**	**0.29**
IMAGE	62.68	0.18	0.32	52.78	**0.14**	0.33	48.55	0.15	**0.28**
LETTER	42.33	0.58	0.58	29.29	**0.07**	0.22	23.76	**0.07**	**0.19**
MAGIC [4]	61.91	0.32	0.33	29.57	**0.11**	**0.21**	25.36	0.12	0.30
PIMA [5]	49.00	0.30	0.37	43.98	**0.17**	0.26	43.26	0.18	**0.25**
RECIDIVISM [6]	56.28	0.24	0.26	46.58	**0.08**	**0.16**	39.34	**0.08**	0.21
SKILLCRAFT [7]	40.62	0.15	0.19	29.95	**0.07**	**0.14**	25.48	**0.07**	0.17
STATLOG	87.80	0.79	0.79	21.77	0.23	0.69	7.76	**0.13**	**0.44**
STEEL	46.05	0.64	0.71	40.79	0.19	0.51	11.86	**0.09**	**0.22**
TAIW. CREDIT [8]	45.52	0.11	0.17	30.07	0.08	0.16	20.08	**0.06**	**0.14**
WINE QUALITY [9]	50.21	0.97	0.96	37.34	0.21	0.51	10.12	**0.06**	**0.13**

in many cases, this increase in fairness comes at minimal loss in the explanatory power of the principal components. There are a few datasets for which (7d) appear superfluous. In general, gains in fairness are stronger with respect to $\mathcal{F}_v$; this is to be expected, as $\mathcal{F}_k$ is a highly sophisticated set, and thus more robust to linear projections. Kernel FPCA may be a better approach to tackling this issue, but we leave this for future work. Additional experiments and a comparison to the method of Calmon et al. are shown in the appendix. We find that our method leads to more fairness on almost all datasets.

6.3 Hyperparameter sensitivity

Next, we consider the sensitivity of our results to hyperparameters δ, μ, for the Wine Quality dataset. The data was split into training (70%) and testing (30%) sets, and the top three fair principle components were found using (7) with only the mean constraint for each candidate δ and using (7) with both constraints for all combinations of candidate δ and μ. All data was normalized to have unit variance in each independent feature. We calculated the percentage of the variance explained by the resulting principle components, and we estimated the fairness level $\Delta(\mathcal{F}_v)$ for the family of linear SVM's. This process was run 10 times for random data splits, and the averaged results are plotted in Figure 2. Here, the solid red line represents (7) with only the mean constraint. On the other hand, the dotted blue lines represent the (7) with both constraints, for the indicated μ.

Adding the covariance constraints and further tightening μ generally improves fairness and decreases the proportion of variance explained. However, observe that the relative sensitivity of fairness to δ is higher than that of the variance explained, at least for this dataset. Similarly, increasing μ decreases the portion of variance explained while resulting in a less discriminatory dataset after the dimensionality reduction. We note that increasing μ past a certain point does not provide much benefit, and so smaller values of μ are to be preferred. We found that increasing μ past 0.1 did not substantively change results further, so the largest μ that we consider is 0.1. In general, hyperparameters may be set with cross-validation, although (4) may serve as guidance.

6.4 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] (Mansouri et al. 2013) [2] (Horton and Nakai 1996) [3] (Tsanas and Xifara 2012) [4] (Bock et al. 2004) [5] (Smith et al. 1988) [6] (Angwin et al. 2016) [7] (Thompson et al. 2013) [8] (Yeh and Lien 2009) [9] (Cortez et al. 2009)
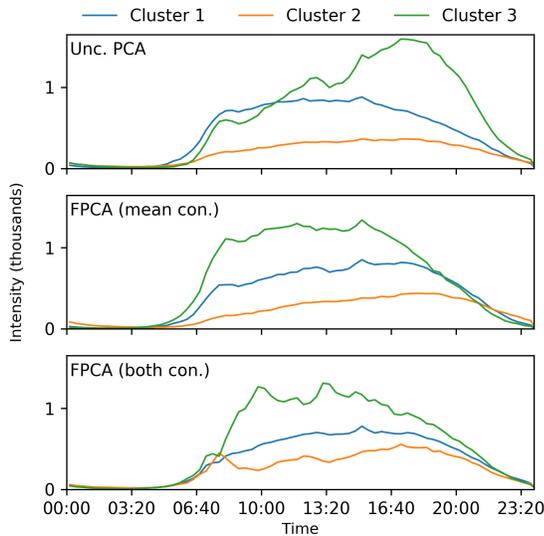
Figure 3: The mean physical activity intensities, plotted throughout a day, of the clusters generated after dimensionality reduction through PCA, FPCA with the mean constraint, and FPCA with both constraints. In each plot, each line represents the average activity level of the members of one cluster.

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 $\delta = 0$ and $\mu = 0.1$ 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.

7 Conclusion

In this paper, we proposed a quantitative definition of fairness for dimensionality reduction, developed convex SDP formulations for fair PCA, and then demonstrated its effectiveness using several datasets. Many avenues remain for future research on fair unsupervised learning. For instance, we believe that our formulations in this paper may have suitable modifications that can be used to develop deflation and regression approaches for fair PCA analogous to those for sparse PCA (d'Aspremont et al. 2007; Zou, Hastie, and Tibshirani 2006).

References

Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2016. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica, May* 23.

Arora, R.; Cotter, A.; and Srebro, N. 2013. Stochastic optimization of pca with capped msg. In *Advances in Neural Information Processing Systems*, 1815–1823.

Berk, R.; Heidari, H.; Jabbari, S.; Kearns, M.; and Roth, A. 2017. Fairness in criminal justice risk assessments: The state of the art. *arXiv preprint arXiv:1703.09207*.

Beutel, A.; Chen, J.; Zhao, Z.; and Chi, E. H. 2017. Data decisions and theoretical implications when adversarially learning fair representations. *arXiv preprint arXiv:1707.00075*.

Bock, R.; Chilingarian, A.; Gaug, M.; Hakl, F.; Hengstebeck, T.; Jiřina, M.; Klaschka, J.; Kotrč, E.; Savický, P.; Towers, S.; et al. 2004. Methods for multidimensional event classification: a case study using images from a cherenkov gamma-ray telescope. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 516(2):511–528.

Boyd, S.; El Ghaoui, L.; Feron, E.; and Balakrishnan, V. 1994. *Linear matrix inequalities in system and control theory*, volume 15. SIAM.

Calders, T.; Kamiran, F.; and Pechenizkiy, M. 2009. Building classifiers with independency constraints. In *Data mining workshops, 2009. ICDMW'09. IEEE international conference on*, 13–18. IEEE.

Calmon, F.; Wei, D.; Vinzamuri, B.; Ramamurthy, K. N.; and Varshney, K. R. 2017. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems*, 3992–4001.

Centers for Desease Control and Prevention (CDC). National Center for Health Statistics (NCHS). 2018. National health and nutrition examination survey data.

Chierichetti, F.; Kumar, R.; Lattanzi, S.; and Vassilvitskii, S. 2017. Fair clustering through fairlets. In *Advances in Neural Information Processing Systems*, 5036–5044.

Chouldechova, A. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *arXiv preprint arXiv:1703.00056*.

Cortez, P.; Cerdeira, A.; Almeida, F.; Matos, T.; and Reis, J. 2009. Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems* 47(4):547–553.

d'Aspremont, A.; El Ghaoui, L.; Jordan, M.; and Lanckriet, G. 2007. A direct formulation of sparse PCA using semidefinite programming. *SIAM Review* 49(3).

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 214–226. ACM.

Feldman, M.; Friedler, S. A.; Moeller, J.; Scheidegger, C.; and Venkatasubramanian, S. 2015. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 259–268. ACM.

Fukuoka, Y.; Zhou, M.; Vittinghoff, E.; Haskell, W.; Goldberg, K.; and Aswani, A. 2018. Objectively measured baseline physical activity patterns in women in the mped trial: Cluster analysis. *JMIR Public Health and Surveillance* 4(1):e10.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, 3315–3323.

Horton, P., and Nakai, K. 1996. A probabilistic classification system for predicting the cellular localization sites of proteins. In *Ismb*, volume 4, 109–115.

Kullback, S. 1997. *Information theory and statistics*. Courier Corporation.

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 self-tracking 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 non-asymptotic 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 (ICML-13)*, 325–333.

Zhang, B. H.; Lemoine, B.; and Mitchell, M. 2018. Mitigating unwanted biases with adversarial learning. *arXiv preprint arXiv:1801.07593*.

Zliobaite, I. 2015. On the relation between accuracy and fairness in binary classification. *arXiv preprint arXiv:1505.05723*.

Zou, H.; Hastie, T.; and Tibshirani, R. 2006. Sparse principal component analysis. *Journal of computational and graphical statistics* 15(2):265–286.