A Appendix

A.1 Proofs

Proof of Proposition 1. First, we use a property of elliptically contoured distributions [Cambanis et al., 1981, Corollary 5] to obtain

$$\begin{aligned}\mathbb{E}\left[\boldsymbol{C}^\top X \mid \boldsymbol{B}^\top X\right] &= \boldsymbol{a} + \operatorname{Cov}\left(\boldsymbol{C}^\top X, \boldsymbol{B}^\top X\right) \operatorname{Var}^{-1}\left(\boldsymbol{B}^\top X\right) \left[\boldsymbol{B}^\top X - \mathbb{E}\left(\boldsymbol{B}^\top X\right)\right] \\ &= \boldsymbol{a} + \boldsymbol{C}^\top \operatorname{Var}(X) \boldsymbol{B} \left(\boldsymbol{B}^\top \operatorname{Var}(X) \boldsymbol{B}\right)^{-1} \boldsymbol{B}^\top \left(X - \mathbb{E}X\right)\end{aligned}$$

for some constant $\boldsymbol{a}$. From condition (1) and the law of total covariance,

$$\begin{aligned}\operatorname{Cov}\left(S, \boldsymbol{C}^\top X\right) &= \operatorname{Cov}\left(f_S\left(\boldsymbol{B}^\top X, \epsilon_S\right), \boldsymbol{C}^\top X\right) \\ &= \mathbb{E}\left[\operatorname{Cov}\left(f_S\left(\boldsymbol{B}^\top X, \epsilon_S\right), \boldsymbol{C}^\top X \mid \boldsymbol{B}^\top X, \epsilon_S\right)\right] + \\ &\quad \operatorname{Cov}\left[\mathbb{E}\left(f_S\left(\boldsymbol{B}^\top X, \epsilon_S\right) \mid \boldsymbol{B}^\top X, \epsilon_S\right), \mathbb{E}\left(\boldsymbol{C}^\top X \mid \boldsymbol{B}^\top X, \epsilon_S\right)\right] \\ &= \operatorname{Cov}\left(f_S\left(\boldsymbol{B}^\top X, \epsilon_S\right), \mathbb{E}\left(\boldsymbol{C}^\top X \mid \boldsymbol{B}^\top X\right)\right) \\ &= \operatorname{Cov}\left(f_S\left(\boldsymbol{B}^\top X, \epsilon_S\right), X\right) \boldsymbol{B} \left(\boldsymbol{B}^\top \operatorname{Var}(X) \boldsymbol{B}\right)^{-1} \boldsymbol{B}^\top \operatorname{Var}(X) \boldsymbol{C}.\end{aligned}$$

Thus, we have $\operatorname{Cov}\left(S, \boldsymbol{C}^\top X\right) = \boldsymbol{0}$ if $\boldsymbol{B}^\top \operatorname{Var}(X) \boldsymbol{C} = \boldsymbol{0}$ which implies that the columns of $\boldsymbol{C}$ lie in the nullspace of $\operatorname{Var}(X) \boldsymbol{B}$. ☐

Proof of Theorem 1. We first show that the basis (8) of the classifier hypothesis RKHS is orthonormal, and then compute the canonical angles using the basis (8) and an orthonormal basis of $\mathcal{F}$. Denote by $\xi_i \coloneqq (\gamma_i - \rho_i \sigma_i) \boldsymbol{QM\Lambda}^{-1/2} \boldsymbol{U}_i + \rho_i \boldsymbol{AT\Omega}^{-1/2} \boldsymbol{V}_i$ the i-th basis function in (8), we have

$$\begin{aligned}\langle \xi_i, \xi_j \rangle &= (\gamma_i - \rho_i \sigma_i)(\gamma_j - \rho_j \sigma_j) \boldsymbol{U}_i^\top \boldsymbol{U}_j + \rho_i \rho_j \boldsymbol{V}_i^\top \boldsymbol{V}_j + 2\rho_i \sigma_i (\gamma_i - \rho_i \sigma_i) \mathbb{1}_{i=j} \\ &= \left[(\gamma_i - \rho_i \sigma_i)^2 + \rho_i^2 + 2\rho_i \sigma_i (\gamma_i - \rho_i \sigma_i)\right] \mathbb{1}_{i=j} \\ &= \mathbb{1}_{i=j},\end{aligned}$$

where the first equality follows from the orthonormal basis (7). This shows that (8) is an orthonormal basis. Using the orthonormal basis $\left\{\psi_i \coloneqq \boldsymbol{\phi QM\Lambda}_i^{-1/2}\right\}_{i=1}^r$ of $\mathcal{F}$, we can use the SVD to compute the canonical angles (see e.g., Algorithm 6.4.3 in [Golub and Van Loan, 2013]) as

$$[\xi_1, \cdots, \xi_d]^\top [\psi_1, \cdots, \psi_r] = \operatorname{diag}(\gamma_i - \rho_i \sigma_i) \boldsymbol{U}^\top + \operatorname{diag}(\rho_i) \boldsymbol{\Sigma U}^\top = \boldsymbol{I}_d \operatorname{diag}(\gamma_i) \boldsymbol{U}^\top. \tag{1}$$

Here, $\operatorname{diag}(d_i)$ denotes the diagonal matrix with diagonal elements d_i. Note that the last term in (1) is the (thin) SVD, and the singular values γ_i are the canonical angles between $\mathcal{M}$ and $\mathcal{F}$. Finally, we relate the canonical angles to the operator norm in (9). Recall that the orthogonal projector can be expressed as the tensor product $\mathcal{P}_\mathcal{F} = \sum_{i=1}^r \psi_i \otimes \psi_i$, and $\mathcal{P}_\mathcal{F} h = \sum_{i=1}^r \langle h, \psi_i \rangle \psi_i$. We have

$$\begin{aligned}\|\mathcal{P}_\mathcal{F} - \mathcal{P}_\mathcal{M}\| &= \|(\mathcal{P}_\mathcal{F} + \mathcal{P}_\mathcal{F})(\mathcal{P}_\mathcal{F} - \mathcal{P}_\mathcal{M})(\mathcal{P}_\mathcal{M} + \mathcal{P}_\mathcal{M})\| \\ &= \|\mathcal{P}_\mathcal{F}\mathcal{P}_\mathcal{M} - \mathcal{P}_\mathcal{F}\mathcal{P}_\mathcal{M}\| \\ &= \sup_{h \in \mathcal{H}_{\kappa,n} \colon \|h\| \le 1} \left(\|\mathcal{P}_\mathcal{F}\mathcal{P}_\mathcal{M} h\| + \|\mathcal{P}_\mathcal{F}\mathcal{P}_\mathcal{M} h\|\right),\end{aligned}$$

where $\mathcal{F}$ and $\mathcal{M}$ represent respectively the orthogonal complements of $\mathcal{F}$ and $\mathcal{M}$. It can be shown that $\mathcal{P}_\mathcal{F}\mathcal{P}_\mathcal{M}$ and $\mathcal{P}_\mathcal{F}\mathcal{P}_\mathcal{M}$ have the same nonzero singular values which are the sines of the principal angles between $\mathcal{F}$ and $\mathcal{M}$ (see e.g., p.249 of Stewart, 2001). From (1), these principal angles are $\arccos(\gamma_i)$. Thus, we obtain $\|\mathcal{P}_\mathcal{F} - \mathcal{P}_\mathcal{M}\| = \sqrt{1 - \min_i \gamma_i^2}$. To obtain (10), one can simply apply the trigonometric identity of sines yielding

$$\|\mathcal{P}_\mathcal{G} - \mathcal{P}_\mathcal{M}\| = \gamma_{\min}\sqrt{1 - \sigma_{\min}^2} - \sigma_{\min}\sqrt{1 - \gamma_{\min}^2} = \max\left\{0, \epsilon\sqrt{1 - \sigma_{\min}^2} - \sigma_{\min}\sqrt{1 - \epsilon^2}\right\},$$

where we denote by $\gamma_{\min} \coloneqq \min_i \gamma_i$. ☐

A.2 Implementation

Algorithm 1 gives the Matlab-style pseudo-code for our approach which can handle multiple protected attributes. This algorithm use the SDR procedure described in Algorithm 2 to compute the desired model representation with a specified trade-off ϵ.

Algorithm 1: $\boldsymbol{E} = \texttt{MBasis}\left(\boldsymbol{K}, \boldsymbol{y}, \boldsymbol{S}, m, d, \epsilon\right)$ — Compute the basis $\boldsymbol{\phi E}$ for $\mathcal{M}$

---

[1] Initialize $\boldsymbol{W} = [\,]$, n with the number of rows of $\boldsymbol{K}$, as well as indices $\texttt{pos} = (\boldsymbol{y} == 1)$ and $\texttt{neg} = (\boldsymbol{y} \neq 1)$.

[2] **foreach** *column* $\boldsymbol{s}$ *of* $\boldsymbol{S}$ **do**

[3]   **if** *EqualizedOdds* **or** *EqualityOfOpportunity* **then**

[4]   Set $\boldsymbol{B} = \boldsymbol{0}_{n\times m}$ and update $\boldsymbol{B}\left(\texttt{pos}\,:\right) = \texttt{SDR}\left(\boldsymbol{K}\left(\texttt{pos}, \texttt{pos}\right), \boldsymbol{s}, m\right)$.

[5]   Append basis $\boldsymbol{B}$ to $\boldsymbol{W}$: $\boldsymbol{W} = [\boldsymbol{W}\ \ \boldsymbol{B}]$.

[6]   **if** *EqualizedOdds* **then**

[7]   Set $\boldsymbol{B} = \boldsymbol{0}_{n\times m}$, then $\boldsymbol{B}\left(\texttt{neg}, :\right) = \texttt{SDR}\left(\boldsymbol{K}\left(\texttt{neg}, \texttt{neg}\right), \boldsymbol{s}, m\right)$.

[8]   Let $\boldsymbol{W} = [\boldsymbol{W}\ \ \boldsymbol{B}]$.

        end

    else

[9]   Compute $\boldsymbol{B} = \texttt{SDR}\left(\boldsymbol{K}, \boldsymbol{s}, m\right)$, and update $\boldsymbol{W} = [\boldsymbol{W}\ \ \boldsymbol{B}]$.

    end

end

[10] **Predictive Subspace:** Compute the predictive subspace as the SDR subspace $\boldsymbol{A} = \texttt{SDR}\left(\boldsymbol{K}, \boldsymbol{y}, d\right)$.

[11] **Fair Subspace:** Obtain $\boldsymbol{K}'$ by subtracting the mean of each column of $\boldsymbol{K}$. Let $\tilde{\boldsymbol{K}} = \boldsymbol{K}'^{\top}\boldsymbol{K}'$, and use QR decomposition to compute $\boldsymbol{Q}$ as the nullspace basis of the column space of $\tilde{\boldsymbol{K}}\boldsymbol{W}$.

[12] Perform the eigenvalue decompositions to obtain Equation (7), and then use Theorem 1 to compute $\boldsymbol{E}$.

Algorithm 2: $\boldsymbol{W} = \texttt{SDR}\left(\boldsymbol{K}, \boldsymbol{s}, m\right)$ — Compute the SDR subspace $\boldsymbol{\phi W}$

---

[1] Sort $\boldsymbol{s}$ such that $s\left(\texttt{idx}\right)$ is non-decreasing. Let $\texttt{invIdx}$ be the inverse of $\texttt{idx}$ satisfying $\texttt{idx}\left(\texttt{invIdx}\right) = 1 : n$, where n is the number of rows of $\boldsymbol{K}$.

[2] Slice $\boldsymbol{s}$ approximately evenly as described in § 3.1 such that entries with the same value are in the same partition. Denote by n_i the size of partition i.

[3] Initialize $\eta = 10^{-4}$, i.e., a small constant. Let $\boldsymbol{K}' \coloneqq \boldsymbol{K}\left(\texttt{idx}, \texttt{idx}\right)$, and solve $\boldsymbol{\Gamma}_n \boldsymbol{K}' \boldsymbol{A}_i = \tau_i \left[\operatorname{diag}\left(\boldsymbol{\Gamma}_{n_i}\right)\boldsymbol{K}' + n\eta \boldsymbol{I}_n\right]\boldsymbol{A}_i$ for $\boldsymbol{A}$.

[4] Return $\boldsymbol{W} = \boldsymbol{A}\left(\texttt{invIdx}, :\right)$.

References

Stamatis Cambanis, Steel Huang, and Gordon Simons. On the theory of elliptically contoured distributions. *Journal of Multivariate Analysis*, 11(3):368–385, 1981.

G.H. Golub and C.F. Van Loan. *Matrix Computations*. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, 2013. ISBN 9781421407944.

G.W. Stewart. *Matrix Algorithms Volume 2: Eigensystems*. Other Titles in Applied Mathematics. Society for Industrial and Applied Mathematics, 2001. ISBN 9780898715033.