Supplements for Contrastive Functional Principal Components Analysis

Eric Zhang, Didong Li

Department of Biostatistics, University of North Carolina at Chapel Hill
eyzhang@unc.edu, didongli@unc.edu

Licensed assets

1. Gait Cycle data - CC BY-4.0.
2. Stock data - Apache Software License.
3. Python- scikit-fda - BSD 3-Clause License.
4. This work is licensed by CC BY-4.0.

Code availability

All codes and datasets can be found and downloaded at https://github.com/ezhang1218/CFPCA. All experiments were conducted on a laptop equipped with a 12th Generation Intel Core i7-1255U processor running at 1.70 GHz, 16 GB of installed RAM, and an Intel Iris Xe Graphics card with 128 MB of dedicated video memory.

Alternative CFPCA method

The following method is another implementation of CFPCA that follows closely the implementation of FPCA in python, and is available in the codes.

Proofs

Proof of Theorem 1

Proof. From Lemma 4.3 in Bosq (2000) and the use of the triangle inequality,

$$\begin{aligned}
&\mathbb{E}\|\widehat{p}_j^{N,M}\widehat{v}_j^{N,M} - v_j\| \\
&\leq \frac{2\sqrt{2}}{\delta_j}\cdot\mathbb{E}\|C^{N,M} - C\|_{\mathcal{L}} \\
&\leq \frac{2\sqrt{2}}{\delta_j}\cdot(\mathbb{E}(\|C_x^N - C_x\|_{\mathcal{L}}) + \alpha\mathbb{E}(\|C_y^N - C_y\|_{\mathcal{L}})) \\
&\leq \frac{2\sqrt{2}}{\delta_j}\cdot(\mathbb{E}(\|C_x^N - C_x\|_{\mathcal{S}}) + \alpha\mathbb{E}(\|C_y^N - C_y\|_{\mathcal{S}}) \\
&= O(N^{-\frac{1}{2}}) + O(M^{-\frac{1}{2}}),
\end{aligned}$$

where $\delta_1 = \lambda_1 - \lambda_2$ and $\delta_j = \min(\lambda_{j-1} - \lambda_j, \lambda_j - \lambda_{j+1})$, $2 \leq j \leq L$. The third inequality follows from the fact $\|\Psi\|_{\mathcal{L}} \leq \|\Psi\|_{\mathcal{S}}$ for operator $\Psi \in \mathcal{L}$. The last equality follows from Theorem 2.1 in Hörmann and Kokoszka (2012).

Algorithm 1: CFPCA 2

1: **Input** target and background data: $\{x_i(t_k)\}_{i=1}^N$, $\{y_j(t_k)\}_{j=1}^M$, $k = 1, \cdots, n$; contrast parameter α, number of components L, time points $t_1, \ldots, t_n$ in interval $\mathcal{T}$.
2: Center the data $\{x_i(t)\}_{i=1}^N$, $\{y_j(t)\}_{j=1}^M$, define $X \in \mathbb{R}^{N\times n}$ and $Y \in \mathbb{R}^{M\times n}$ where each row is an observation and the columns are the discretized time points, and let $I_n \in \mathbb{R}^{n\times n}$ be the identity matrix.
3: **for** $i = 1$ to n **do**
4: $\quad W[i] \leftarrow \text{SimpsonIntegral}(I_n[i,:], \mathcal{T})$
5: **end for**
6: Define $D \in \mathbb{R}^{n\times n}$ such that $(D)_{(i,i)} = W[i], (D)_{(i,j)} = 0$ for $i \neq j$.
7: Matrix multiply: $X_{\text{new}} = X \cdot D$, $Y_{\text{new}} = Y \cdot D$
8: Perform Cholesky decomposition: $D = LL^T$.
9: Solve for B_1, B_2 through:

$$L^T B_1 = X_{\text{new}}^T$$

$$L^T B_2 = Y_{\text{new}}^T$$

10: Center B_1^T, B_2^T, and compute $V_x = \frac{1}{N}B_1^T B_1$, $V_y = \frac{1}{M}B_2^T B_2$
11: Find eigenvectors $(u_1, \ldots, u_L)$ and eigenvalues $(\widehat{\lambda}_1, \ldots, \widehat{\lambda}_L)$ of

$$V = V_x - \alpha V_y$$

12: Solve for $\widehat{v}_l$

$$L\widehat{v}_l = u_l, \quad l = 1, \ldots, L$$

13: Return $(\widehat{v}_1, \ldots, \widehat{v}_L)$ and $(\widehat{\lambda}_1, \ldots, \widehat{\lambda}_L)$.

For the second claim, from Lemma 4.2 in Bosq (2000), we have

$$\mathbb{E}\Big[\sup_{1\le j\le L}|\widehat{\lambda}_j-\lambda_j|\Big]\le \mathbb{E}\|C^{N,M}-C\|_{\mathcal{L}} = O(N^{-\frac{1}{2}})+O(M^{-\frac{1}{2}}),$$

where the equality follows from the first half of the proof. □

Proof of Theorem 2

Proof. Under the model assumption in Equations (5), we can rewrite our data as

$$x(t)=\mathbf{W(t)A}+\mathbf{S(t)H}+\epsilon(t),$$
$$y(t)=\mathbf{S(t)\Gamma}+\epsilon(t),$$

where

$$\mathbf{W}(t)=(w_1(t),\dots,w_L(t))^T,\quad \mathbf{S}(t)=(s_1(t),\dots,s_K(t))^T,$$
$$\mathbf{A}=(a_{i1},\dots,a_{iL}),\quad \mathbf{\Gamma}=(\gamma_{j1},\dots,\gamma_{jK}),$$
$$\mathbf{H}=(\eta_{j1},\dots,\eta_{jK}).$$

Recall

$$\mathbf{A}\sim\mathcal{N}(0,\mathbf{I}_L),\quad \mathbf{\Gamma}\sim\mathcal{N}(0,\mathbf{I}_K),\quad \mathbf{H}\sim\mathcal{N}(0,\mathbf{I}_K).$$

Then it follows that

$$\begin{aligned}&\text{Cov}(x(t),x(t'))\\ &=\text{Cov}(\mathbf{W(t)A}+\mathbf{S(t)\Gamma}+\epsilon(t),\mathbf{W(t')A}+\mathbf{S(t')\Gamma}+\epsilon(t')).\end{aligned}$$

Using the independence of $\mathbf{A}$, $\mathbf{\Gamma}$, and $\epsilon_i(t)$, and the linearity of covariance, we have:

$$=\text{W(t) W(t')}^T+\mathbf{S(t)S(t')}^T+\delta(t-t')f(t).$$

Here, $\delta(t-t')$ is the Kronecker delta, which equals 1 if $t=t'$ and 0 otherwise. Similarly,

$$\text{Cov}(y(t),y(t'))=\mathbf{S(t)S(t')}^T+\delta(t-t')f(t)$$

which implies

$$\begin{aligned}c_x(t,t')-c_y(t,t')&=\text{Cov}(x(t),x(t'))-\text{Cov}(y(t),y(t'))\\ &=\mathbf{W(t)W(t')}^T.\end{aligned}$$

This holds for any $t,t'>0$. Therefore w_j is the jth eigenfunction of $C=C_x-C_y$. Applying Theorem 1, the result follows. □

Proof of Theorem 3

Proof. For clarity, we omit the sign determinant term. By the triangle inequality,

$$\begin{aligned}&\mathbb{E}\left[\left\|\hat{v}_j^{N,M,n}-v_j\right\|\right]\\ &\le\mathbb{E}\left[\left\|v_j-\hat{v}_j^{N,M}\right\|\right]+\mathbb{E}\left[\left\|\hat{v}_j^{N,M}-\hat{v}_j^{N,M,n}\right\|\right].\end{aligned}$$

For the first term, we observe that

$$\begin{aligned}&\left(\frac{T}{n}\sum_{q=1}^{n}\left(v_j(t_q)-\hat{v}_j^{N,M}(t_q)\right)^2\right)^{1/2}\\ &=\left(\int\left|v_j(t)-\hat{v}_j^{N,M}(t)\right|^2dt\right)^{1/2}+O(n^{-1}),\end{aligned}$$

which is a well known result from Riemann sum approximation (Burden and Faires 2011). This implies that

$$\begin{aligned}\mathbb{E}\left[\left\|v-\hat{v}_j^{N,M}\right\|\right]&\le\mathbb{E}\left[\left\|v_j-\hat{v}_j^{N,M}\right\|\right]+O(n^{-1})\\ &=O(N^{-\frac{1}{2}})+O(M^{-\frac{1}{2}})+O(n^{-1}),\end{aligned}$$

where the last step follows from Theorem 1. Similarly, for the second term, combining Lemma 4.3 from Bosq (2000), Corollary 3 from Yu, Wang, and Samworth (2015), and Riemann sum approximation, we can approximate the empirical eigenfunction with a discrete eigenvector of the discretized covariance matrix,

$$\mathbb{E}\left[\left\|\hat{v}_j^{N,M}-\hat{v}_j^{N,M,n}\right\|\right]\le O(n^{-\frac{1}{2}}).$$

Combining both terms, we get the desired result. For the second part of the Theorem,

$$\begin{aligned}&\mathbb{E}\left[\left\|\hat{v}_j^{N,M,n}-w_j\right\|\right]\\ &\le\mathbb{E}\left[\left\|\hat{v}_j^{N,M,n}-\hat{v}_j^{N,M}\right\|\right]+\mathbb{E}\left[\left\|\hat{v}_j^{N,M}-w_j\right\|\right]\\ &=O(n^{-\frac{1}{2}})+O(N^{-\frac{1}{2}})+O(M^{-\frac{1}{2}}).\end{aligned}$$

where the first term follows from before and the second follows from Theorem 2.

□

Additional simulation details

For each function, 100 time points were generated from an interval $[-1,1]$. The background dataset was generated from a linear combination of the shared functions with $\mathcal{N}(0,1)$ coefficients. The foreground dataset was generated from a linear combination of the shared functions plus a linear combination of the foreground-specific functions, also with $\mathcal{N}(0,1)$ coefficients. The error function used for both groups was $\epsilon(t)=\mathcal{N}(0,1)\cdot\sin(2\pi t)$ for $t\in[-1,1]$.

The following plots demonstrate the convergence of the CFPCs from Simulations 1,3, and 4. The best fit line and its slope are also reported here.

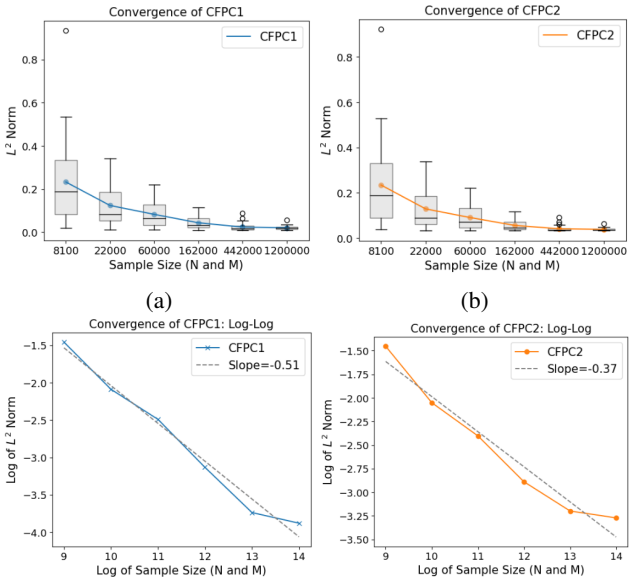


Figure 1: Convergence plots from simulation 1.

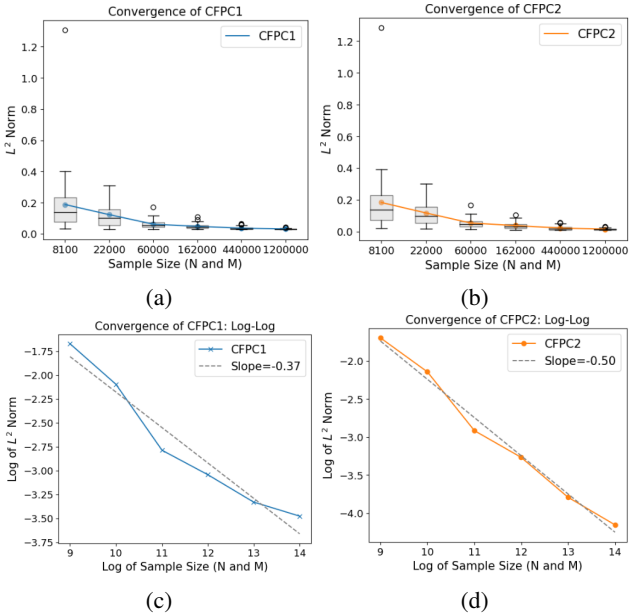


Figure 2: Convergence plots from simulation 3.

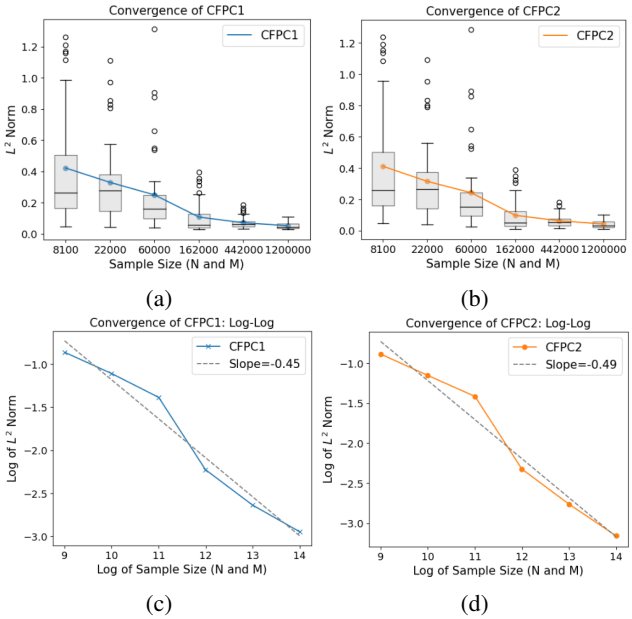


Figure 3: Convergence plots from simulation 4.

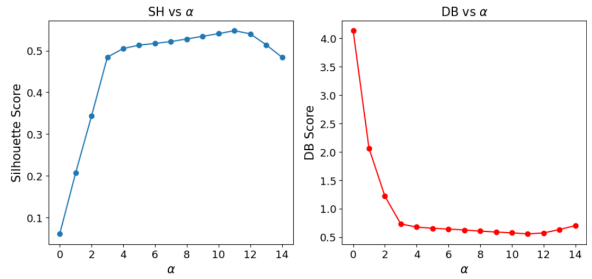
Additional application details

The purpose of this section is to first present the results from the hyperparameter selection of α and second, present events that have occurred that explain the differences in subgroups and ultimately how CFPCA is able to uncover them for Gait Cycle and Stock Market data.

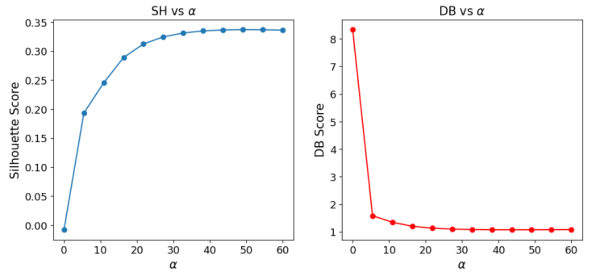
Hyperparameter tuning of α

Fig. 4 shows the tuning process for α in both applications in Application. In each process, we selected an α not only for its high SS and low DB, but also for its interpretability. Specifically, the CFPCs displayed in Fig, 4f. and Fig. 5f. were chosen because they illustrate the variations in trends among subgroups. This enhanced understanding is crucial for identifying and analyzing the distinctive characteristics of each subgroup. In general for α, we recommend setting $\alpha = 1$ under the model setup, as this directly recovers the foreground-specific functions, as demonstrated in our simulations. α controls how much background information is subtracted from the foreground: a higher α removes more shared information, which is useful when substantial overlap exists between groups. If users believe significant shared information exists, we suggest increasing α accordingly.

For selecting L, we generally suggest choosing L as 1 or 2 for straightforward visualization of high-dimensional data. Additionally, standard PCA techniques use scree plots to identify when eigenvalues plateau. Examining the fraction of variance explained by the CFPCs also aids in choosing an optimal L; we recommend selecting L where further components add only marginally to the explained variance.



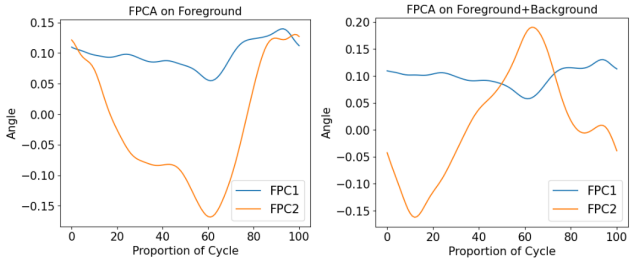
(a) Gait Cycle. $\alpha = 5$ was selected in Fig. 4e.



(b) Stock Market. $\alpha = 30$ was selected in Fig. 5e.

Figure 4: Cross-validation results of α

Additional gait cycle results



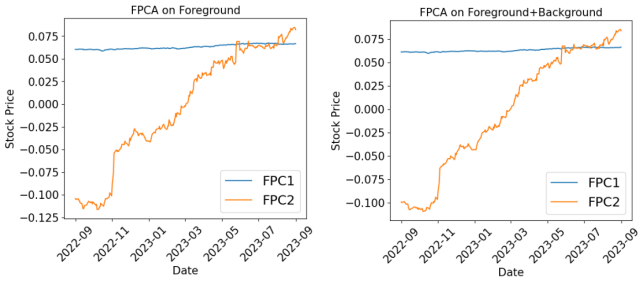
(a) (b)

Figure 5: (a) First two FPCs on braced subjects, corresponding to Fig. 4a (b) First two FPCs on braced + unbraced subjects, corresponding to Fig. 4c.

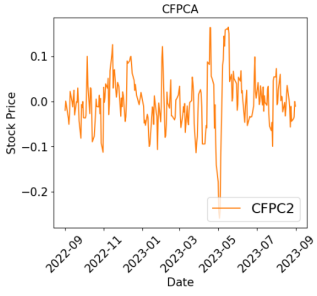
In both plots, the first FPC is difficult to interpret exactly, while the second FPC shows resemblance to the general angle displacement trend. However, neither plot is able to provide an adequate interpretation for the difference in subgroups within the foreground. This limitation indicates a need for further analysis to discern the underlying differences within these subgroups.

Additional stock market results

For the stock market dataset, a logarithmic transformation was applied to stabilize the variance (Lütkepohl and Xu 2012).



(a) (b)



(c)

Figure 6: (a) First two FPCs on tech companies, corresponding to Fig. 5a. (b) First two FPCs on tech + nontech companies, corresponding to Fig. 5c (c) Second CFPC, corresponding to Fig. 5e.

In both plots, the first FPC is represented as a straight line, which doesn't yield any valuable insights. The second FPC, however, bears a resemblance to the overall trend observed in stock prices, suggesting it captures more relevant variability. Despite this, neither plot successfully facilitates an interpretation regarding the variations among subgroups in the foreground.

As previously discussed in Application, the hardware and semiconductor industries have shown robust growth in stock, driven by products innovations and strategic expansions. Here, we provide a more structured discussion of these developments, contrasting them with trends in the cloud computing sector.

The **hardware and semiconductor** sectors have been characterized by a flurry of new product launches and facility expansions, contributing to notable stock performance improvements:

- **Product Innovations:**
  - **Dell** released new gaming laptops, monitors, and desktops (Jain 2023).
  - **Microchip** developed new ethernet devices for industrial automation (Microchip 2022).
  - **Corning** unveiled a new smartphone glass design, used by Samsung (Corning 2023).
  - **Arista** expanded its network hardware portfolio (Arista 2022).
  - **KLA** announced a new imaging system for electronic manufacturing (KLA 2022).
  - **HP** introduced products for hybrid work and gaming (HP 2022).
  - **Advanced Micro Devices** released a new series of graphics cards (Advanced Micro Devices 2022).
  - **Onsemi** launched a new series of MOSFET devices for automotive applications (Onsemi 2022).
- **Capacity Expansions:**
  - **Texas Instruments** started production in its new 300mm wafer fab (Texas Instruments 2022).
  - **Micron** initiated mass production of advanced RAM chips (Reuters 2022).
  - **STMicroelectronics** introduced new silicon carbide power modules (Randall 2022).
  - **ASML** announced capacity adjustments and increases for chip-making equipment (ASML 2022).
  - **Skyworks Solutions** launched a new network synchronization solutions portfolio (Business Wire 2022).
- **Strategic Acquisitions and Partnerships:**
  - **Lam Research Corp.** acquired SEMSYSCO for advanced chip packaging (Lam Research 2022).
- **Market Expansion:**
  - **Qualcomm** increased its automotive design-win pipeline to $30 billion, reflecting growth in automotive business (Zachs Equity Research 2022).
  - **NXP Semiconductors**, **Keysight**, and **Fortive Corporation** experienced high demand and launched new products (NXP Semiconductors 2023; Keysight Technologies 2022; Fortive Corporation 2022).
  - **Lattice Semiconductor** saw growth in its industrial and automotive segments (Lattice Semiconductor 2023).

Conversely, companies centered around **cloud computing** experienced a downturn, with several facing challenges that impacted stock values:

- Growth slowdowns and Lack of demand:
  - **Microsoft** announced a lackluster forecast in sales growth in its cloud-computing business (Bass 2022).
  - **Amazon** also faced a slowdown due to concerns related to their growth in its cloud business (Eswaran, Soumya 2023).
  - **Twilio** laid off many of its employees in its cloud services due to lack of demand (Reuters 2022).
- Partner-related impacts:
  - **Snowflake** significantly fell also due to its partner in Amazon reporting slow growth (Quast 2022).

References

Advanced Micro Devices, I. 2022. AMD Unveils World's Most Advanced Gaming Graphics Cards, Built on Groundbreaking AMD RDNA 3 Architecture with Chiplet Design. AMD.

Arista. 2022. Arista Delivers Next Generation Switching for Compute and Storage.

ASML. 2022. ASML provides updated view on demand outlook, capacity plans and business model at Investor Day meeting. ASML.

Bass, D. 2022. Microsoft Plunges on Forecast for Lackluster Azure Growth.

Bosq, D. 2000. *Linear processes in function spaces: theory and applications*, volume 149. Springer Science & Business Media.

Burden, R. L.; and Faires, J. D. 2011. *Numerical Analysis*. Brooks/Cole, 9th edition.

Business Wire. 2022. Skyworks' Advanced Synchronization Solutions Support Next-Generation 5G Deployments. Business Wire.

Corning. 2023. Corning Reports Fourth-Quarter and Full-Year 2022 Financial Results.

Eswaran, Soumya. 2023. Why Amazon.com (AMZN) Declined. Yahoo Finance.

Fortive Corporation. 2022. Fortive Reports Strong Third Quarter 2022 Results; Narrows and Raises Full Year 2022 Outlook. Fortive Investor Relations.

Hörmann, S.; and Kokoszka, P. 2012. Functional time series. In *Handbook of statistics*, volume 30, 157–186. Elsevier.

HP. 2022. HP at CES 2023: Leading in Hybrid and Gaming Experiences. HP.

Jain, S. 2023. Dell unveils new gaming laptops, monitors, Aurora desktop, and more ahead of CES 2023.

Keysight Technologies. 2022. Keysight Technologies Reports Fourth Quarter and Fiscal Year 2022 Results. Keysight Technologies.

KLA. 2022. KLA Launches New Double-Sided Direct Imaging Platform Supporting Continued Innovation for PCBs and IC Substrates.

Lam Research. 2022. Lam Research Acquires SEMSYSCO to Advance Chip Packaging. Lam Research.

Lattice Semiconductor. 2023. Lattice Semiconductor Reports Fourth Quarter and Full Year 2022 Results.

Lütkepohl, H.; and Xu, F. 2012. The role of the log transformation in forecasting economic variables. *Empirical Economics*, 42: 619–638.

Microchip. 2022. New Industrial Gigabit Ethernet Transceivers Offer Precision Timing Protocol to Optimize Process Automation Functionality.

NXP Semiconductors. 2023. NXP Semiconductors Reports First Quarter 2023 Results. NXP Semiconductors - Newsroom.

Onsemi. 2022. onsemi Launches MOSFETs With Innovative Top-Cool Packaging. Onsemi.

Quast, J. 2022. Why Snowflake Stock Fell 6% in October.

Randall, C. 2022. STMicroelectronics presents new SiC power modules. electrive.

Reuters. 2022. Cloud Communications Platform Twilio to Cut Staff by 11%. Reuters.

Reuters. 2022. Micron Begins Mass Production of Advanced DRAM Memory Chip in Japan. Gadgets 360.

Texas Instruments. 2022. TI's new 300-millimeter wafer fab in Richardson, Texas, begins initial production. Texas Instruments.

Yu, Y.; Wang, T.; and Samworth, R. J. 2015. A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika*, 102(2): 315–323.

Zachs Equity Research. 2022. Qualcomm Automotive Design-Win Pipeline Swells to 30B. Yahoo Finance.