MARMARA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

SKALER-FONKSİYON LOJİSTİK REGRESYON MODELİNİN TAHMİNİ İÇİN DİRENÇLİ BİR KESTİRİM YÖNTEMİ

BERKAY AKTÜRK

YÜKSEK LİSANS TEZİ
İstatistik Anabilim Dalı

DANIŞMAN
Doç. Dr. Ufuk BEYAZTAŞ

ISTANBUL, 2023

MARMARA UNIVERSITY
INSTITUTE FOR GRADUATE STUDIES
IN PURE AND APPLIED SCIENCES

A ROBUST ESTIMATION METHOD FOR SCALAR-ON-FUNCTION LOGISTIC REGRESSION MODEL

BERKAY AKTÜRK

MASTER THESIS
Department of Statistics

Thesis Supervisor
Doç. Dr. Ufuk BEYAZTAŞ

ISTANBUL, 2023

TEŞEKKÜR

Bu süreç boyunca bana destek olan tüm insanlara olan minnettarlığımı ifade etmek benim için büyük bir onurdur. Bu tezde bana yardımcı olan ve kişisel gelişimime katkıda bulunan insanlarla tanışmış olmaktan dolayı kendimi şanslı hissediyorum.

Öncelikle, danışmanım Doç. Dr. Ufuk BEYAZTAŞ'a, yüksek lisans eğitimim süresince desteklerini esirgemeyen Prof. Dr. Müjgan Tez, Prof. Dr. Birsen EYGİ ERDOĞAN, Prof. Dr. Deniz İNAN, Prof. Dr. Ufuk YOLCU ve Doç. Dr. Özge CAĞCAĞ YOLCU'ya en derin minnettarlığımı sunmak istiyorum. Bu bilge ve deneyimli akademisyenlerle çalışmak büyük bir onur ve harika bir fırsattı. Onlar sadece ilham verici araştırmacılar değil, aynı zamanda mükemmel öğretmenlerdir ve istatistik alanındaki geniş bilgi ve deneyimleri sayesinde çok şey öğrenme fırsatı buldum. Ayrıca onların bilgelikleri ve canlılıklarından da büyük ölçüde yararlandım. Öğrencilik dönemimde, derslerini ve anlamlı tartışmalarını dört gözle bekledim. Onların öğrencisi olmak büyük bir zevk ve fayda sağladı.

Marmara Üniversitesi'ndeki yıllarım boyunca akademik bakış açımı genişletmek ve kişisel gelişimime katkıda bulunmak için çaba gösteren Doç. Dr. Yakup Arı'ya minnettarım. Hayatım boyunca sevgi ve desteklerini esirgemeyen, varlıklarıyla güç veren ve her zaman yanımda olan değerli annem Ayşe Ünal AKTÜRK, babam Mehmet Şükrü AKTÜRK, kardeşlerim Burçak AKTÜRK ve Bora AKTÜRK'e en içten teşekkürlerimi sunarım.

Bu tez Türkiye Bilimsel ve Teknolojik Araştırma Kurumu (TÜBİTAK) tarafından desteklenmiştir (Burs No: 120F270).

Haziran 2023 **Berkay AKTÜRK**

İÇİNDEKİLER

TEŞEKKÜR ..... **i**

İÇİNDEKİLER ..... **ii**

ÖZET ..... **iv**

ABSTRACT ..... **v**

KISALTMALAR ..... **vi**

ŞEKİL ..... **vii**

TABLO LİSTESİ ..... **viii**

1. GİRİŞ **1**

2. FONKSİYONEL VERİ ANALİZİ **4**

2.1. Hilbert Uzayı **5**

2.3. B-Spline Temel Fonksiyonları **6**

2.4. Fonksiyonel Temel Bileşen Analizi **7**

3. FONKSİYONEL LOJİSTİK REGRESYON **9**

3.1. FLogitR Modeli İçin Temel Genişletme Fonksiyonları **10**

3.2. FLogitR Modeli İçin FPCA Yaklaşımı **11**

4. FlogitR MODELİ İÇİN ÖNERİLEN YÖNTEM **13**

4.1. RFPCA Yöntemi **14**

4.2. Flogitr Modeli İçin Bianco ve Yohai Tahmincisi **15**

5. UYGULAMA **17**

5.1. Monte-Carlo Denemeleri **18**

5.1.1. Süreç 1 ..... **18**
5.1.2. Süreç 2 ..... **20**
5.2. Bulgular **22**
5.3. Gerçek Veri Analizi: El Radyografisi Verisi **28**
6. TARTIŞMA VE SONUÇ **31**
7. KAYNAKLAR **33**
ÖZGEÇMİŞ **36**

ÖZET

SKALER-FONKSİYON LOJİSTİK REGRESYON MODELİNİN TAHMİNİ İÇİN DİRENÇLİ BİR KESTİRİM YÖNTEMİ

Fonksiyonel lojistik regresyon modeli, yanıtın 0 ve 1 gibi ikili değerler aldığı ve açıklayıcı değişkenin rastgele eğrilerden oluştuğu durumlarda, ikili yanıt ve fonksiyonel açıklayıcı değişken arasındaki doğrusal ilişkiyi modellemek için sıkça kullanılır. Bu modelde parametre tahmini için kullanılan yöntemlerin çoğu, aykırı değerlere karşı hassastır. Aykırı değerlerin varlığı, yanlış parametre tahminleri ve artan hatalı sınıflandırma olasılığına yol açabilir. Bu tezde, fonksiyonel lojistik regresyon modelinde parametre fonksiyonlarının aykırı değerlere karşı dirençli bir şekilde tahminlenmesi için yeni bir yaklaşım önerilmektedir. Önerilen yöntemde, fonksiyonel açıklayıcı değişken elemanları dirençli bir fonksiyonel temel bileşenler analizi ile sonlu boyutlu uzaya yakınsanır; bu sayede, aykırı değerlerin etkisi azaltılır. Böylece, model üzerinde olumsuz etki yaratabilecek aykırı değerlerin etkileri en aza indirilir. Sonrasında, ikili yanıt ve dirençli temel bileşen skorlarıyla kurulan lojistik regresyon modelinin parametreleri, M-tipi bir tahmin yöntemi kullanılarak tahmin edilir. Bu yaklaşım, bağımlı değişkendeki aykırı değerlerin etkisini hafifletir. Dolayısıyla, önerilen yöntem hem bağımlı hem de bağımsız değişkenlerdeki aykırı değerlerin etkilerini en aza indirgeyerek parametre fonksiyonu için dirençli tahminler sağlar. Önerilen yöntemin kestirim ve sınıflandırma performansı ile hesaplama hızı, Monte-Carlo denemeleri ve gerçek veri setleri üzerinde yapılan analizler ile kapsamlı bir şekilde değerlendirilmiştir. Sonuçlar, mevcut dirençli ve dirençli olmayan yöntemlerle karşılaştırılmıştır. Elde edilen bulgular, önerilen yöntemin aykırı değerlerin varlığında mevcut yöntemlere göre daha başarılı sonuçlar sağladığını ve veri setinde aykırı gözlem bulunmadığında ise mevcut yöntemlerle rekabetçi performans sergilediğini göstermektedir.

Anahtar Kelimeler: Fonksiyonel Veri Analizi, Fonksiyonel Lojistik Regresyon, Dirençli Fonksiyonel Temel Bileşen Regresyonu, Bianco ve Yohai tahmincisi, El Radyografisi veri seti, Monte Carlo Denemeleri.

ABSTRACT

A ROBUST ESTIMATION METHOD FOR SCALAR-ON-FUNCTION LOGISTIC REGRESSION MODEL

The functional logistic regression model is commonly applied to model the linear relationship between a binary response, which takes binary values like 0 and 1, and a functional explanatory variable comprised of random curves. Most of the existing methods used to estimate the parameters of this method are hindered by outliers. The presence of outliers may lead to erroneous parameter estimates and increased probabilities of missclassification. In this thesis, we introduce a novel approach for the robust estimation of parameter functions in the functional logistic regression model. In this proposed method, observations of the functional explanatory variable are projected onto a finite-dimensional space via a robust functional principal component analysis, that reduces the impact of outliers. Consequently, the effects of outliers, which could potentially worsen the model, are minimized. The parameters of the logistic regression model, constructed using the binary response and robust principal component scores, are estimated using an M-type estimation method. This strategy leads to decrease the effects of outliers in the binary response variable. Hence, the proposed method provides robust estimates for the parameter function, reducing the effects of outliers in both the response and explanatory variables. The estimation and classification performance as well as the computing power of the proposed method is evaluated via a series of Monte-Carlo experiments and an empirical data analysis. The results are compared with existing robust and not-robust methods. The findings demonstrate that, compared with existing methods, the proposed method produces improved performance in the presence of outliers. Furthermore, when no outlier is present in the data, the proposed method produces competitive performance with the existing methods.

Keywords: Functional Data Analysis, Functional Logistic Regression, Robust Functional Principal Component Regression, Bianco and Yohai Estimator, Hand Radiograph Data Set, Monte Carlo Simulations.

KISALTMALAR

AUC : ROC Eğrisi Altında Kalan Alan

BY : Bianco ve Yohai

CCR : Doğru Sınıflandırma Oranı

FlogitR : Fonksiyonel Lojistik Regresyon

FPCA : Fonksiyonel Temel Bileşen Analizi

FPLS : Fonksiyonel Kısmi En Küçük Kareler

IMSE : İntegral Hata Kareler Ortalaması

LASSO : En Küçük Mutlak Daralma ve Seçim Operatörü

RFPCA : Dirençli Fonksiyonel Temel Bileşen Analizi

RFPLS : Dirençli Fonksiyonel Kısmi En Küçük Kareler

ŞEKİL

Şekil 5.1 Süreç 1 altında yaratılan verilerin temsili görselleri.

Şekil 5.2 Süreç 2 altında yaratılan verilerin temsili görselleri.

Şekil 5.3 FPLS, FPCA, RFPLS ve RFPCA yöntemleri için süreç 1 altında hesaplanan AUC (sol panel) ve CCR (sağ panel) değerlerinin kutu grafikleri.

Şekil 5.4 FPLS, FPCA, RFPLS ve RFPCA yöntemleri için süreç 1 altında hesaplanan IMSE değerlerinin kutu grafikleri

Şekil 5.5 Süreç 1 altında 200 Monte Carlo deneyinden elde edilen tahmini parametre fonksiyonları (gri çizgiler), ortalamaları (kırmızı çizgiler) ve gerçek parametre fonksiyonlarının (mavi çizgiler) temsili görseli.

Şekil 5.6 FPLS, FPCA, RFPLS ve RFPCA yöntemleri için süreç 2 altında hesaplanan AUC (sol panel) ve CCR (sağ panel) değerlerinin kutu grafikleri.

Şekil 5.7 El radyografisi için oluşturulan el modelleri.

Şekil 5.8 El radyografisi için iyi el verileri (sol panel), kötü el verileri (orta panel) ve sınıflarına göre tespit edilen aykırı değerlerin (sağ panel) grafik gösterimi.

Şekil 5.9 El radyografisi verileri için hesaplanan CCR ve AUC değerlerinin kutu grafiği.

Şekil 5.10 El radyografisi verileri için tahmin edilen parametre fonksiyonlarını (gri çizgiler) ve tahmin edilen parametre fonksiyonlarının ortalamasını (kırmızı çizgiler) temsil etmektedir.

TABLO LİSTESİ

Tablo 5.1 Süreç 1 için FPLS, FPCA, RFPLS ve RFPCA yöntemlerinin temel bileşen sayısına göre tek bir döngüyü hesaplama süreleri (saniye cinsinden). (Gözlem Sayısı = 1000).

Tablo 5.2 Süreç 1 için FPLS, FPCA, RFPLS ve RFPCA yöntemlerinin gözlem sayısına göre tek bir döngüyü hesaplama süreleri (saniye cinsinden). (Temel Bileşen Sayısı = 15).

Tablo 5.3 Süreç 2 için FPLS, FPCA, RFPLS ve RFPCA yöntemlerinin temel bileşen sayısına göre tek bir döngüyü hesaplama süreleri (saniye cinsinden). (Gözlem Sayısı = 1000).

Tablo 5.4 Süreç 2 için FPLS, FPCA, RFPLS ve RFPCA yöntemlerinin gözlem sayısına göre tek bir döngüyü hesaplama süreleri (saniye cinsinden). (Temel Bileşen Sayısı = 15).

1. GİRİŞ

Fonksiyonel veri analizi, elemanları zaman, eğri, görüntü ya da benzeri koordinat sistemleri üzerinde gözlemlenen verilerin modellenmesini sağlayan analitik bir yöntemdir. Bu yöntem, verilerin fonksiyonel formunu tahmin etmek, sınıflandırmak ya da regresyon gibi çeşitli amaçlar için kullanılabilir. Bu tez, yanıtın 0 ve 1 gibi ikili sonuca sahip olduğu, açıklayıcı değişkenin ise rastgele eğrilerden oluştuğu fonksiyonel lojistik regresyon (FlogitR) modelinin parametrelerinin tahmin edilmesine odaklanmaktadır.

FlogitR modelinde parametre tahmini ve yanıt değişkeninin sınıflandırılması amacı ile yapılan ilk çalışmalar fonksiyonel açıklayıcı değişkenin ve regresyon parametre fonksiyonunun temel genişletme fonksiyonları kullanılarak sonlu boyutlu bir uzaya yakınsanmaları ile gerçekleştirilmiştir. Örneğin, (Ratcliffe ve ark., 2002) fonksiyonel açıklayıcı değişken ve regresyon parametre fonksiyonunun Fourier açılımları kullanılarak FlogitR için bir tahmin yöntemi önermiştir. (Araki ve ark., 2009) Gaussian temel genişletme yöntemini kullanarak FlogitR modelinin parametre tahminini gerçekleştirmiştir. (Mousavi ve Sørensen, 2017) temel genişletme yaklaşımlarını ve LASSO yöntemini, FlogitR modelinde regresyon parametre fonksiyonunun tahmin edilmesi için kullanmıştır.

Daha sonra, verideki değişkenliğin daha iyi yakalanabilmesi ve model parametrelerinin daha etkin bir şekilde tahmin edilebilmesi için boyut indirgeme yöntemlerinden olan Fonksiyonel Temel Bileşen Analizi (FPCA) ve Fonksiyonel En Küçük Kareler (FPLS) yöntemleri FlogitR modelinin tahminine genişletilmiştir. FPCA, ilk olarak sonsuz boyutlu uzayda gözlenen fonksiyonel bağımsız değişkenleri, genel temel genişletme yöntemlerine benzer bir biçimde sonlu boyutlu bir uzaya yansıtır ve ortonormal temel bileşenler ile bu temel bileşenlere karşılık gelen temel bileşen skorlarını elde eder. Bu yöntemde temel bileşenler elde edilirken, bağımsız değişkenler arasındaki kovaryansın maksimum yapılması prensibi göz önünde bulundurulur. Daha sonra, FlogitR modelindeki regresyon parametre fonksiyonu, temel bileşen skorları ile ikili yanıt değişkeni arasında kurulan lojistik regresyon modeli yardımıyla yaklaştırılır. FPLS yöntemi de, FPCA yöntemine benzer bir biçimde fonksiyonel değişken öğelerini sonlu boyutlu bir uzaya yansıtır. Ancak, FPCA yönteminin aksine, FPLS yönteminde bileşenler elde edilirken yanıt ve bağımsız değişkenler arasındaki kovaryansın maksimum yapılması prensibi dikkate alınır. (Escabias ve ark., 2004), (Escabias

ve ark., 2005), (Leng ve Müller, 2006), (Aguilera ve ark., 2008) ve (Wei ve ark., 2014) FlogitR modelinde parametre tahmini, sınıflandırma ve model seçimi için FPCA yöntemine dayalı çeşitli yöntemler önermiş ve elde ettikleri bulguları klasik lojistik regresyon modelleri ile kıyaslamışlardır. (Preda ve ark., 2007) ve (Escabias ve ark., 2007), FlogitR modelinin tahminlenmesi için FPLS yöntemine dayalı yöntemler önermişlerdir. Ayrıca, (Mousavi ve Sørensen, 2018), FlogitR modeli için FPCA, FPLR ve LASSO yöntemlerine dayalı tahmin edicilerin performanslarını karşılaştırmıştır.

Yukarıda bahsedilen çalışmalardan elde edilen bulgular, FPCA ve FPLS yöntemlerinin, FlogitR modelinin tahminlenmesinde temel genişletme yöntemlerine kıyasla daha etkin olduklarını göstermiştir. Bunun nedeni, temel genişletme yöntemlerine dayalı yaklaşımların performanslarının, modeldeki değişken sayısı ve kullanılacak temel genişletme fonksiyonlarının sayısından önemli derecede etkilenmesidir. Modelde çok sayıda fonksiyonel değişken olduğunda ve/veya fonksiyonel değişkenler sonlu boyutlu uzaya yansıtılırken çok sayıda temel genişletme fonksiyonu kullanıldığında, bu yöntemler kararsız sonuçlar verebilir (Matsui ve Konishi, 2009), (Beyaztas ve Shang, 2020), (Mutis ve ark., 2022). Bu tür durumlarda, temel genişletme yöntemlerine dayalı tahmin yöntemleri FlogitR modeli için yanıltıcı parametre tahminleri yapabilir ve bu da yüksek yanlış sınıflandırma olasılıklarının elde edilmesine yol açabilir.

Mevcut FPCA ve FPLS tabanlı tahmin yöntemleri, verilerin düzgün bir veri oluşturma sürecini izlediği durumlarda etkili sonuçlar verse de, bu yöntemler veri setindeki potansiyel aykırı gözlemlere karşı duyarlıdırlar. Aykırı değerlerin varlığı durumunda, bu yöntemler yanıltıcı parametre tahminleri ve zayıf model uyumu üretebilirler. Bildiğimiz kadarıyla, bu sorunun çözümüne yönelik sadece iki yöntem önerilmiştir. (Denhere ve Billor, 2016) çalışmalarında minimum kovaryans determinantı yaklaşımını FPCA yöntemine entegre ederek temel bileşen skorlarını elde etmişlerdir. Daha sonra, ikili yanıt ve temel bileşen skorları arasında kurulan lojistik regresyon modelinde parametre tahminleri en büyük olabilirlik yöntemi kullanılarak elde edilmiş ve FlogitR modelindeki regresyon parametre fonksiyonunu yaklaştırmışlardır. Ancak, bu yöntem tam olarak dirençli bir yöntem değildir çünkü en büyük olabilirlik yöntemi parametre fonksiyonunun yaklaştırılmasında kullanılmıştır. (Mutis ve ark., 2022), FlogitR modelinin aykırı gözlemlere karşı dirençli bir şekilde tahmin edilmesi amacıyla FPLS tabanlı yeni bir yöntem önermiştir. Önerilen

yöntemde, ilk olarak FPLS bileşenleri ağırlıklı olabilirlik prensibine göre oluşturulmuştur. Sonrasında, ikili yanıt ve FPLS skorları ile kurulan lojistik regresyon modelinin parametreleri, (Alqallaf ve Agostinelli, 2016) tarafından önerilen ağırlıklı olabilirlik tabanlı genel lineer model yardımıyla tahmin edilmiş ve regresyon parametre fonksiyonu yaklaştırılmıştır. (Mutis ve ark., 2022), sayısal analizlerinden elde ettikleri bulgularla, önerdikleri yöntemin hem bağımsız hem de yanıt değişkenindeki aykırı gözlemlere karşı dirençli tahminler ürettiğini göstermiştir. Bu yöntemin bir dezavantajı, ağırlıklı olabilirlik yönteminde kullanılan iterasyon tabanlı tahmin yönteminin hesaplama süresini önemli derecede arttırmasıdır. Örneğin, 5. Bölümde detaylı olarak sunulan sonuçlarımız, (Mutis ve ark., 2022) tarafından önerilen yöntemin dirençli olmayan yöntemlere kıyasla 40 ile 1200 kat daha fazla hesaplama süresi gerektirdiğini göstermiştir.

Bu tezde, FLogitR modelinin hem yanıt hem de açıklayıcı değişkenlerdeki aykırı gözlemlere karşı dirençli bir şekilde tahmin edilmesi amacıyla, hesaplama süresi açısından etkin yeni bir tahmin yöntemi önerilmektedir. Önerilen yöntemde ilk olarak, sonsuz boyutlu bir uzayda gözlemlenen açıklayıcı değişkenin elemanları, (Bali ve ark., 2011) tarafından önerilen dirençli temel bileşenler analizi (RFPCA) yöntemi ile sonlu boyutlu bir uzaya yansıtılmaktadır. RFPCA yöntemi, (Croux ve Ruiz-Gazen, 1996) tarafından önerilen dirençli bir projeksiyon yaklaşımı ile dirençli bir ölçek tahmincisini kullanarak fonksiyonel gözlemleri sonlu boyutlu bir uzaya dirençli bir şekilde yansıtmaktadır. Sonrasında, elde edilen RFPCA skorları ve ikili yanıt değişkeni arasında kurulan lojistik regresyon modelinin parametreleri, (Croux ve Haesbroeck, 2003) tarafından önerilen tahminleme prosedürü ile tahmin edilmiştir. (Croux ve Haesbroeck, 2003)'ün prosedürü, (Bianco ve Yohai, 1996) tarafından önerilen Bianco ve Yohai tahmincisini kullanmaktadır. Bu tahminci, hem yanıt hem de açıklayıcı değişkenlerdeki aykırı gözlemlere karşı dirençlidir. Dolayısıyla, önerilen yöntemimiz, hem yanıt hem de açıklayıcı değişkenlerdeki aykırı gözlemlere karşı dirençli tahminler üretmektedir. Önerilen yöntemimiz, ağırlıklı olabilirlik yönteminin aksine, parametre tahmininde ileri seviye iterasyonlar gerektirmediğinden, hesaplama zamanı açısından etkin bir yöntem olarak kullanılabilir. Sayısal analizlerimizde elde ettiğimiz bulgular, veri setinde aykırı gözlem bulunmadığında, önerilen yöntemin dirençli olmayan yöntemlere kıyasla rekabetçi tahminler ürettiğini, ayrıca, aykırı değerlerin varlığı durumunda ise klasik yöntemlere kıyasla daha iyi parametre tahmini ve daha iyi doğru

sınıflandırma olasılıkları ürettiğini göstermektedir. Ayrıca, sonuçlarımız, önerilen yöntemin mevcut dirençli yönteme kıyasla, rekabetçi ve hatta daha iyi parametre tahmini ve doğru sınıflandırma olasılık değerlerini 4 ile 150 kat daha az hesaplama süresi ile ürettiğini göstermektedir. Dolayısıyla önerilen yöntem, fonksiyonel lojistik regresyon modellerinde, aykırı değerlerin varlığı durumunda, model tahminini etkin ve hızlı bir şekilde gerçekleştirmeye olanak tanımaktadır.

2. FONKSİYONEL VERİ ANALİZİ

Fonksiyonel Veri Analizi (FDA), gözlemlenen noktasal verilerin pürüzsüz bir eğri veya fonksiyon olarak ifade edilmesine olanak sağlayan bir istatistiksel analiz yöntemidir. FDA, geleneksel veri analizindeki skaler veya vektörel veriyi, sürekli bir fonksiyon olarak modeller. Bu modelde ki her bir eğri fonksiyonel gözlemler ile oluşturulur ve Hilbert uzayında tanımlanır.

FDA, bu verileri analiz ederken, genellikle verilerin düzgünlüğünü ve sürekliliğini korumak için pürüzsüz bir eğri oluşturur. Bu eğrilerin oluşturulması için B-spline, Fourier, Gaussian radial ve polinom gibi temel genişletme fonksiyonları sıklıkla kullanılmaktadır. Ancak B-spline fonksiyonları, her bir alt aralığa uygun ve ayrı bir parçalı polinom olduğu için, verinin karmaşık yapılarını ve değişkenliklerini belirlemede etkilidir. Bu durumdan dolayı B-spline fonksiyonları en sık kullanılan temel genişletme yöntemlerinden biridir. Belirlenen temel fonksiyonlar, ayrık $t, (t_1, t_2, \dots t_j) \in \mathcal{T}$ kapalı ve sınırlı noktalarda gözlemlenen stokastik süreci tanımlar. Bu süreçte yer alan $\mathcal{X}_i(t)$ açıklayıcı değişkenler ile $\beta(t)$ regresyon katsayısı fonksiyonu aşağıdaki gibi tanımlanır.

$$\mathcal{X}_i(t) \approx \sum_{k=1}^{K} a_{ik}\phi_k(t). \tag{2.1}$$

$$\beta(t) \approx \sum_{k=1}^{K} \beta_k\phi_k(t). \tag{2.2}$$

$\beta(t)$ regresyon katsayısı fonksiyonu, K adet $\phi(t)$ temel fonksiyonların doğrusal bir kombinasyonunu ve a_{ik}, temel fonksiyonlara karşılık gelen katsayıları temsil eder. Buradaki amaç $\mathcal{L}_2$ ayrılabilir Hilbert uzayında tanımlı integrallenebilir rastgele fonksiyonlar yardımı ile $\beta(t)$ regresyon katsayısı fonksiyonunun düzgün bir şekilde tahminlenmesidir.

2.1. Hilbert Uzayı

$\mathcal{H}$ bir boş değerden farklı vektör uzayı olsun. $\mathcal{H}$ üzerindeki iç çarpımı, $\langle .,. \rangle : d: H \times H \to \mathbb{R}$ olacak şekilde gösteririz. Bu d dönüşümü her $x, y, z \in H$ ve $\alpha, \beta \in \mathbb{R}$ için,

i) $d\langle x, x\rangle \geq 0$ ve $d\langle x, x\rangle = 0, x = y$.

ii) $d\langle x, y\rangle = \overline{d\langle y, x\rangle}$.

iii) $d\langle \alpha x + \beta z, y\rangle = \alpha\langle x, y\rangle + \beta\langle z, y\rangle$.

sırasıyla pozitiflik, simetrilik ve doğrusallık özelliklerine sahip bir fonksiyon olduğu varsayılır. Bu özellikler göz önünde bulundurulduğunda $(\mathcal{H}, \langle .,. \rangle)$ çiftine iç çarpım uzayı denir. Bu durumda aşağıdaki eşitsizlik elde edilebilir.

$$|\langle x, y\rangle|^2 \leq \langle x, x\rangle\langle y, y\rangle. \tag{2.3}$$

Burada $|.| : \mathcal{H} \to \mathbb{R} \to [0, \infty)$ fonksiyonu bir norm olarak tanımlanır ve $\mathcal{H}$ iç çarpımın normu $|.|$ ile gösterilir. Bu durumda, $(H, \langle .,. \rangle)$ çifti bir normlu uzaydır. Bu $(\mathcal{H}, \langle .,. \rangle)$ içindeki iç çarpım uzayı Cauchy dizisi normuna göre yakınsak ise bu iç çarpım uzayı bir Hilbert uzayı $(\mathcal{H})$ olarak adlandırılır.

2.2. L_2 Normu

$\mathcal{H}$ uzayında tanımlı olan sınırlı ve doğrusal operatörlerin uzayına $\mathcal{L}$ diyelim. $L^2(\mathcal{T})$, $\mathcal{T} = [a, b]$ aralığında tanımlı ve $\int_{\mathcal{T}} x^2(t_1)dt_1 < \infty$ koşulunu sağlayan ölçülebilir fonksiyonların oluşturduğu bir uzaydır. Bu uzay, aşağıdaki Hilbert-Schmidt operatörü gibi tanımlanır.

$$\mathcal{C}(x)(t_1) = \int_{\mathcal{T}} \mathcal{C}(t_1, t_2)x(t_2)dt_2, \quad x \in \mathcal{L}^2. \tag{2.4}$$

Burada $\Psi : \mathcal{H} \to \mathcal{H}$ kompakt bir operatördür. Bu integrallenebilen operatörün Hilbert-Schmidt operatörü olabilmesi için aşağıdaki koşulu sağlaması gerekmektedir:

$$\iint_{\mathcal{T}} \Psi^2(t_1, t_2)dt_1dt_2 < \infty. \tag{2.5}$$

Bu koşul sağlandığında, operatör aşağıdaki normu karşılar:

$$|\mathcal{C}|_{t_2}^2 = \iint_{\mathcal{T}} \mathcal{C}^2(t_1, t_2)dt_1dt_2. \tag{2.6}$$

Eğer $\mathcal{C}(t_2, t_1) = \mathcal{C}(t_1, t_2)$ ve $\iint_{\mathcal{T}} \mathcal{C}(t_1, t_2)x(t_1)x(t_2)dt_1dt_2 \geq 0$ ise, $\mathcal{C}$ integral operatörü simetrik ve pozitif tanımlıdır. Bu durumda, operatör aşağıdaki gibi ifade edilebilir:

$$\mathcal{C}(t_1, t_2) = \sum_{p=1}^{\infty} \lambda_p \psi_p(t_1)\psi_p(t_2), \quad \in \mathcal{L}_2(\mathcal{T} \times \mathcal{T}). \tag{2.7}$$

Bu ifade Mercer Teoremi olarak bilinir. Bu durumda, λ_p özdeğerleri ve ψ_p özfonksiyonlarıdır. Mercer Teoremi'nin bu özelliği, özellikle yüksek boyutlu verilerle çalışırken, verileri daha düşük boyutlu bir uzaya indirgemek ve bu şekilde daha karmaşık veri modellerini analiz etmek için kullanılır.

2.3. B-Spline Temel Fonksiyonları

B-Spline temel fonksiyonları, veri setindeki noktasal gözlemlere istenilen düzgünlüğü sağlama ve aynı zamanda bu düzgünlüğü muhafaza etmek adına daha az sayıda fonksiyon kullanabilen matematiksel işlevlerdir. (De Boor, 2001) tarafından geliştirilen B-Spline temel fonksiyonları, hızlı hesaplama kabiliyeti sayesinde, bilimsel hesaplamalar ve fonksiyonel veri analizi konularında büyük bir avantaj sağlar (Beyaztas ve Shang, 2022). Ayrıca, Fourier analizine kıyasla, B-Spline temel fonksiyonlarının daha esnek bir yapıya sahip olduğu belirtilmiştir (Ramsay ve ark., 2009). Bu esneklik, karmaşık veri trendlerine düzgün eğrileri uydurma yeteneğiyle kendini gösterir. Bu durum, B-Spline temel fonksiyonlarının verinin özfonksiyonlarını ve özdeğerlerini daha doğru bir şekilde belirlemesini sağlar.

B-Spline fonksiyonları, verideki ölçüm aralığının alt aralıklara ayrılmasıyla oluşturulur. Bu alt aralıkların (düğüm noktaların), sınırlarının belirlendiği maksimum ve minimum noktalara ise, kırılma noktası denir. Bu temel fonksiyonlar, her bir segmentin belirli bir j'inci derecede parçalı fonksiyon olduğu ve farklı segmentler arasındaki birleşme noktalarında düzgün bir şekilde bağlandığı spline'lar olarak tanımlanabilir. Bu bağlamda, bir spline'ın düzgünlüğü genellikle belirli bir derece ile ilişkilendirilir ve bu derece, belirli bir sayıda kesme noktasına sahip olan ve belirli bir sıra ile düzgünleştirilmiş B-Spline'ların, baz fonksiyonlarına dayalı olarak elde edilen doğrusal kombinasyonları şeklinde ifade edilir.

B-Spline baz fonksiyonlarının oluşturulması, belirli bir aralıkta tanımlanan düğüm noktaları dizisine dayanır. Bu düğüm noktaları $t = t_1, \dots, t_{j+1}$ olarak adlandırılır ve t aralığında tanımlanır. Buna ek olarak, t düğümlerin bir kesme noktaları dizisi, $t_0 < t_1 < \cdots < t_j$ olarak

gösterilir. Bu durumda, k'ıncı B-Spline temel fonksiyonu, $\phi_k(t) = \phi_{k,j}(t)$ şeklinde tanımlanabilir.

B-Spline tabanlı bir fonksiyon oluştururken, düğüm dizisi sınırları olan $[t_0, t_j]$ aralığının dışında ek düğümlere ihtiyaç duyulur. Bu nedenle, $\cdots < t_{-1} < t_0 < t_1 < \cdots < t_j < t_{j+1} < \cdots$ gibi bir düğüm dizisi, genişletilmiş veya arttırılmış bir düğüm dizisi olarak tanımlanır. Bu genişletilmiş düğüm dizisi, B-Spline temel fonksiyonlarının oluşturulması için gereklidir ve bu temel fonksiyonlar genişletilmiş düğüm dizisine dayanarak tanımlanır. Bu, B-Spline temel fonksiyonlarının esnekliğini artırır ve karmaşık veri yapılarını yakalamak için daha fazla adaptasyon sağlar. Bu durumlar göz önünde bulundurulduğunda B-Spline temellerini aşağıdaki duruma bağlı olarak tanımlayabiliriz.

$$\phi_{k,1}(t) = \begin{cases} 1, & t_k \leq t \leq t_{k+1} \\ 0, & \text{diğer durumlarda.} \end{cases} \tag{2.8}$$

Sabit B-Spline açılımların t düğüm noktasına bağlı olarak j'inci dereceye sahip k'ıncı B-Spline eşitliği aşağıdaki gibi tanımlanır.

$$\phi_{k,j}(t) = \frac{t - t_k}{t_{k+j-1} - t_k} B_{k,j-1}(t) + \frac{t_{k+j} - t}{t_{k+j} - t_{k+1}} B_{k+1,j-1}(t), \quad t_k \in t, \quad j \geq 2. \tag{2.9}$$

B-Spline temel fonksiyonları elde edildikten sonra $\mathcal{X}_i(t)$ fonksiyonlarının düzgünleştirilmesi için a katsayısı temel bileşen analizi ile belirlenebilir.

2.4. Fonksiyonel Temel Bileşen Analizi

Temel Bileşen Analizi (PCA), çok değişkenli veriler üzerinde boyut indirgeme işlemi gerçekleştirmek için kullanılan önemli tekniklerden birisidir (Jolliffe, 2002). PCA'nın ana prensipleri ilk kez (Karhunen, 1946) ve (Loève, 1946) tarafından tanımlansada, Fonksiyonel Temel Bileşenler Analizi (FPCA) için istatistiksel çıkarım uygulamalarında daha geniş bir çerçeve ilk defa (Kleffe, 1973) tarafından belirlenmiştir. O günden bu yana, FPCA, FDA içerisinde en sık kullanılan araçlardan biri haline gelmiştir. Bu popülarite, FPCA'nın, doğası gereği sonsuz boyuta sahip olan fonksiyonel veriyi sonlu boyutlu bir rastgele skor vektörüne dönüştürme yeteneğinden kaynaklanmaktadır.

Özellikle, $\mathcal{X}_i(t)$'nin, ilgili süreçlerin kovaryans operatörünün özfonksiyonlarına dayanan bir fonksiyonel baz üzerinden genişletilmesi yoluyla boyut indirgeme sağlanmaktadır. Bu süreçler, kovaryans fonksiyonunun herhangi bir $\mathcal{C}(t_1, t_2) \in L_2$ fonksiyonu için $\int_{\mathcal{T}} \int_{\mathcal{T}} \mathcal{C}(t_1, t_2) dt_1 dt_2$ biçiminde tanımlanır, burada hem kovaryans operatörü hem de kovaryans fonksiyonu aynı süreci kullanır. Bu süreç, kovaryans işlemlerinin genel tanımını ve fonksiyonların nasıl işlediğini belirtirken kullanılır. İntegral formundan ötürü, (lineer) kovaryans fonksiyonu iz sınıfına aittir ve bu sebeple kompakt bir Hilbert-Schmidt operatörüdür (Conway, 2019). Simetrik ve pozitif kesin olmasından dolayı, gerçek değerli ve negatif olmayan özdeğerlere sahiptir. Bazı varsayımlar altında, Mercer teoremi $\mathcal{C}(t_1, t_2) = \sum_{p \geq 1} \lambda_p \psi_p(t_1) \psi_p(t_2)$ eşitliğini sağlar, burada t_1 ve $t_2 \in \mathcal{C}$ için düzgün bir yakınsama vardır. λ_k özdeğerler düşen sıradadır ve ψ_P ilgili ortogonal özfonksiyonlardır.

$$\mathcal{X}_i(t) = \sum_{p=1}^{P} \xi_{ip} \psi_p(t). \tag{2.10}$$

Burada, $\xi_p = \int_{\mathcal{T}} \mathcal{X}(t) \psi_p(t) dt$, $\mathcal{X}i(t)$'nin FPC skorları olarak adlandırılır. ξ_p değerleri, i boyunca bağımsız ve p boyunca korelasyon göstermezler, ve bu koşullar için $E(\xi_{ip}) = 0$ ve $\text{var}(\xi_{ip}) = \lambda_p$ eşitlikleri geçerlidir. Denklem (2.10)'daki genişleme işlemi, yeterince büyük bir p değeri için ilk p terimi, sonsuz toplamı, sonlu boyutlu uzaya yakınsanır ve $\mathcal{X}_i$ açıklayıcı değişkene iyi bir yaklaşım sağlar. Bu, boyut indirgeme sürecini kolaylaştırır çünkü bu durumda $\mathcal{X}_i$'deki değerler esasen p boyutlu $\xi_i = (\xi_{i1}, \ldots, \xi_{ip})$ vektörüne taşınır ve bu vektör, yaklaşık süreçleri temsil eder.

Fonksiyonel veri, doğası gereği sonsuz boyutludur ve bu, analizini karmaşıklaştırır. FPCA, yukarıdaki koşullar altında bu verileri anlamlı ve yönetilebilir bir şekilde analiz etme yeteneği sağlar. Bahsi geçen özdeğerler ve karşılık gelen özfonksiyonlar, verinin temsilini oluşturan ana bileşenlerdir ve genellikle bu bileşenlerin sadece yüksek özdeğerlere sahip olanlar toplam varyansın büyük bir bölümünü açıklar. Böylelikle, yüksek boyutlu bir veri seti, daha düşük boyutlu bir alt uzaya indirgenebilir. Bu durum, verinin daha kolay anlaşılmasını ve analiz edilmesini sağlar, ayrıca hesaplama süresini ve veri saklama yükünü hafifletir. Ek olarak, bu yöntem genellikle veri setindeki gürültüyü azaltır ve bu sayede daha açık ve net sonuçlar elde edilir. Bu nitelikler, FPCA'nın fonksiyonel veri analizi için neden bu denli kullanışlı ve güçlü bir araç olduğunu ortaya koymaktadır.

3. FONKSİYONEL LOJİSTİK REGRESYON

Rastgele seçilmiş $\{\mathcal{Y}_i, \mathcal{X}_i(t): i = 1,2, \dots, n\}$ bir örneklem setinden bağımsız ve benzer dağılıma sahip $\{\mathcal{Y}, \mathcal{X}\}$ değişkenlerini kullanarak oluşturulmuş FLogitR modelini düşünelim. Burada yanıt değişkeni $\mathcal{Y} \in \{0,1\}$ ikili sınıfı temsil ederken, bağımsız değişken $\mathcal{X} = \{\mathcal{X}(t)_{t\in\mathcal{T}}\}$ ise, sınırlı ve kapalı bir $t \in \mathcal{T}$ aralığında tanımlı bir stokastik bir süreci ifade etmektedir. Bu süreç, $\mathcal{L}_2$ ayrılabilir Hilbert uzayında tanımlı ve $\mathcal{T} \subset \mathbb{R}$ üzerinde integrallenebilir bir rastgele fonksiyon $\mathcal{X}: \mathcal{T} \to \mathbb{R}$ olarak varsayılmaktadır. FLogitR modelinde $\{\mathcal{Y}, \mathcal{X}\}$ arasındaki ilişki (Ratcliffe ve ark., 2002) tarafından aşağıdaki gibi tanımlanmaktadır:

$$\mathcal{Y}_i = \pi_i + \varepsilon_i, \quad i = 1, \dots, n. \tag{3.1}$$

Burada ε_i, sıfır ortalamalı ve $\pi_i(1-\pi_i)$ varyanslı bağımsız rastgele hata terimini temsil ederken π ise fonksiyonel açıklayıcı değişken verildiğinde yanıt değişkeninin 1 değerini alma olasılığını göstermektedir ve aşağıdaki gibi tanımlanmaktadır:

$$\pi_i = \mathbb{P}[\mathcal{Y} = 1|\mathcal{X}(t) = \mathcal{X}_i(t)] = \frac{exp\{\alpha+\int_{\mathcal{T}} \mathcal{X}(t)\beta(t)dt\}}{1+exp\{\alpha+\int_{\mathcal{T}} \mathcal{X}(t)\beta(t)dt\}}. \tag{3.2}$$

Burada α sabit terimi gösterirken $\beta(t)$ regresyon parametre fonksiyonunu temsil etmektedir. (3.2) denklemin çözümlenebilmesi için logit dönüşümü uygulanır ve (3.3) ile tanımlı FLogitR modeli logit dönüşümü ile aşağıdaki gibi ifade edilir:

$$l_i = ln\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \int_{\mathcal{T}} \mathcal{X}_i(t)\beta(t)dt, \quad i = 1, \dots, n. \tag{3.3}$$

(3.3) eşitliği, tahmin edilen fonksiyonların etkisini, büyüklüğünü ve yönünü ifade eder. Bu sayede, modeldeki değişkenin etkilerini anlamak ve tahminler yapmak mümkün hale gelir. Daha ayrıntılı bir şekilde, $\beta(t)$ fonksiyonunun μ gibi bir sabit ile çarpılan integrali, fonksiyonel açıklayıcı değişkenin $\mathcal{T}$ boyunca μ birim arttırıldığında $\mathcal{Y} = 1$ olasılığındaki çarpımsal değişiklik olarak yorumlanabilir (Escabias ve ark., 2005) ve (Escabias ve ark., 2007).

3.1. FLogitR Modeli İçin Temel Genişletme Fonksiyonları

Model (3.3)'te yer alan fonksiyonel açıklayıcı değişken, sonsuz boyutlu bir uzayda tanımlı iken, bu değişkenin elemanları pratikte kesikli zaman noktaları üzerinde gözlemlenmektedir. Dolayısıyla, Model (3.3)'ün direkt bir çözümü mümkün değildir. Bu sorunun çözümüne pratikte, ilk olarak kesikli zaman noktalarında elde edilen gözlemlerin fonksiyonel formlarının yakınsaması ile başlanır. Kesikli zaman noktalarında elde edilen i-inci gözlemin fonksiyonel formu, K temel fonksiyon $\phi_1(t), \dots, \phi_K(t)$ ve bu temel fonksiyonlara karşılık gelen genişletme katsayılarının $a_{i1}, \dots, a_{iK}$ doğrusal bir kombinasyonu ile aşağıdaki gibi elde edilir:

$$\mathcal{X}_i(t) \approx \sum_{k=1}^{K} a_{ik}\phi_k(t) = \mathbf{a_i}^{\top}\boldsymbol{\phi}(t). \tag{3.4}$$

Benzer şekilde, regresyon parametre fonksiyonu da aynı temel genişletme fonksiyonları kullanılarak temsil edilebilir:

$$\beta(t) \approx \sum_{k=1}^{K} \beta_k\phi_k(t) = \boldsymbol{\beta}^{\top}\boldsymbol{\phi}(t). \tag{3.5}$$

Eşitlik (3.4) ve (3.5), Eşitlik (3.6)'da yerine konulduğunda, FLogitR modelinin logit dönüşümü sonlu boyutlu uzayda aşağıdaki gibi yakınsar:

$$l_i = \alpha + \int_{\mathcal{T}} \mathcal{X}_i(t)\beta(t)dt = \alpha + \sum_{k=1}^{K}\sum_{j=1}^{K} a_{ik}\Phi_{kj}\beta_j, \quad i = 1, \dots, n. \tag{3.6}$$

Burada $\phi_{kj} = \int_{\mathcal{T}} \phi_k(t)\Phi_j(t)dt, \quad k, j = 1, \dots, K,$ temel fonksiyonların iç çarpımını göstermektedir. Daha genel bir ifadeyle, FLogitR modelinin logit dönüşümü matris formunda aşağıdaki gibi ifade edilir:

$$\mathbf{L} = \alpha\mathbf{1} + \mathbf{A\Phi\beta}. \tag{3.7}$$

Burada $\mathbf{A} = (a_{ik})n \times K, \quad \mathbf{\Phi} = (\Phi kj)_{K\times K}, \quad \boldsymbol{\beta} = [\beta_1, \dots, \beta_K]^{\top}, \quad \mathbf{L} = [l_1, \dots, l_n]^{\top}$ ve $\mathbf{1} = [1, \dots, 1]^{\top}$ olarak tanımlanır. Özetle, sonsuz boyutlu olan FLogitR modeli, $\mathbf{L}$ ve rastgele $\mathbf{A\Phi}$ matrisleri kullanılarak kurulan bir lojistik regresyon modeli ile sonlu boyutlu uzayda yakınsar. $\widehat{\boldsymbol{\beta}} = [\hat{\beta}_1, \dots, \hat{\beta}_K]^{\top}$, $\mathbf{L}$ ve $\mathbf{A\Phi}$ matrisleri ile kurulan lojistik regresyon modelinden elde edilen parametre tahminini gösterdiğinde, FLogitR modelindeki regresyon parametre fonksiyonu aşağıdaki gibi yakınsanmış olacaktır:

$$\hat{\beta}(t) = \sum_{k=1}^{K} \hat{\beta}_k^{\top} \phi_k(t). \quad (3.8)$$

Bağımsız değişkenler arasında yüksek derecede bağımlılık söz konusu olduğunda, **AΦ** matrisi yüksek korelasyon içerebileceğinden, sonlu boyutlu uzayda kurulacak lojistik regresyon modelinde çoklu doğrusal bağlantı problemi ortaya çıkabilir. Böyle bir durumda, eşitlik (3.8) ile regresyon parametre fonksiyonu için yanlı tahminler elde edilebilir (Escabias ve ark., 2004). Bu sorunun çözümü için FPCA analizini de içeren boyut indirgeme yöntemleri en çok kullanılan yöntemlerdir.

3.2. FLogitR Modeli İçin FPCA Yaklaşımı

Fonksiyonel bağımsız değişken $\mathcal{X}(t)$ için kovaryans fonksiyonunu $\mathcal{C}(t_1, t_2) = \mathrm{Cov}[\mathcal{X}(t_1), \mathcal{X}(t_2)]$ olarak ve $\int_{\mathcal{T}} \int_{\mathcal{T}} \mathcal{C}(t_1, t_2) dt_1 dt_2 < \infty$ koşulunu sağlayacak şekilde tanımlayalım. Mercer teoremine göre kovaryans fonksiyonu aşağıdaki gibi ayrıştırılabilir:

$$\mathcal{C}(t_1, t_2) = \sum_{p \geq 1} \lambda_p \psi_p(t_1) \psi_p(t_2), \quad \forall t_1, t_2 \in \mathcal{L}_2(\mathcal{T}). \quad (3.9)$$

Burada $\lambda_p : p = 1,2, \ldots, \lambda_p > \lambda_{p+1}$ olacak şekilde negatif olmayan özdeğerleri gösterirken, $\psi_p(t) : p = 1,2, \ldots$ ise $\mathcal{L}_2(\mathcal{T})$ üzerinde tanımlı özdeğerlere karşılık gelen özfonksiyonların ortonormal temellerini göstermektedir. Pratikte, fonksiyonel açıklayıcı değişkendeki değişimin büyük bir bölümü sonlu sayıda P özfonksiyon ile yakalanır. Dolayısıyla, kovaryans fonksiyonu önceden belirlenmiş kesme sabiti olan P ile tahmin edilir, $\mathcal{C}(t_1, t_2) = \sum p = 1^{P} \lambda_p \psi_p(t_1) \psi_p(t_2)$. Ayrıca, özfonksiyonların ortonormal temelleri pratikte bilinmemektedir ve bir temel genişletme yöntemi kullanılarak yaklaşık olarak elde edilir. Bu tezde, basitliği ve yüksek performansından dolayı özfonksiyonların ortonormal temelleri B-spline temel genişletme yöntemi kullanılarak yaklaştırılmıştır.

Açıklayıcı değişkenin FPCA ayrıştırılması aşağıdaki gibi elde edilir:

$$\mathcal{X}(t) = \sum_{p=1}^{P} \xi_p \psi_p(t). \quad (3.10)$$

Burada ξ_p, p-inci özfonksiyonun ortonormal temeline karşılık gelen temel bileşen skorunu gösterir ve aşağıdaki gibi hesaplanır:

$$\xi_p = \int_{\mathcal{T}} \mathcal{X}(t) \psi_p(t) dt. \quad (3.11)$$

Pratikte, temel bileşen skorları hesaplanırken ilk olarak fonksiyonel açıklayıcı değişken bir temel genişletme fonksiyonu yardımı ile sonlu boyutlu uzaya yaklaştırılır. Elemanları $\phi_1(t), \dots, \phi_K(t)$ temel genişletme fonksiyonları ile sonlu boyutlu uzaya yansıtılmış fonksiyonel açıklayıcı değişkenin temel bileşen skorları, $\mathbf{A}\mathbf{\Phi}^{\frac{1}{2}}$ matrisinden standart temel bileşenler analizi yöntemi ile elde edilen temel bileşen skorlarına eşittir (Escabias ve ark., 2004). Burada $\mathbf{\Phi}^{\frac{1}{2}}$, temel fonksiyonların iç çarpımlarının matrisinin karekökünü göstermektedir. Daha detaylı bir ifadeyle, $\mathbf{G}$, kolonları $\mathbf{A}\mathbf{\Phi}^{\frac{1}{2}}$ matrisine standart temel bileşenler analizi uygulandığında elde edilen özvektörleri içeren matrisi gösterir. Burada temel bileşen skorlarının matrisi, $\mathbf{\Xi} = [\xi_1, \dots, \xi_P]^\top$, $\mathbf{\Xi} = \left(\mathbf{A}\mathbf{\Phi}^{\frac{1}{2}}\right)\mathbf{G}$ olarak elde edilirken, temel bileşenleri tanımlayan özfonksiyonların ortonormal temelleri $\psi_p(t) = \sum_{k=1}^{K} \psi_{kp}\phi_k(t)$ olarak elde edilir. Burada $\psi_{kp} = \mathbf{\Phi}^{-\frac{1}{2}}\mathbf{G}$ ifadesi, p-inci özfonksiyonun ortonormal temellerine ait temel genişletme katsayılarını göstermektedir.

Sonsuz boyutlu uzayda tanımlı FlogitR modeli için $\mathcal{X}(t)$'nin FPCA ayrıştırılması kullanılarak elde edilen yakınsak model aşağıdaki gibi ifade edilebilir:

$$l_i = \alpha + \int_{\mathcal{T}} \mathcal{X}_i(t)\beta(t)dt = \alpha + \sum_{k=1}^{K} \xi_{ik}\gamma_k = \mathbf{\Xi}_i^\top \boldsymbol{\gamma}, \quad i = 1, \dots, n. \tag{3.12}$$

Burada $\gamma_k = \int_{\mathcal{T}} \psi_k(t)\beta(t)dt$, $k = 1, \dots, K$. Eşitlik (3.12) ile verilen model daha detaylı bir şekilde aşağıdaki gibi ifade edilebilir:

$$l_i = \alpha + \sum_{k=1}^{K} \xi_{ik}\gamma_k = \alpha + \sum_{k=1}^{K} \left[\int_{\mathcal{T}} \mathcal{X}_i(t)\psi_k(t)dt\right]\gamma_k$$

$$l_i = \alpha + \int_{\mathcal{T}} \mathcal{X}_i(t)[\sum_{k=1}^{K} \psi_k(t)\gamma_k]dt, \quad i = 1, \dots, n. \tag{3.13}$$

Dolayısı ile regresyon parametre fonksiyonu $\beta(t) = \sum_{k=1}^{K} \psi_k(t)\gamma_k$ olarak ifade edilir. Eşitlik (3.12) ile verilen ve temel bileşen skorları ile kurulan model için asimtotik olarak en etkin tahminci, en büyük olabilirlik tahmincisidir. $\mathbf{Z}_i = (1, \mathbf{\Xi}_i^\top)^\top$ olarak ifade edilirse, (3.12) modeline ait en büyük olabilirlik tahmincisi aşağıdaki gibi tanımlanır:

$$\hat{\boldsymbol{\gamma}}^{\mathrm{ML}} = argmax \log L(\boldsymbol{\gamma}; \mathbf{Z}^\top) = \operatorname{argmin} \sum_{i=1}^{n} d(\mathbf{Z}_i^\top \boldsymbol{\gamma}; \mathcal{Y}_i). \tag{3.14}$$

Burada $\log L(\boldsymbol{\gamma}; \mathbf{Z}^{\top})$ koşullu log-olabilirlik fonksiyonunu gösterirken $d(\mathbf{Z}_i^{\top}\boldsymbol{\gamma}; \mathcal{Y}_i)$ ise sapma bileşenini göstermektedir ve aşağıdaki gibi elde edilir:

$$d(\mathbf{Z}_i^{\top}\boldsymbol{\gamma}; \mathcal{Y}_i) = -\mathcal{Y}_i \log F(\mathbf{Z}_i^{\top}\boldsymbol{\gamma}) - (1 - \mathcal{Y}_i) \log[1 - F(\mathbf{Z}_i^{\top}\tilde{\boldsymbol{\gamma}})]. \quad (3.15)$$

Burada $F(u) = \frac{1}{[1+\exp(-u)]}$ logit model fonksiyonunu temsil etmektedir. Sonuç olarak, regresyon parametre fonksiyonunun temel genişletme katsayısı $\widehat{\boldsymbol{\beta}} = \boldsymbol{\Phi}^{-\frac{1}{2}}\mathbf{G}\hat{\boldsymbol{\gamma}}^{\mathrm{ML}}$ olarak elde edilir. FlogitR modelindeki regresyon parametre fonksiyonu ise FPCA yöntemi ile $\hat{\beta}(t) = \widehat{\boldsymbol{\beta}}\boldsymbol{\phi}(t)$ olarak elde edilir.

FPCA yöntemi, temel genişletme yöntemine kıyasla, FlogitR modelindeki regresyon parametre fonksiyonunu olası çoklu bağlantı probleminden etkilenmeden etkin bir şekilde tahmin eder. Diğer taraftan, bu yöntem verideki olası aykırı gözlemlerden oldukça etkilenir. Aykırı gözlemlerin varlığı durumunda, bu gözlemlerin etkileri $\mathbf{A}\boldsymbol{\Phi}^{\frac{1}{2}}$ matrisinden elde edilen temel bileşenlere ve temel bileşen skorlarına yansır. Benzer şekilde, temel bileşenler üzerinden elde edilen regresyon parametre fonksiyonunun tahminini de ($\widehat{\boldsymbol{\beta}}$) etkiler. Aynı zamanda, en büyük olabilirlik tahmincisi aykırı gözlemlerden oldukça etkilenir ve aykırı gözlemlerin varlığı durumunda bu tahminci $\boldsymbol{\gamma}$ için yanlı tahminler üretir. Dolayısıyla, veri setinde aykırı gözlem olduğunda, FPCA yöntemi yanlı parametre tahminleri ile yüksek yanlış sınıflandırma olasılığına sebep olabilir. Bu tezde, aykırı gözlemlerin etkilerini en aza indirecek şekilde FlogitR modelinin tahminini etkin bir şekilde elde etmek için yeni bir yaklaşım öneriyoruz.

4. FlogitR MODELİ İÇİN ÖNERİLEN YÖNTEM

RFPCA yaklaşımı, aykırı gözlemlerin etkilerini en aza indirecek şekilde FlogitR modelini tahmin etmek için kullanılır. Bu yaklaşım, (Bali ve ark., 2011) tarafından önerilmiştir ve modelin temel bileşenlerini ve temel bileşen skorlarını dirençli bir şekilde elde eder. Sonrasında, ikili yanıt ve dirençli temel bileşen skorları ile kurulan lojistik regresyon modelinin parametreleri, (Croux ve Haesbroeck, 2003) tarafından önerilen Bianco ve Yohai (BY) tahmincisi ile tahmin edilir. Dolayısıyla, önerilen yöntem hem açıklayıcı değişken hem de ikili yanıt değişkenindeki aykırı gözlemlere karşı dirençlilik sağlar.

4.1. RFPCA Yöntemi

(Bali ve ark., 2011) tarafından önerilen RFPCA yöntemi, klasik FPCA yöntemine benzer bir şekilde çalışır fakat varyans yerine aykırı gözlemlerin varlığına karşı dirençli bir M-tipi ölçeklendirme fonksiyonu kullanır. Başlangıç olarak, $\| \alpha \|^2 = \langle \alpha, \alpha \rangle$, $\langle \cdot, \cdot \rangle$ iç çarpımı tarafından yaratılan normu göstersin. Ayrıca, $\mathcal{F}$ ve $\mathcal{F}[\alpha]$ sırasıyla; $\mathcal{X}$ ve $\langle \alpha, \mathcal{X} \rangle$ ifadelerinin dağılımlarını göstersin. Verilen bir M-ölçek fonksiyoneli $\sigma_M(\mathcal{F})$ için özfonksiyonların ortogonal temelleri $\tilde{\psi}$, (Bali ve ark., 2011) tarafından aşağıdaki gibi tanımlanmıştır:

$$\begin{cases} \tilde{\psi}_p(\mathcal{F}) = \underset{\|\alpha\|^2=1}{argmax}\, \sigma_M(\mathcal{F}[\alpha]), & p = 1, \\ \tilde{\psi}_p(\mathcal{F}) = \underset{\|\alpha\|^2=1, \alpha \in \mathcal{B}_p}{argmax}\, \sigma_M(\mathcal{F}[\alpha]), & p \geq 2. \end{cases} \tag{4.1}$$

Burada $\mathcal{B}_p = \left\{ \alpha \in \mathcal{L}_2(\mathcal{T}) : \left\langle \alpha, \tilde{\psi}_p(\mathcal{F}) \right\rangle = 0, 1 \leq p \leq \ P - 1 \right\}$. P-inci en büyük özdeğer $\tilde{\lambda}$ aşağıdaki gibi tanımlanmıştır:

$$\tilde{\lambda}_p(\mathcal{F}) = \sigma_M^2\left(\mathcal{F}\left[\tilde{\psi}_p\right]\right) = \max_{\|\alpha\|^2=1, \alpha \in \mathcal{B}_p} \sigma_M^2(\mathcal{F}[\alpha]). \tag{4.2}$$

σ_M için fonksiyoneli $\sigma_M(\mathcal{F}_n[\alpha])$ olarak tanımlayalım. $s_n^2 : \mathcal{L}_2[0,1] \to \mathbb{R}$, ampirik M-ölçek fonksiyonelinin fonksiyonunu göstersin; $s^2(\alpha) = \sigma_M^2(\mathcal{F}[\alpha])$. Bu durumda, $\mathcal{X}(s)$ için özfonksiyonların ortogonal temellerinin RFPCA tahminleri aşağıdaki gibi elde edilir.

$$\begin{cases} \hat{\tilde{\psi}}_p(t) = \underset{\|\alpha\|^2=1}{argmax}\, s_n(\alpha), & p = 1, \\ \hat{\tilde{\psi}}_p(t) = \underset{\alpha \in \widehat{\mathcal{B}}_p}{argmax}\, s_n(\alpha), & p \geq 2. \end{cases} \tag{4.3}$$

Burada $\widehat{\mathcal{B}}_p = \left\{ \alpha \in \mathcal{L}_2(\mathcal{T}) : \| \alpha \| = 1, \left\langle \alpha, \hat{\tilde{\psi}}_p \right\rangle = 0, \forall\, 1 \leq p \leq \ P - 1 \right\}$. Karşılık gelen özdeğerler ise aşağıdaki gibi elde edilir:

$$\hat{\tilde{\lambda}}_p = s_n^2\left(\hat{\tilde{\psi}}_p\right), \quad p \geq 1. \tag{4.4}$$

4.2. Flogitr Modeli İçin Bianco ve Yohai Tahmincisi

RFPCA yöntemi ile birlikte, sonlu boyuta indirgenmiş model matrislerini kullanarak fonksiyonel verilerde lojistik regresyon parametre tahminleri için ikili sınıflandırmaya dayalı FlogitR modeli uygulanabilir. 0-1 ikili yanıt değişkeninin sınıflandırılması için, (Croux ve Haesbroeck, 2003) çalışmasında sunulan Bianco ve Yohai (BY) maksimum olabilirlik (ML) tahmincisine başvurulmuştur. Bu yöntem, maksimum olabilirlik tahmincisine dayalı mevcut sağlam tahmincilerin hesaplanması için hızlı ve kararlı bir algoritma sunar. Maksimum olabilirlik tahmin denklemlerinin fonksiyonel veriye uygun bir şekilde tanımlanmış fonksiyonları, bağımsız Bernoulli dağılımına uygun bir maksimizasyon modeli ile değiştirilerek elde edilir. Bu yaklaşım, maksimum olabilirlik tahmin denklemlerinden türetilen fonksiyonel verilerin, daha uygun ve istatistiksel olarak daha sağlam tahminler üretmesine olanak sağlar.

Sonsuz boyutlu uzayda tanımlı FlogitR modelinin RFPCA ayrıştırılması ile elde edilen yakınsak modeli düşünelim:

$$l_i = \alpha + \sum_{k=1}^{K} \tilde{\xi}_{ik}\tilde{\gamma}_k = \tilde{\Xi}_i^{\top}\tilde{\boldsymbol{\gamma}}. \tag{4.5}$$

Burada $\tilde{\xi}_k = \int_{\mathcal{T}} \mathcal{X}(t)\tilde{\psi}_k(t)dt$ ve $\tilde{\gamma}_k = \int_{\mathcal{T}} \tilde{\psi}_k(t)\beta(t)dt$ olarak tanımlıdır. $\tilde{\mathbf{Z}}_i = (1, \tilde{\Xi}_i^{\top})^{\top}$ olsun. (Croux ve Haesbroeck, 2003) tarafından gösterildiği üzere $\boldsymbol{\gamma}$'nın dirençli bir tahmini eşitlik (3.14) ile verilen en çok olabilirlik tahmincisinde kullanılan d fonksiyonunun başka bir kayıp fonksiyonu ile değiştirilmesi ile elde edilir:

$$\hat{\tilde{\boldsymbol{\gamma}}} = \operatorname{argmin} \sum_{i=1}^{n} \rho\left(\tilde{\Xi}_i^{\top}\tilde{\boldsymbol{\gamma}}; \mathcal{Y}_i\right). \tag{4.6}$$

Burada ρ pozitif ve her yerde türevlenebilir bir fonksiyondur. Verilen herhangi bir skor $\kappa = \tilde{\Xi}i^{\top}\tilde{\boldsymbol{\gamma}}$ değeri için $\rho(\kappa; 0) = \rho(-\kappa; 1)$ eşitliğini sağlaması gerekir. Genelliği bozmadan, tek değişkenli, azalmayan ve $\lim\kappa \to -\infty\, \rho(\kappa) = 0$ koşulunu sağlayan $\rho(\kappa) = \rho(\kappa; 0)$ fonksiyonunu düşünelim. Burada $\mathcal{Y} = 0$'a karşılık gelen bir gözlem için $\rho(\kappa)$ değeri, κ skorunun eşitlik (4.5) ile verilen amaç fonksiyonundaki değer üzerine etkisini göstermektedir. Eşitlik (4.5)'den hareketle, $\hat{\tilde{\boldsymbol{\gamma}}}$ bir M-tipi tahmincidir ve ilgili birinci-derece koşulu aşağıdaki gibidir:

$$\frac{1}{n}\sum_{i=1}^{n}\rho'\left(\widetilde{\Xi}_i^{\top}\tilde{\boldsymbol{\gamma}};\mathcal{Y}_i\right)\widetilde{\Xi}_i^{\top}=0. \tag{4.7}$$

Burada $\rho'(\kappa;0)=\frac{\partial\rho(\kappa;0)}{\partial\kappa}$ kayıp fonksiyonu ρ'nun türevini gösterir ve $\rho'(\kappa;1)=-\rho'(-\kappa;0)$ eşitliğini sağlar. (Croux ve Haesbroeck, 2003) tarafından önerilen tahminci, (Bianco ve Yohai, 1996) tarafından önerilen tutarlı tahminciyi dikkate almaktadır:

$$\hat{\tilde{\boldsymbol{\gamma}}}=\underset{\tilde{\boldsymbol{\gamma}}}{\operatorname{argmin}}\sum_{i=1}^{n}\left\{\rho\left[d\left(\widetilde{\Xi}_i^{\top}\tilde{\boldsymbol{\gamma}};\mathcal{Y}_i\right)\right]+C\left(\widetilde{\Xi}_i^{\top}\tilde{\boldsymbol{\gamma}}\right)\right\}. \tag{4.8}$$

Burada $C\left(\widetilde{\Xi}_i^{\top}\tilde{\boldsymbol{\gamma}}\right)$ yanlılık düzeltme terimidir ve aşağıdaki gibi hesaplanır:

$$C(\kappa)=D[F(\kappa)]+D[1-F(\kappa)]+D(1). \tag{4.9}$$

Burada $D(u)=\int_0^u\rho'(-\ln u)du$ olarak ifade edilir. Eşitlik (4.8)'den görüldüğü üzere BY tahmincisi kullanlılan kayıp fonksiyonu ρ'ya bağlıdır ve (Croux ve Haesbroeck, 2003) tarafından önerilen kayıp fonksiyon aşağıda sunulmuştur:

$$\rho(u)=\begin{cases}ue^{-\sqrt{c}}, & \text{eğer } u\le c\\ -2e^{-\sqrt{u}}\left(1+\sqrt{u}\right)+e^{-\sqrt{c}}\left(2\left(1+\sqrt{c}\right)+c\right), & \text{diğer durumlarda.}\end{cases} \tag{4.10}$$

Burada c sabiti dirençlilik ve etkinlik arasındaki dengeyi sağlamaktadır. c'nin değeri arttıkça tahmincinin etkinliği artarken dirençliliği azalır, azalırkende tam tersi durum söz konusu olur. Bu tezde c değeri (Croux ve Haesbroeck, 2003) tarafından önerildiği üzere $c=0.5$ seçilmiştir. Bu durumda kayıp fonksiyonunun türevi,

$$\rho'(u)=\begin{cases}e^{-\sqrt{c}}, & \text{eğer } u\le c\\ e^{-\sqrt{u}}, & \text{diğer durumlarda.}\end{cases} \tag{4.11}$$

olarak elde edilirken $D(u)$ fonksiyonu da

$$D(u)=\begin{cases}ue^{-\sqrt{-\ln u}}+e^{1/4}\sqrt{\pi}\varphi(\sqrt{2}(\frac{1}{2}+\sqrt{-\ln u}))-e^{1/4}\sqrt{\pi}, & \text{eğer } u\le e^{-c}\\ e^{-\sqrt{c}}u+e^{1/4}\sqrt{\pi}+e^{1/4}\sqrt{\pi}\varphi(\sqrt{2}(\frac{1}{2}+\sqrt{c})), & \text{diğer durumlarda.}\end{cases} \tag{4.12}$$

olarak elde edilir. Burada φ normal dağılıma ait birikimli dağılım fonksiyonunu göstermektedir. RFPCA ve BY yaklaşımları ile birlikte elde edilen regresyon parametre

fonksiyonunun tahmini birlikte düşünüldüğünde; eğer fonksiyonel açıklayıcı değişkende aykırı gözlem varsa RFPCA yöntemi ile bu değişkenlerin etkileri en aza indirgenerek temel bileşenler ve temel bileşen skorları dirençli bir şekilde elde edilir. Ayrıca, BY tahmincisi ile, eğer ikili yanıt değişkeninde aykırı gözlem varsa bu gözlemlerin etkileri en aza indirgenerek sonlu boyutlu uzayda kurulan yakınsak lojistik modelinin parametreleri dirençli bir şekilde tahmin edilir, $\hat{\tilde{\boldsymbol{\gamma}}}$. Dolayısıyla, sonsuz boyutlu FlogitR modeli için elde edilen regresyon parametresinin tahmini $\hat{\beta}(t) = \sum_{k=1}^{K} \tilde{\psi}_k(t)\hat{\tilde{\boldsymbol{\gamma}}}$, hem açıklayıcı değişkendeki hem de ikili yanıt değişkenindeki aykırı gözlemlerin etkileri en aza indirgenerek tahmin edilir.

5. UYGULAMA

Bu bölümde, önerilen RFPCA yönteminin tahmin, sınıflandırma ve hesaplama süresi performansını çeşitli Monte Carlo denemeleri ve bir gerçek veri uygulaması (El Radyografisi verisi) ile değerlendirmekteyiz. Önerilen yöntemin performansı, klasik FPLS ve FPCA ile (Mutis ve ark., 2022) tarafından önerilen dirençli FPLS (RFPLS) yöntemleriyle karşılaştırılmaktadır. Monte Carlo denemelerinde, (Escabias ve ark., 2007), (Mousavi ve Sørensen, 2018) ve (Mutis ve ark., 2022) tarafından kullanılan iki farklı veri yaratım sürecini göz önünde bulunduruyoruz. Her iki veri yaratım süreci için 200 Monte Carlo denemesi gerçekleştirilmiş ve her bir denemede aşağıdaki süreç tekrarlanmıştır.

İlk olarak, $n = 1000$ gözlemden oluşan veri seti oluşturulmuş, rastgele seçilen 700 gözlem eğitim seti olarak ayrılmış ve FPCA, FPLS, RFPLS ve RFPCA yöntemleri bu set üzerinde uygulanmıştır. Ardından, kurulan modellerle bağlantılı olarak geri kalan 300 gözlemden oluşan test kümesi üzerinde ikili yanıt değişkeninin değerleri tahmin edilmiştir.

Her iki veri yaratım süreci için iki farklı senaryo ele alınmıştır. İlk senaryoda, veriler düzgün bir veri oluşturma sürecinden türetilmiş ve veride herhangi bir aykırı gözlem bulunmamaktadır. Bu senaryo ile amaç, önerilen yöntemin tahmin ve sınıflandırma tutarlılığını test etmektir. İkinci senaryoda ise, eğitim kümesinden rastgele seçilen $[\%1, \%5, \%10]$'luk gözlemler aykırı gözlemlerle kirletilmiştir. Bu senaryodaki amaç, önerilen yöntemin aykırı gözlemlerin varlığı durumunda mevcut yöntemlere kıyasla performansını değerlendirmektir. Ayrıca, her iki veri yaratım süreci için de önerilen yöntemin hesaplama süresi açısından etkinliği test edilmektedir.

Monte-Carlo denemeleri sırasında, tüm yöntemler için veri setindeki değişkenliğin en az %95'ini açıklayacak şekilde 5 temel bileşen kullanılarak modeller oluşturulmuştur. Ayrıca, tüm yöntemler için $K = 15$ B-spline temel genişletme fonksiyonu kullanılarak temel bileşenler ve parametre fonksiyonları yakınsanmıştır. İlk veri yaratım süreci, belirli bir parametre fonksiyonuna bağlı olarak oluşturulmuş ve yöntemlerin tahmin performansları integral hata kareler ortalaması (IMSE) kullanılarak karşılaştırılmıştır:

$$\text{IMSE} = \int_{\mathcal{T}} \left[\beta(t) - \hat{\beta}(t)\right]^2 dt. \tag{5.1}$$

Yöntemlerin sınıflandırma performansları ise, doğru sınıflandırma oranı (CCR) ve ROC eğrisi altında kalan alan (AUC) metrikleri kullanılarak karşılaştırılmıştır. İkinci veri yaratım süreci, iki farklı veri üretme sürecine bağlı olarak oluşturulmuş ve yöntemlerin sadece sınıflandırma performansları CCR ve AUC metrikleri kullanılarak değerlendirilmiştir. Sayısal analizler, 12. Nesil Intel(R) Core (TM) i5-12450H, 2.00 GHz işlemci ve 16 GB RAM içeren Huawei D16 (2022) model bilgisayarda ve performans modu aktifken gerçekleştirilmiştir. Bu tezde bahsi geçen tüm yöntemlerle ilgili örnek bir R koduna, bu adresten ulaşılabilir: (https://github.com/berkayakturk/F_Logit_R).

5.1. Monte-Carlo Denemeleri

Monte Carlo denemeleri boyunca, iki farklı veri oluşturma süreci Süreç 1 ve Süreç 2 olarak ele alınmaktadır. Süreçlerden elde edilen sonuçlar ise bulgular (5.2.) bölümünde sunulmaktadır.

5.1.1. Süreç 1

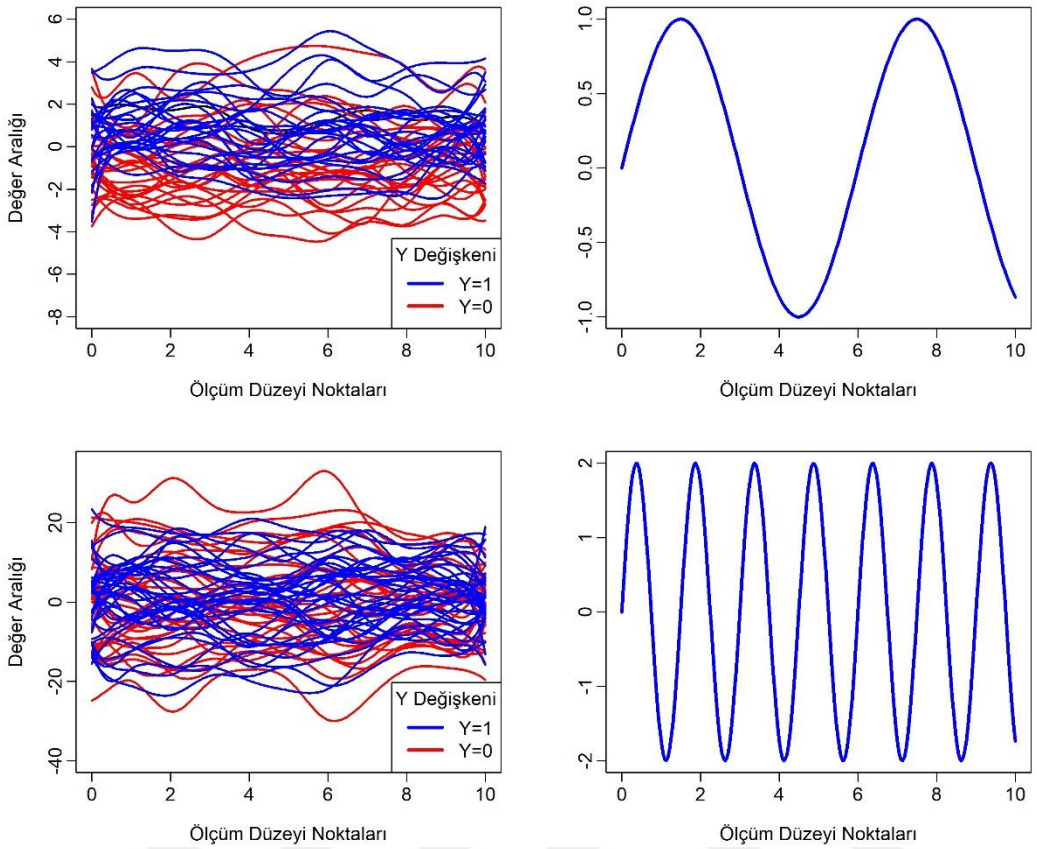
İlk süreçte, (Mousavi ve Sørensen, 2018) ve (Mutis ve ark., 2022) tarafından kullanılan veri yaratım sürecini ele alıyoruz. Bu veri yaratım sürecinde ilk olarak fonksiyonel açıklayıcı değişkenin elemanları $\mathcal{T} = [1,10]$ aralığında 256 eşit parçada oluşturulmuştur. Bu amaçla, 13 adet kübik B-spline temel genişletme fonksiyonu $\phi_k(t)_{k=1}^{13}$, dokuz eşit parçada $[0,10]$ aralığında tanımlanmıştır. Temel fonksiyon katsayıları ise rastgele olarak $C = ZU$ formunda oluşturulmuştur. Burada, Z adlı matris 1000×13 boyutunda olup, elemanları standart normal dağılımdan elde edilmiştir. U ise, 13×13 boyutlu rastgele bir matristir ve elemanları $[0,1]$ aralığında tekdüze dağılımdan üretilmiştir. Bu matrislere bağlı olarak

fonksiyonel açıklayıcı değişkenin değerleri $\mathcal{X}_i(t_j) = \sum_{k=1}^{13} c_{ik}\phi_k(t_j)$ $(j \in [0,10])$ olarak elde edilmiştir. Regresyon parametre fonksiyonu $\beta(t) = \sin(t\pi/3)$ olarak seçilmiş ve ikili yanıt değişkeninin değerleri aşağıdaki gibi açıklayıcı değişkenin ve parametre fonksiyonunun doğrusal bir bileşeni olarak elde edilmiştir:

$$l_i = \int_0^{10} \mathcal{X}_i(t)\beta(t)dt, \quad i = 1, \dots, 1000. \tag{5.1}$$

Yukarıda belirtildiği gibi, bu süreç için iki senaryo düşünülmüştür. İlk senaryoda, veriler düzgün bir şekilde bahsi geçen veri yaratım sürecinden elde edilmiş ve veri setinde herhangi bir aykırı gözlem bulunmamaktadır. Bu durum, önerilen yöntemin kestirim ve sınıflandırma doğruluğunu göstermek amacıyla kullanılmıştır.

İkinci senaryoda ise, eğitim seti olarak ayrılan 700 gözlemden rastgele seçilen $[\%1, \%5, \%10]$'luk gözlemler, yukarıda bahsi geçen veri yaratım sürecinde U'nun $[0,5]$ aralığında tekdüze dağılımdan elde edildiği ve parametre fonksiyonunun $\beta(t) = 2\sin(4t\pi/3)$ olarak belirlendiği şekilde yaratılan aykırı gözlemler ile değiştirilmiştir. Ayrıca, aykırı gözlem olarak değiştirilen açıklayıcı gözlem değerlerine karşılık gelen ikili yanıt değişkeninin değerleri tam tersi sınıf olarak değiştirilmiştir. Böylece, bu senaryo altında yaratılan veri setinde hem açıklayıcı değişkene hem de yanıt değişkenine aykırı gözlemler eklenmiştir. Bu senaryo altında, önerilen yöntemin klasik dirençli olmayan yöntemlere göre daha iyi kestirim ve sınıflandırma performansı sergilemesi beklenmektedir. Aynı zamanda, mevcut dirençli yönteme kıyasla benzer sonuçları daha az hesaplama süresi ile üretmesi öngörülmektedir. Süreç 1 için yaratılan örnek veri setlerinin temsili görselleri Şekil 5.1'de sunulmaktadır.



Şekil 5.1 Süreç 1 altında yaratılan verilerin temsili görselleri. İlk senaryoda yaratılan gözlemler 1. satırda, ikinci senaryoda yaratılan gözlemler 2. satırda gösterilmektedir. Sol paneldeki görseller, açıklayıcı değişkenin elemanlarını gösterirken sağ paneldeki görseller seçilen regresyon parametre fonksiyonunu göstermektedir.

5.1.2. Süreç 2

İkinci süreçte ise, (Ferraty ve Vieu, 2006) ve (Escabias ve ark., 2007) tarafından kullanılan ve sınıflandırma amacıyla önerilen veri yaratım sürecini ele alıyoruz. Bu süreçte, $\mathcal{Y} = 0$ ve $\mathcal{Y} = 1$ sınıfları için 500'er toplamda 1000 gözlem [1,21] aralığında 101 eşit noktada yanıt değişkeni oluşturulmuştur. Fonksiyonel açıklayıcı değişkenin gözlemleri aşağıdaki süreç yardımıyla elde edilmiştir:

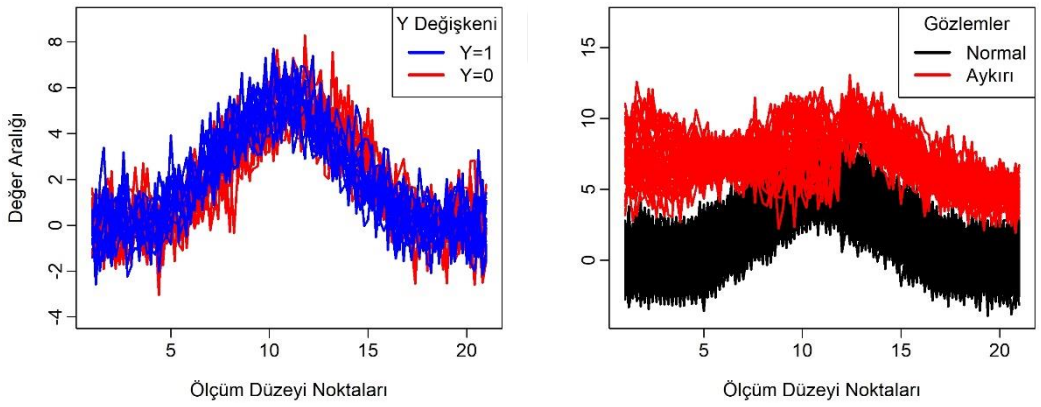
$$\mathcal{X}(t) = \begin{cases} uh_1(t) + (1-u)h_2(t) + \varepsilon(t), & \mathcal{Y} = 0 \\ uh_1(t) + (1-u)h_3(t) + \varepsilon(t), & \mathcal{Y} = 1 \end{cases} \tag{5.2}$$

Burada u, $[0,1]$ aralığında tekdüze dağılımdan rastgele yaratılan değişkeni gösterirken, $\varepsilon(t)$ ise standart normal dağılımdan yaratılan değişkeni göstermektedir. $h_1(t) = max\{6 - |t - 11|,0\}$, $h_2(t) = h_1(t-4)$ ve $h_3(t) = h_1(t+4)$ olarak belirlenmiştir.

İkinci süreçteki aykırı gözlemlerin oluşturulması, ilk süreçteki aykırı gözlem yaratım sürecine benzer bir yöntemle gerçekleştirilir. Eğitim setindeki fonksiyonel açıklayıcı değişkenlerden rastgele seçilen $[\%1, \%5, \%10]$'luk gözlemler aşağıdaki veri yaratım süreci ile yer değiştirilir:

$$\mathcal{X}(t) = \begin{cases} uh_4(t) + (1-u)h_2(t) + \varepsilon^*(t), & \mathcal{Y} = 0 \\ uh_4(t) + (1-u)h_3(t) + \varepsilon^*(t), & \mathcal{Y} = 1 \end{cases} \tag{5.3}$$

Burada $\varepsilon^*(t) \sim \mathcal{N}(5,10)$ dağılımından üretilmiş olup, $h_4(t) = h_1(t+50)$ şeklinde seçilmiştir. Açıklayıcı değişkendeki aykırı gözlemlere karşılık gelen ikili yanıt değişkeninin değerleri ters sınıfa değiştirilerek hem yanıt hem de açıklayıcı değişkenlerde aykırı gözlemlerin mevcut olduğu durum sağlanmıştır. Süreç 2 için yaratılan örnek veri setlerinin temsili görselleri Şekil 5.2'de sunulmaktadır.

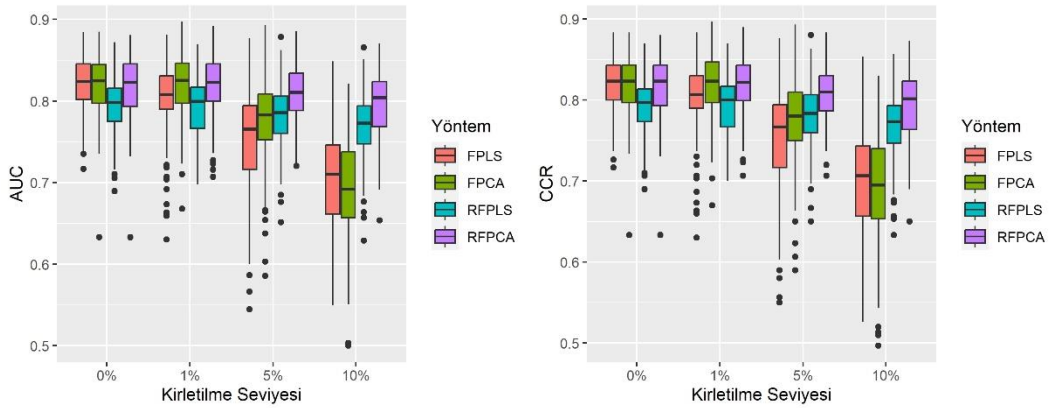


Şekil 5.2 Süreç 2 altında yaratılan verilerin temsili görselleri. İlk senaryoda yaratılan gözlemler sol panelde sunulmuşken ikinci senaryoda yaratılan gözlemler sağ panelde sunulmuştur.

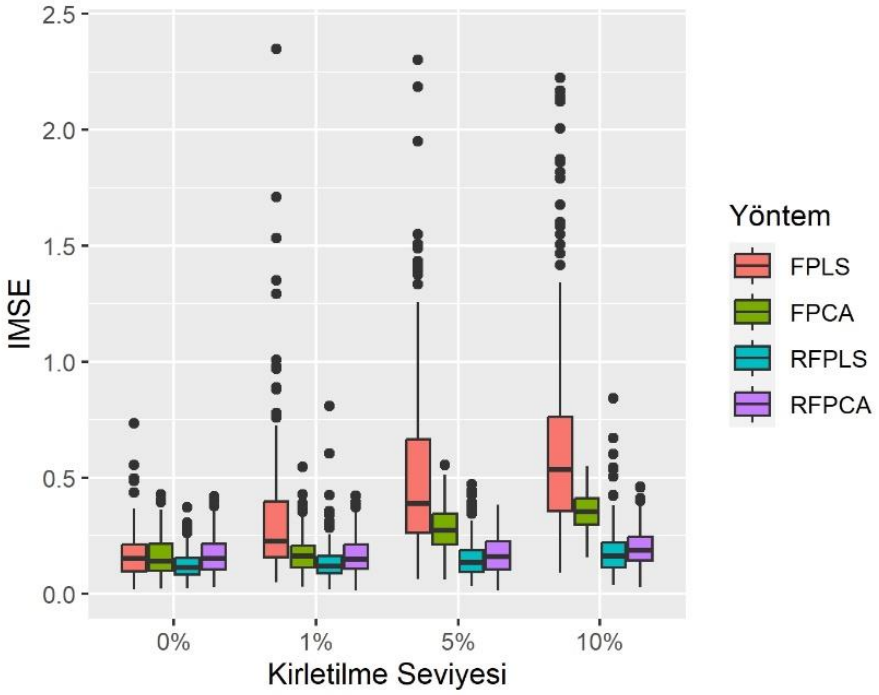
5.2. Bulgular

İlk süreçten elde edilen sonuçlar Şekil 5.3 ve Şekil 5.4'te sunulmuştur. İkinci süreçten elde edilen sonuçlar ise Şekil 5.6'da sunulmuştur. Şekil 5.3 incelendiğinde, veri setinde aykırı gözlem bulunmadığı durumda tüm yöntemler benzer ve yüksek doğru sınıflandırma (AUC ve CCR) performansına sahiptir. Diğer taraftan, aykırı gözlemlerin varlığı durumunda klasik FPCA ve FPLS yöntemleri zayıf doğru sınıflandırma sonuçları üretirken RFPLS ve önerilen RFPCA yüksek doğru sınıflandırma değerleri üretmeye devam etmişlerdir. Klasik ve dirençli yöntemlerin performansları arasındaki fark aykırı gözlem oranı arttıkça daha belirgin olarak gözlemlenmektedir. RFPLS yöntemine kıyasla, önerilen RFPCA yöntemi hem normal süreçte hem de aykırı gözlem durumunda daha yüksek doğru sınıflandırma sonuçları üretmeyi başarmıştır.

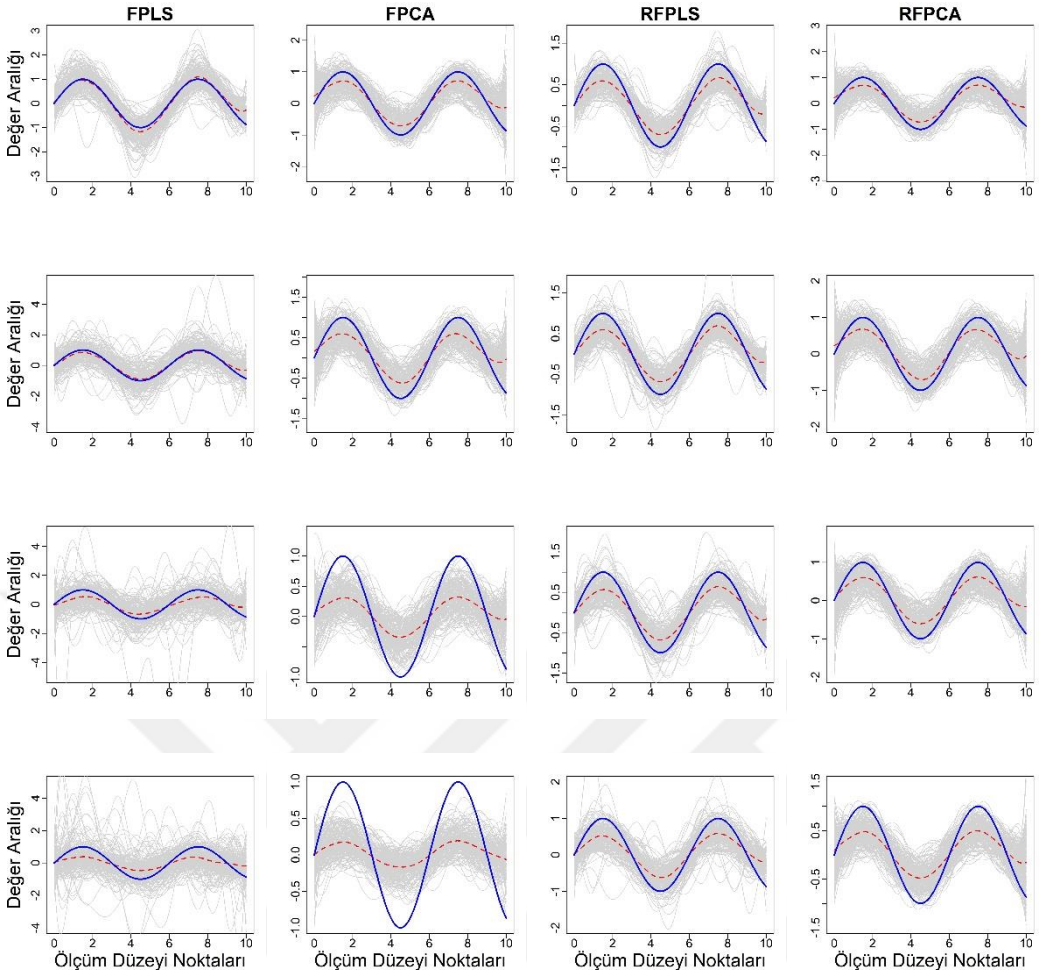
Şekil 5.4'te sunulan IMSE değerleri incelendiğinde ise, veri setinde aykırı gözlem olmadığı durumda tüm yöntemlerin benzer IMSE değerleri dolayısı ile benzer parametre tahminleri ürettikleri gözlemlenmiştir. Yaratılan veri seti aykırı gözlemler tarafından kirletildiği durumda ise, klasik yöntemler (özellikle FPLS) ciddi oranda aykırı gözlemlerden etkilenmekte ve yüksek IMSE değerleri üretmektedir. Diğer taraftan RFPLS ve RFPCA yöntemleri, aykırı gözlemlerin etkilerini sınırlandırarak aykırı gözlem olmayan duruma yakın IMSE değerleri üretmektedirler. RFPLS ve RFPCA yöntemleri kıyaslandığında, RFPLS yöntemi özellikle yüksek kirlilik seviyeleri için bazı durumlarda yüksek IMSE değerleri üretirken, önerilen yöntem neredeyse tüm denemelerde benzer ve kararlı IMSE değerleri üretmektedir. Bu nedenle önerilen yöntem, aykırı değerlerin varlığında klasik yöntemlere göre daha başarılı parametre tahminleri sağlarken, mevcut dirençli yönteme göre daha kararlı parametre tahminleri sunmaktadır. Tüm yöntemler için 200 Monte-Carlo denemesinden elde edilen parametre tahminleri, bu tahminlerin ortalamaları ve gerçek parametre fonksiyonunu gösteren bir görseli Şekil 5.5'te sunulmaktadır. Bu görseldeki sonuçlar, Şekil 5.4'te sunulan bulguları desteklemekte ve neredeyse her durum altında önerilen yöntemin gerçek parametre fonksiyonuna benzer ve kararlı sonuçlar ürettiği görülmektedir.



Şekil 5.3 FPLS, FPCA, RFPLS ve RFPCA yöntemleri için süreç 1 altında hesaplanan AUC (sol panel) ve CCR (sağ panel) değerlerinin kutu grafikleri.



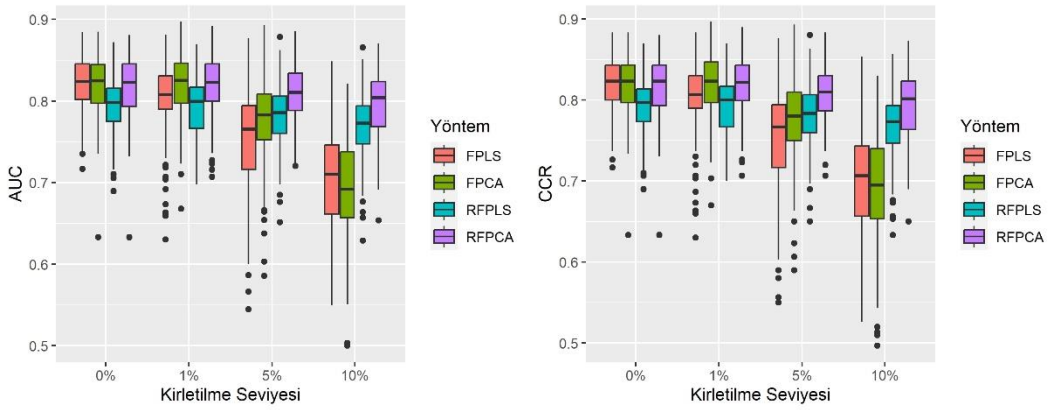
Şekil 5.4 FPLS, FPCA, RFPLS ve RFPCA yöntemleri için süreç 1 altında hesaplanan IMSE değerlerinin kutu grafikleri.



Şekil 5.5 Süreç 1 altında 200 Monte Carlo deneyinden elde edilen tahmini parametre fonksiyonları (gri çizgiler), ortalamaları (kırmızı çizgiler) ve gerçek parametre fonksiyonlarının (mavi çizgiler) temsili görseli. 1., 2., 3. ve 4. satırdaki görseller sırasıyla veride aykırı gözlemin bulunmadığı, verinin %1 oranda kirletildiği, verinin %5 oranda kirletildiği ve verinin %10 oranda kirletildiği durumları göstermektedir.

İkinci süreç altında elde edilen AUC ve CCR değerlerinin sonuçları Şekil 5.6'da sunulmaktadır. Bu şekil incelendiğinde, aykırı gözlemlerin veri setinde bulunmadığı durumda tüm yöntemlerin benzer sınıflandırma performansına sahip oldukları gözlemlenmektedir. Diğer taraftan, yaratılan veri setleri aykırı gözlemler tarafından kirletildiğinde ise klasik yöntemlerin (özellikle FPLS) aykırı değerlerden ciddi oranda etilendiği ve düşük AUC ve CCR değerleri ürettikleri görülmektedir. RFPLS ve RFPCA yöntemleri ise aykırı gözlemlerden etkilenmeyerek hem aykırı gözlemlerin varlığı hem de normal veri yaratım süreci altında benzer ve yüksek doğru sınıflandırma değerleri

üretmektedir. Önerilen RFPCA yöntemi, RFPLS yöntemine kıyasla biraz daha yüksek doğru sınıflandırma değerleri sağlamıştır.



Şekil 5.6 FPLS, FPCA, RFPLS ve RFPCA yöntemleri için süreç 2 altında hesaplanan AUC (sol panel) ve CCR (sağ panel) değerlerinin kutu grafikleri.

Son olarak yöntemlerin hesaplama etkinliklerini bir dizi Monte-Carlo denemesi ile karşılaştırdık. Bu amaçla her iki veri yaratım süreci için iki farklı senaryo düşünüldü. İlk senaryoda, $n = 1000$ gözlem yaratılarak $K = [5,10,15,20,25,30]$ B-spline temel genişletme fonksiyonunun kullanılması durumunda tek bir Monte-Carlo denemesinden elde edilen hesaplama süreleri kaydedildi ve temel genişletme fonksiyon sayısının yöntemlere göre hesaplama performansları üzerindeki etkileri araştırıldı. İkinci senaryoda ise, $K = 15$ B-spline temel genişletme fonksiyonu ile $n = [100,300,500,1000,1500,2000]$ gözlem oluşturularak gözlem sayısının yöntemlere göre hesaplama performansları üzerindeki etkileri araştırıldı. İlk veri yaratım süreci altında elde edilen hesaplama süreleri Tablo 5.1 ve Tablo 5.2'de sunulmaktadır, ikinci süreç altında elde edilen hesaplama süreleri ise Tablo 5.3 ve Tablo 5.4'te sunulmuştur. Bu sonuçlar, yöntemlerin hesaplama etkinliği üzerinde temel genişletme fonksiyon sayısı ve gözlem sayısının etkilerini incelememize yardımcı olmuştur.

Tablo 5.1'de sunulan hesaplama süreleri incelendiğinde klasik FPLS ve FPCA yöntemleri, dirençli yöntemlere kıyasla çok daha az hesaplama sürelerine sahip oldukları gözlemlenmektedir. Bu tabloda gösterilen bulgular, düşünülen tüm K değerleri için önerilen RFPCA yönteminin 1 saniyenin altında sonuç üretirken, RFPLS yönteminin RFPCA'ya kıyasla 29 kat ile 151 kat arasında daha fazla hesaplama süresi gerektirdiğini göstermektedir.

Tablo 5.2'de sunulan sonuçlar incelendiğinde, farklı örneklem büyüklükleri için, klasik yöntemlerin dirençli yöntemlere kıyasla çok daha az hesaplama süreleri gerektirdiği ve önerilen RFPCA yönteminin ise RFPLS yöntemine kıyasla 20 kat ile 103 kat arasında değişen daha az hesaplama süreleri gerektirdiği gözlemlenmektedir. İkinci veri yaratım süreci altında elde edilen hesaplama süreleri de ilk veri yaratım süreci altında elde edilen hesaplama sürelerine benzerdir. Klasik yöntemler dirençli yöntemlere kıyasla çok daha az hesaplama süreleri gerektirmektedir. RFPLS'ye kıyasla önerilen RFPCA yöntemi farklı K değerleri için 12 kat ile 43 kat arasında değişen ve farklı örneklem büyüklükleri için ise 3 ile 32 kat arasında değişen daha az hesaplama süreleri gerektirmektedir.

Tablo 5.1 Süreç 1 için FPLS, FPCA, RFPLS ve RFPCA yöntemlerinin temel bileşen sayısına göre tek bir döngüyü hesaplama süreleri (saniye cinsinden). (Gözlem Sayısı = 1000).

Temel Bileşen Sayısı	FPLS	FPCA	RFPLS	RFPCA
5	0.08	0.02	14.62	0.50
10	0.17	0.04	18.39	0.39
15	0.14	0.03	25.16	0.32
20	0.41	0.03	41.49	0.50
25	0.56	0.05	66.51	0.44
30	0.73	0.07	92.10	0.61

Tablo 5.2 Süreç 1 için FPLS, FPCA, RFPLS ve RFPCA yöntemlerinin gözlem sayısına göre tek bir döngüyü hesaplama süreleri (saniye cinsinden). (Temel Bileşen Sayısı = 15).

Gözlem Sayısı	FPLS	FPCA	RFPLS	RFPCA
100	0.03	0.02	3.09	0.15
300	0.12	0.03	8.29	0.34
500	0.20	0.03	22.29	0.48
1000	0.15	0.03	35.12	0.34
1500	0.19	0.03	52.39	1.39
2000	0.36	0.05	97.20	1.47

Tablo 5.3 Süreç 2 için FPLS, FPCA, RFPLS ve RFPCA yöntemlerinin temel bileşen sayısına göre tek bir döngüyü hesaplama süreleri (saniye cinsinden). (Gözlem Sayısı = 1000).

Temel Bileşen Sayısı	FPLS	FPCA	RFPLS	RFPCA
5	0.05	0.02	6.31	0.53
10	0.10	0.03	9.51	0.56
15	0.22	0.04	17.36	1.28
20	0.25	0.03	25.45	0.84
25	0.28	0.05	29.70	1.41
30	0.41	0.04	33.58	0.78

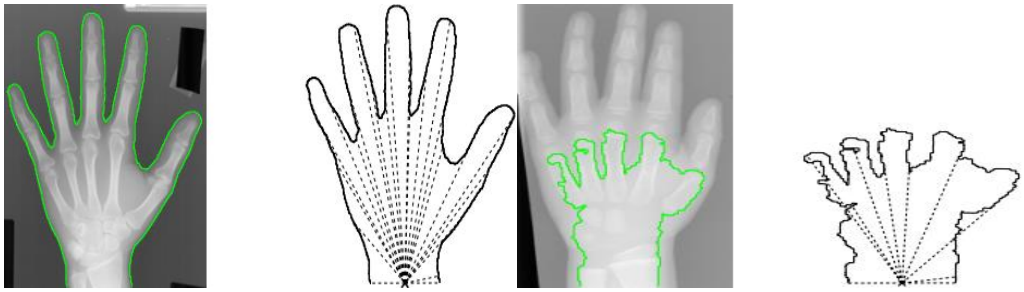
Tablo 5.4 Süreç 2 için FPLS, FPCA, RFPLS ve RFPCA yöntemlerinin gözlem sayısına göre tek bir döngüyü hesaplama süreleri (saniye cinsinden). (Temel Bileşen Sayısı = 15).

Gözlem Sayısı	FPLS	FPCA	RFPLS	RFPCA
100	0.05	0.02	2.14	0.58
300	0.09	0.03	5.20	0.22
500	0.23	0.03	7,17	0.68
1000	0.19	0.03	17.96	1.14
1500	0.27	0.04	32.87	1.18
2000	0.42	0.07	49.38	1.50

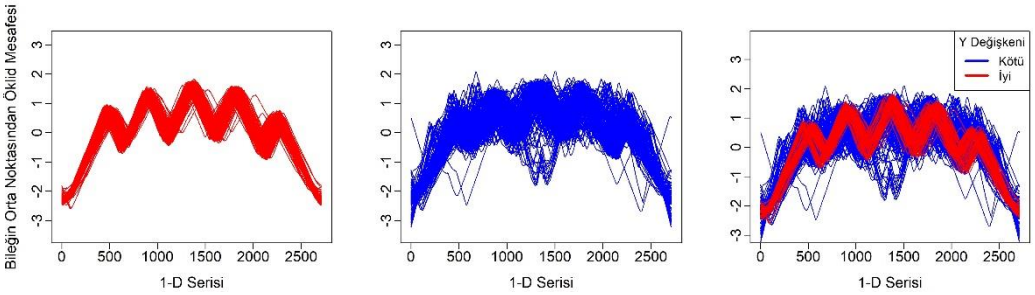
Sonuç olarak, Monte-Carlo denemelerinden elde edilen bulgular, önerilen RFPCA yönteminin klasik yöntemlere kıyasla aykırı gözlem olmadığı durumda benzer kestirim ve sınıflandırma performansı ürettiğini göstermektedir. Aykırı gözlem varlığında ise, önerilen yöntem daha iyi kestirim ve sınıflandırma performansı sunmaktadır. Bunlara ek olarak, bulgular önerilen yöntemin, mevcut RFPLS yöntemine kıyasla benzer veya daha iyi kestirim ve sınıflandırma sonuçlarını çok daha az hesaplama süreleri ile ürettiğini ortaya koymaktadır.

5.3. Gerçek Veri Analizi: El Radyografisi Verisi

Gerçek veri analizi olarak (Davis ve ark., 2012) tarafından kullanılan el çizim verisini düşünüyoruz. Veri seti https://www.timeseriesclassification.com adresinde mevcuttur. Bu veri seti, el radyografisi görüntülerini kullanarak kemik yaşının tahmin edilmesine yönelik yapılan kapsamlı bir projenin bir parçasıdır. Kemik yaşının belirlenmesi, baskın olmayan el kemiklerinin gelişimini ölçerek, bir hastanın röntgen görüntüsünden beklenen yaşını tahmin etmeyi içerir (Davis ve ark., 2012). Bu, bir çocuğun veya bireyin kemiklerinin belirli bir zamanda ne kadar geliştiğini değerlendirmek için kullanılır. Ancak bu sonuca ulaşmak için radyografik görüntülerin iyi çekilmiş olması gerekmektedir. Görüntülerden oluşan bu veri seti, röntgen filmlerinin yapısı bozulmadan sayısal hale getirilerek oluşturulmuştur. Bir el modeli, dynamic time warping (DTW) yoluyla 1-D serilerinden elde edilen sonuçlara göre oluşturulmuştur. 1-D serileri oluşturulurken noktasal Öklid mesafe matrisi kullanılarak, röntgenler sayısallaştırılmıştır. Ancak bir sonuca ulaşmak için radyografik görüntülerin iyi çekilmiş olması gerekmektedir. Bu veri setinde amaç, elde edilen görüntülerden, yaş tahminini başarılı bir şekilde yapmak için bir çizimin iyi olup olmadığını kontrol ederek sınıflandırmaktır. Elde edilen iyi ve kötü el modellerine ait örnek görsel Şekil 5.7'de ve el modellerine ait Öklid uzaklıklarına ait örnek bir görsel ise Şekil 5.8' de sunulmaktadır.



Şekil 5.7 El radyografisi için oluşturulan el modelleri. Birinci panel iyi el modelini gösterirken üçüncü panel kötü el modelini gösterir. İkinci ve dördüncü panel sırasıyla 1-D serisine dönüştürülen iyi el ve kötü el modelini temsil eder.

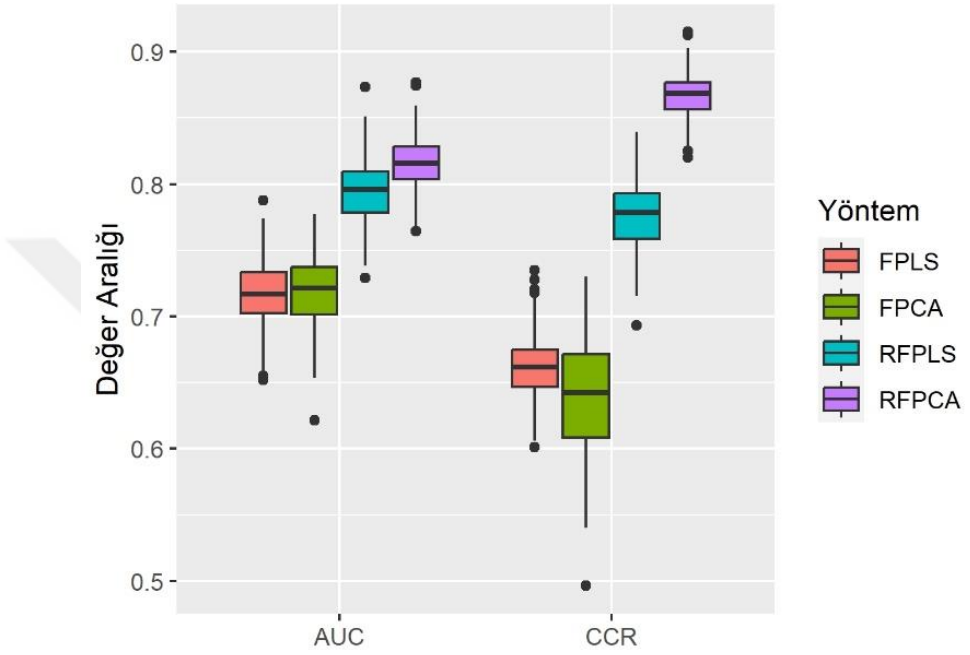


Şekil 5.8 El radyografisi için iyi el verileri (sol panel), kötü el verileri (orta panel) ve sınıflarına göre tespit edilen aykırı değerlerin (sağ panel) grafik gösterimi.

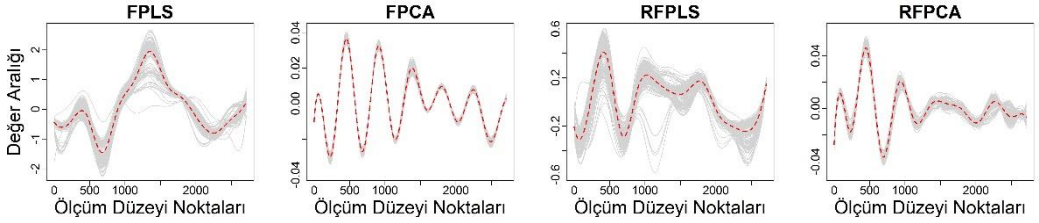
Veri setinde 495 kötü el modeli ve 975 iyi el modeli olmak üzere toplamda 1370 el modeline ait gözlem bulunmaktadir. Öklid uzunluk mesafeleri ise [0,2709] aralığında toplamda 2709 eşit noktada hesaplanmıştır. Veri setinde aykırı gözlemlerin varlığını belirlemek adına (Sun ve Genton, 2011) tarafından önerilen ve "fdaoutlier" R paketinde mevcut olan fonksiyonel kutu grafiği yöntemi kullanılmış ve sonuçlar veri setinin toplamda 309 (%22.5) aykırı gözlem olduğunu göstermiştir (Şekil 5.8'e bakınız). Bu durum, önerilen yöntemimizin bu veri setine uygulanması söz konusu olduğunda, klasik yöntemlere göre daha üstün sınıflandırma performansı sunarken, RFPLS yöntemiyle karşılaştırıldığında ise benzer ya da daha iyi sonuçlar elde etme potansiyeline sahip olduğunu ve bunu daha kısa sürede başarabileceğini bizlere gösterecektir. Bu nedenle, önerilen yöntemimizin kullanılması bizi motive edecektir.

Veri analizinde 200 Monte-Carlo denemesi gerçekleştirilmiş ve her bir denemede tüm gözlemlerin %70'i (959 gözlem) eğitim kümesi olarak ve %30'u (411) ise test kümesi olarak ikiye ayrılmıştır. Eğitim kümesi üzerinde modeller kurularak test kümesi üzerindeki ikili yanıt değişkeninin değerleri tahmin edilmiş ve AUC ve CCR değerleri hesaplanmıştır. Elde edilen sonuçlar Şekil 5.9'da sunulmaktadır. Elde edilen bulgular, RFPLS ve RFPCA yöntemlerinin klasik FPLS ve FPCA yöntemlerine göre daha iyi sınıflandırma performansı (daha yüksek AUC ve CCR değerleri) sergilediklerini göstermiştir. Ayrıca bulgularımız, RFPLS yöntemine kıyasla önerilen yöntemin daha iyi sınıflandırma performansına sahip olduğunu göstermektedir. 200 Monte-Carlo denemesinden elde edilen parametre fonksiyonlarının grafiği Şekil 5.10'da sunulmuştur. Grafiği incelediğimizde, FPCA tabanlı yöntemlerin, FPLS tabanlı yöntemlere göre daha benzer ve kararlı parametre fonksiyonları oluşturduğunu görmekteyiz. Ek olarak, tek bir Monte-Carlo denemesi üzerinden yöntemlerin

gerektirdiği hesaplama sürelerini kaydettik ve FPLS, FPCA, RFPLS ve RFPCA yöntemleri için sırasıyla 0.20, 0.24, 67.03 ve 0.91 saniye hesaplama sürelerini elde ettik. Sonuç olarak, Monte-Carlo denemeleriyle elde edilen bulgulara paralel olarak, önerilen yöntemin gerçek veri analizinde hem klasik hem de dirençli FPLS yöntemlerine göre daha iyi sınıflandırma ve kestirim performansı sergilediği görülmektedir.



Şekil 5.9 El radyografisi verileri için hesaplanan CCR ve AUC değerlerinin kutu grafiği.



Şekil 5.10 El radyografisi verileri için tahmin edilen parametre fonksiyonlarını (gri çizgiler) ve tahmin edilen parametre fonksiyonlarının ortalamasını (kırmızı çizgiler) temsil etmektedir. FPLS (birinci sütun), FPCA (ikinci sütun), RFPLS (üçüncü sütun) ve RFPCA (dördüncü sütun).

6. TARTIŞMA VE SONUÇ

Yanıt değişkeninin ikili değerler (0 ve 1) aldığı ve açıklayıcı değişkenin rastgele eğrilerden oluştuğu FlogitR modeli bağlamında, temel genişletme yöntemleri olan FPLS ve FPCA temelli algoritmalar üzerine çeşitli yaklaşımlar öne sürülmüştür. Mevcut yöntemler, verilerin düzgün bir süreçten kaynaklandığı durumlarda tatmin edici sınıflandırma performansı sergilerken, veri setindeki aykırı gözlemler nedeniyle yanlı parametre tahminleri sonucu yüksek yanlış sınıflandırma olasılıkları ortaya çıkmaktadır. Literatürdeki dirençli yöntemlerden biri olan RFPLS, aykırı gözlemlerin varlığında klasik yöntemlere göre daha iyi sınıflandırma performansı sunsa da, hesaplama süresi açısından verimli bir yaklaşım değildir. Bu tezde, aykırı gözlemler söz konusu olduğunda bu gözlemlerden etkilenmeksizin yüksek doğruluklu sınıflandırma performansı sunan ve hesaplama süresi açısından etkin olan RFPCA yöntemini önermekteyiz.

Önerilen yöntemin kestirim ve sınıflandırma performansı, çeşitli Monte Carlo denemeleri ve bir gerçek veri analizi ile incelendi. Önerilen yöntemle elde edilen sonuçlar, mevcut dirençli olmayan FPLS ve FPCA yöntemleri ile dirençli FPLS yöntemi sonuçları ile mukayeseli olarak değerlendirildi. Elde edilen bulgular, önerilen yöntemin klasik yöntemlerle kıyaslandığında, veri setinde aykırı gözlemler bulunmadığı durumlarda benzer kestirim ve sınıflandırma performansı sergilediğini, aykırı gözlemlerin mevcut olduğu durumlarda ise daha üstün kestirim ve sınıflandırma performansına sahip olduğunu ortaya koydu. Ayrıca, bulgularımız önerilen yöntemin, mevcut RFPLS yöntemine kıyasla benzer hatta daha iyi sonuçları çok daha kısa hesaplama süreleri ile elde ettiğini gösterdi. Bu durum, bilim insanları, araştırmacılar ve veri bilimcileri için zaman faktörünün kritik öneme sahip olduğu göz önünde bulundurulduğunda, veri analizi süreçlerinde daha pratik ve hızlı sonuçlar elde etmek adına önemli bir avantaj sağlamaktadır.

Önerilen yöntemimiz, gelecekteki çalışmalar için çeşitli açılardan geliştirilebilir. Örneğin, mevcut tezde, yalnızca tek bir fonksiyonel açıklayıcı değişken düşünülmüştür. Bununla birlikte, ikili yanıt değişkeninin sınıflandırılmasında birden fazla açıklayıcı değişken kullanılarak daha iyi sonuçlar elde edilebilir. Bu bağlamda, önerilen yöntemimiz, modelde birden fazla açıklayıcı değişkenin bulunduğu daha genel bir modele uyarlanabilir. Ayrıca, FlogitR modelinde sadece ikili yanıt değişkeni ele alınmaktadır. Önerilen RFPCA yöntemi, hedef değişkenin ikiden fazla sınıfa sahip olduğu durumlarda da kullanılabilir hale

getirilerek daha geniş bir yelpazede sınıflandırma problemleri için uygulanabilir. Mevcut çalışmada modelde yalnızca fonksiyonel açıklayıcı değişken kullanılmıştır. Ancak, birçok bilim alanında ikili yanıt değişkeni skaler açıklayıcı değişkenlere bağlı olabilir. Sağlık alanında kullanılan ağırlık, boy, yaş gibi ölçütler veya finans alanında kullanılan döviz kuru, hisse senedi değeri, faiz oranı gibi skaler değişkenlere bağlı olarak model, hem fonksiyonel hem de skaler açıklayıcı değişkenlerin bulunduğu kısmi FlogitR modeline genişletilebilir. Son olarak, fonksiyonel regresyon modellerinde, cezalandırma terimi ile fonksiyonel değişkenlerin düzgünlüğü kontrol edilerek modelin tahmin performansı artırılabilir. Bu bağlamda, önerilen yöntemimizin cezalandırılmış bir versiyonu geliştirilerek daha iyi sınıflandırma ve tahmin performansı sağlanabilir. Bu tür geliştirmeler, önerilen yöntemin uygulanabilirliğini ve etkinliğini artırarak, bilim insanları, araştırmacılar ve veri bilimcileri için daha geniş bir kullanım alanı sunacaktır.

7. KAYNAKLAR

Aguilera, A. M., Escabias, M. and Valderrama, M. J. (2008), 'Discussion of different logistic models with functional data. application to systemic lupus erythematosus', Computational Statistics & Data Analysis 53(1), 151–163.

Alqallaf, F. and Agostinelli, C. (2016), 'Robust inference in generalized linear models', Communica- tions in Statistics - Simulation and Computation 45(9), 3053–3073.

Araki, Y., Konishi, S., Kawano, S. and Matsui, H. (2009), 'Functional logistic discrimination via regularized basis expansions', Communications in Statistics—Theory and Methods 38(16-17), 2944– 2957.

Bali, J. L., Boente, G., Tyler, D. E. and Wang, J.-L. (2011), 'Robust functional principal components: A projection-pursuit approach', The Annals of Statistics 39(6), 2852 – 2882.

Beyaztas, U. and Shang, H. L. (2020), 'On function-on-function regression: Partial least squares approach', Environmental and Ecological Statistics 27(1), 95–114.

Beyaztas, U. and Shang, H. L. (2022), 'A comparison of parameter estimation in function-on- function regression', Communications in Statistics-Simulation and Computation 51(8), 4607–4637.

Bianco, A. M. and Yohai, V. J. (1996), Robust estimation in the logistic regression model, Vol. 109, Springer.

Conway, J. B. (2019), A course in functional analysis, Vol. 96, Springer.

Croux, C. and Haesbroeck, G. (2003), 'Implementing the bianco and yohai estimator for logistic regression', Computational statistics & data analysis 44(1-2), 273–295.

Croux, C. and Ruiz-Gazen, A. (1996), 'High breakdown estimators for principal components: the projection-pursuit approach revisited', Journal of Multivariate Analysis 95(1), 206–226.

Davis, L. M., Theobald, B.-J., Lines, J., Toms, A. and Bagnall, A. (2012), 'On the segmentation and classification of hand radiographs', International journal of neural systems 22(05), 1250020.

De Boor, C. (2001), 'Calculation of the smoothing spline with weighted roughness measure', Mathematical Models and Methods in Applied Sciences 11(01), 33–41.

Denhere, M. and Billor, N. (2016), 'Robust principal component functional logistic regression', Communications in Statistics-Simulation and Computation 45(1), 264–281.

Escabias, M., Aguilera, A. M. and Valderrama, M. J. (2007), 'Functional pls logit regression model', Computational Statistics & Data Analysis 51(10), 4891–4902.

Escabias, M., Aguilera, A. M. and Valderrama, M. J. (2004), 'Principal component estimation of functional logistic regression: discussion of two different approaches', Journal of Nonparametric Statistics 16(3-4), 365–384.

Escabias, M., Aguilera, A. and Valderrama, M. (2005), 'Modeling environmental data by functional principal component logistic regression', Environmetrics: The official journal of the International Environmetrics Society 16(1), 95–107.

Ferraty, F. and Vieu, P. (2006), Nonparametric functional data analysis: theory and practice, Vol. 76, Springer; New York.

Jolliffe, I. T. (2002), Principal component analysis for special types of data, Springer.

Karhunen, K. (1946), 'Zur spektraltheorie stochastischer prozesse', Ann. Acad. Sci. Fennicae, AI 34.

Kleffe, J. (1973), 'Principal components of random variables with values in a seperable hilbert space', Mathematische Operationsforschung und Statistik 4(5), 391–406.

Leng, X. and Mu¨ ller, H.-G. (2006), 'Classification using functional data analysis for temporal gene expression data', Bioinformatics 22(1), 68–76.

Loève, M. (1946). ‘Fonctions aléatoires à décomposition orthogonale exponentielle’. La Revue Scientifique, 84, 159-162.

Matsui, H. and Konishi, S. (2009), ‘Regularized functional regression modeling for functional response and predictors’, Journal of Math-for-Industry 53, 17–25.

Mousavi, S. N. and Sørensen, H. (2017), ‘Multinomial functional regression with wavelets and lasso penalization’, Econometrics and statistics 1, 150–166.

Mousavi, S. N. and Sørensen, H. (2018), ‘Functional logistic regression: a comparison of three methods’, Journal of Statistical Computation and Simulation 88(2), 250–268.

Mutis, M., Beyaztas, U., Simsek, G. G. and Shang, H. L. (2022), ‘A robust scalar-on-function logistic regression for classification’, Communications in Statistics-Theory and Methods pp. 1–17.

Preda, C., Saporta, G. and Le´ve´der, C. (2007), ‘Pls classification of functional data’, Computational Statistics 22(2), 223–235.

Ramsay, J., Hooker, G., Graves, S., Ramsay, J., Hooker, G. and Graves, S. (2009), ‘Introduction to functional data analysis’, Functional data analysis with R and MATLAB pp. 1–19.

Ratcliffe, S. J., Heller, G. Z. and Leader, L. R. (2002), ‘Functional data analysis with application to periodically stimulated foetal heart rate data. ii: Functional logistic regression’, Statistics in medicine 21(8), 1115–1127.

Sun, Y. and Genton, M. G. (2011), ‘Functional boxplots’, Journal of Computational and Graphical Statistics 20(2), 316–334.

Wei, P., Tang, H. and Li, D. (2014), ‘Functional logistic regression approach to detecting gene by longitudinal environmental exposure interaction in a case-control study’, Genetic epidemiology 38(7), 638–651.

ÖZGEÇMİŞ

Adı ve Soyadı: Berkay AKTÜRK

Bildiği Yabancı Diller: İngilizce

Eğitim:

Derece	**Bölüm/Program**	**Üniversite**
Lisans	İstatistik	Mimar Sinan Güzel Sanatlar Üniversitesi
Y. Lisans	İstatistik	Marmara Üniversitesi

Projelerde Yaptığı Görevler:

Tübitak\Bursiyer.