PLSR ve PCR Tekniklerinin Monte Carlo Simülasyonu ile Karşılaştırılması

Gamze Güven

YÜKSEK LİSANS TEZİ

İstatistik Anabilim Dalı

Aralık 2015

Comparison of PLSR and PCR Techniques with Monte Carlo Simülation

Gamze Güven

MASTER OF SCIENCE THESIS

Department of Statistics

December 2015

PLSR ve PCR Tekniklerinin Monte Carlo Simülasyonu ile Karşılaştırılması

Gamze Güven

Eskişehir Osmangazi Üniversitesi
Fen Bilimleri Enstitüsü
Lisansüstü Yönetmeliği Uyarınca
İstatistik Anabilim Dalı
Uygulamalı İstatistik Bilim Dalında
YÜKSEK LİSANS TEZİ
Olarak Hazırlanmıştır

Danışman: Doç. Dr. Hatice ŞAMKAR

Aralık 2015

ONAY

İstatistik Anabilim Dalı Yüksek Lisans öğrencisi Gamze Güven'in YÜKSEK LİSANS tezi olarak hazırladığı "PLSR ve PCR Tekniklerinin Monte Carlo Simülasyonu ile Karşılaştırılması" başlıklı bu çalışma, jürimizce lisansüstü yönetmeliğin ilgili maddeleri uyarınca değerlendirilerek kabul edilmiştir.

Danışman **:** Doç. Dr. Hatice Şamkar

İkinci Danışman **:** ----------

Yüksek Lisans Tez Savunma Jürisi:

Üye : Doç. Dr. Hatice ŞAMKAR

Üye : Prof. Dr. Zeki YILDIZ

Üye : Prof. Dr. Berna YAZICI

Fen Bilimleri Enstitüsü Yönetim Kurulu'nun ..... tarih ve ..... sayılı kararıyla onaylanmıştır.

Prof. Dr. Hürriyet ERŞAHAN
Enstitü Müdürü

ETİK BEYAN

Eskişehir Osmangazi Üniversitesi Fen Bilimleri Enstitüsü tez yazım kılavuzuna göre, Doç. Dr. Hatice Şamkar danışmanlığında hazırlamış olduğum "PLSR ve PCR Tekniklerinin Monte Carlo Simülasyonu ile Karşılaştırılması" başlıklı YÜKSEK LİSANS tezimin özgün bir çalışma olduğunu; tez çalışmamın tüm aşamalarında bilimsel etik ilke ve kurallara uygun davrandığımı; tezimde verdiğim bilgileri, verileri akademik ve bilimsel etik ilke ve kurallara uygun olarak elde ettiğimi; tez çalışmamda yararlandığım eserlerin tümüne atıf yaptığımı ve kaynak gösterdiğimi ve bilgi, belge ve sonuçları bilimsel etik ilke ve kurallara göre sunduğumu beyan ederim. 14/12/2015

Gamze Güven

İmza

ÖZET

Çoklu doğrusal regresyon, bir bağımlı değişken ile bir ya da daha fazla bağımsız değişken arasındaki ilişkiyi modellemek için yaygın olarak kullanılan istatistiksel bir yöntemdir. Bu yöntemde regresyon katsayılarını tahmin etmek için En Küçük Kareler (Least Squares-LS) tekniği kullanılır. Ancak çoklu doğrusal regresyonda güvenilir sonuçlar elde etmek için LS tekniğinin belli başlı varsayımlarının sağlanması gerekir. Bu varsayımlardan bir tanesi bağımsız değişkenler arasında ilişki bulunmaması gerektiğidir. Bağımsız değişkenler arasındaki ilişki, çoklu bağlantı sorununa sebep olur. Çoklu bağlantı sorunu, parametre tahminleri üzerinde olumsuz sonuçlar doğurur. Bu sorunu ortadan kaldırmak için çeşitli yollara başvurulabilir. Bu yollardan en yaygın olarak kullanılanı yanlı tahmin teknikleridir.

Bu tezde veri indirgemesi yaparak çoklu bağlantıyı ortadan kaldıran yanlı tahmin tekniklerinden Kısmi En Küçük Kareler Regresyonu (Partial Least Squares Regression – PLSR) ve Temel Bileşenler Regresyonu (Principal Component Regression – PCR) ele alınmıştır. Bu iki teknik farklı çoklu bağlantı dereceleri, farklı gözlem sayıları, farklı değişken sayıları ve farklı standart sapma değerleri için Çapraz Geçerliliğin Hata Kareler Ortalamasının Karekökü (Root Mean Square Error Cross Validation – RMSECV) kriterine göre ve bileşen sayısına göre Monte Carlo simülasyonu ile karşılaştırılmıştır. Benzer RMSECV değerlerini veren bileşen sayıları açısından simülasyon sonuçları değerlendirildiğinde bazı durumlarda PLSR tekniği PCR'den daha iyi performans gösterirken, bazı durumlarda her iki teknik benzer sonuçlar vermiştir.

Anahtar Sözcükler: Çoklu Bağlantı, PLSR, PCR, RMSECV, Bileşen

SUMMARY

Multiple Lineer Regression is a statistical method commonly used to model relationships between the dependent variable and one or more independent variables. In this method, Least Squares (LS) technique is used for prediction of regression coefficients. However, to obtain reliable results in multiple linear regression, some assumptions need to be provided for the LS method. One of the these assumptions is that there is no relationship between independent variables. The relationship between independent variables lead to multicollinearity problem. Multicollinearity problem results in negative consequences on parameter estimations. Various ways can be applied to remove this problem. One of the most widely used ways is biased estimation techniques.

In this thesis, biased estimation techniques called Partial Least Squares Regression (PLSR) and Principal Component Regression (PCR) that remove multicollinearty problem by reducing were discussed. These two techniques were compared with Monte Carlo simulation study with respect to RMSECV criterion and number of component in the case of different multicollinearity degrees, different number of observations, different number of variables and different number of standart deviation values. When the simülation results are evaluated in terms of component numbers which give similar RMSECV values although in some cases PLSR technique shows better performance than PCR, in some cases both techniques give similar results.

Key Words: Multicollinearity, PLSR, PCR, RMSECV, Component

TEŞEKKÜR

Tez çalışmam boyunca bana hem bilimsel alanda yol gösteren hem psikolojik anlamda desteğini esirgemeyen, her türlü imkanı sağlayarak beni motive eden saygıdeğer danışmanım ve hocam Doç. Dr. Hatice ŞAMKAR'a

Yüksek lisansım boyunca bana destek olan ve her türlü problemde yardımlarını esirgemeyen başta çok değerli bölüm başkanı hocam sayın Prof. Dr. Zeki YILDIZ'a ve çok değerli istatistik bölümü hocalarıma

Her danıştığımda yüksünmeden cevap veren ve yardımcı olan, dostluğunu ve kardeşliğini her anlamda hissettiren Mehmet SANDAL'a

Hayatımın her anında bana destek olan, her üzüldüğümde ve bunaldığımda beni toparlayıp ayağa kaldıran, gece gündüz benimle çabalayan, bütün hayatını bana adayan, gerektiğinde arkadaş gibi, gerektiğinde kardeş gibi davranan ve hep yanımda olan canım anneme sonsuz şükranlarımı sunar, teşekkürü borç bilirim.

İÇİNDEKİLER

Sayfa

ÇİZELGELER DİZİNİ

Çizelge **Sayfa**

SİMGELER VE KISALTMALAR DİZİNİ

Simgeler	**Açıklama**
λ	Özdeğer
λ_{max}	En büyük özdeğer
λ_{min}	En küçük özdeğer

Kısaltmalar	**Açıklama**
CV	Çapraz Geçerlilik
LOOCV	Birini Dışarda Bırakarak Çapraz Geçerlilik
LS	En Küçük Kareler
MSE	Hata Kareler Ortalaması
NIPALS	Doğrusal Olmayan Yinelemeli En Küçük Kareler
PCR	Temel Bileşenler Regresyonu
PLSR	Kısmi En Küçük Kareler Regresyonu
PRESS	Kestirim Hata Kareler Toplamı
RMSE	Hata Kareler Ortalamasının Karekökü
RMSEC	Ayarlamanın Hata Kareler Ortalamasının Karekökü
RMSECV	Çapraz Geçerliliğin Hata Kareler Ortalamasının Karekökü
RMSEP	Kestirimin Hata Kareler Ortalamasının Karekökü
SIMPLS	PLS yönteminin İstatistiksel olarak Esinlenilmiş Değişikliğinin Basit bir uygulaması
SVD	Tekil Değer Ayrışımı
VIF	Varyans Şişirme Çarpanı

1. GİRİŞ VE AMAÇ

Uzun yıllardan beri yapılan çalışmalarda değişkenler arasındaki fonksiyonel ilişkiyi tanımlayabilmek büyük bir önem kazanmıştır. Genel olarak bu ilişkiyi tanımlamak için en sık kullanılan yöntem regresyon analizidir. Regresyon analizi modern uygulamalı istatistikte, tıptan sosyal bilimlere kadar birçok alanda kullanılmaktadır. Regresyon analizinde, bağımlı değişken y ile bağımsız değişkenler $x_1, x_2, ..., x_p$ arasındaki ilişkiyi fonksiyonel olarak ifade etmek ve bu ilişkiyi bir modelle tanımlayabilmek amaçlanmaktadır (Ryan, 2008). Diğer bir ifadeyle bağımlı değişken üzerinde hangi bağımsız değişken veya değişkenlerin etkili olduğunu belirlemek ve bağımsız değişkenlerin değerlerini kullanarak mümkün olan en iyi şekilde bağımlı değişkenin değerini tahmin etmek amaçlanmaktadır (Kaşko, 2007).

n gözlem ve p tane bağımsız değişken içeren çoklu doğrusal regresyonun matematiksel modeli,

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_p x_{pi} + \varepsilon_i \ , \qquad i = 1, 2, ..., n \qquad (1.1)$$

olarak veya matris notasyonuyla

$$y = X\beta + \varepsilon \ , \qquad (1.2)$$

eşitliği ile ifade edilir. Burada,

y : $(nx1)$ boyutlu bağımlı değişken vektörü;

X : (nxk) boyutlu stokastik olmayan bağımsız değişkenlerin gözlenen matrisi ($k = p+1$);

β : $(kx1)$ boyutlu bilinmeyen regresyon katsayıları vektörü;

ε : $(nx1)$ boyutlu rasgele hatalar vektörüdür.

(1.1) veya (1.2)'deki çoklu doğrusal regresyon modelinde, model parametrelerinin, yani regresyon katsayılarının tahmini için genellikle en küçük kareler (Least Squares - LS) tekniği kullanılır. LS, artık kareler toplamının ilgili parametreye göre minimize edilmesi esasına dayanan bir tekniktir. β regresyon katsayılarının LS tahmin edicisi $\hat{\beta}$ yansız bir tahmin edicidir ve,

$$\frac{\partial}{\partial \hat{\beta}}\left[\left(y - X\hat{\beta}\right)'\left(y - X\hat{\beta}\right)\right] = 0 \tag{1.3}$$

eşitliği ile elde edilir. (1.3) nolu denklem çözüldüğünde

$$\hat{\beta} = \left(X'X\right)^{-1} X'y \tag{1.4}$$

bulunur.

$\hat{\beta}$ tahmin edicisine ilişkin beklenen değer $E(\hat{\beta}) = \beta$ ve varyans kovaryans matrisi $\text{var}(\hat{\beta}) = \sigma^2 (X'X)^{-1}$ eşitlikleri ile hesaplanır (Bulut, 2010).

Ancak LS tekniği ile regresyon katsayılarını tahmin etmek için bazı varsayımların sağlanması gerekir. Bu varsayımlar aşağıdaki gibi sıralanabilir:

- Hata terimleri normal, bağımsız ve aynı dağılımlı, sıfır ortalama ve σ^2 varyansına sahiptir.
- Hata terimleri ve bağımsız değişkenler arasındaki kovaryans sıfırdır.
- Bağımlı değişken ile hata terimleri arasında korelasyon yoktur.
- Bağımsız değişkenlerin aldığı değerler, tekrarlanan örneklemlerde rastgele değil belirlenmiş sabit değerlerdir.
- Gözlem sayısı bağımsız değişken sayısından fazladır.

- Bağımsız değişkenler arasında doğrusal ya da doğrusala yakın ilişki bulunmamaktadır (Wold vd., 1984; Khalaf, 2013; Ebegil vd., 2006; Gujarati, 2012; Albayrak, 2012).

Çoklu regresyon modelinde iki ya da daha fazla bağımsız değişken arasında orta ya da yüksek derecede doğrusal ya da doğrusala yakın bir ilişki bulunması, çoklu bağlantı problemini meydana getirir.

Çoklu bağlantının ortaya çıkma sebepleri kısaca aşağıda sıralanmıştır:

- Dummy değişkenlerinin doğru bir şekilde kullanılmaması

- Regresyon denkleminde bulunan bağımsız değişkenlerden bir veya birkaçının, diğer bağımsız değişkenleri kullanarak hesaplanması

- Yığını temsil eden örneğin yeterince uygun olmaması

- Değişken sayısının gözlem sayısından büyük olması

- Araştırmacının örneklemi seçerken yaptığı hataların sonucunda çoklu bağlantı problemi ortaya çıkmaktadır.

Çoklu bağlantının olması durumunda:

- Modeldeki bazı bağımsız değişken katsayı tahminleri pozitif olması gerekirken negatif, negatif olması gerekirken pozitif çıkabilir. Böylece katsayı tahminleri kararsızlık gösterir.

- LS tahmin edicileri büyük varyansa ve dolayısıyla büyük standart hatalara sahip olur, bu da doğruya yakın tahmin yapmayı zorlaştırır.

- Katsayılar için güven aralıkları oldukça büyük ve t istatistikleri oldukça küçük çıkma eğilimindedir.

- Regresyon katsayılarının çoğu veya tümü anlamsız iken, belirleme katsayısı yüksek ve model anlamlı çıkmaktadır.

- LS tahmin edicileri veri üzerinde yapılan küçük değişikliklerden çok fazla etkilenmektedir (Gujarati, 2012)

Yukarıda da ifade edildiği gibi çoklu bağlantının varlığında LS tekniğini kullanmak hatalı sonuçlar doğurur. Bunun için öncelikle çoklu bağlantının tespit edilmesi gerekir.

Çoklu bağlantının tespit edilmesinde kullanılan birçok yol vardır. Bunlardan en belirgin olanları aşağıdaki gibi sıralanabilir:

- Bağımsız değişkenler arasındaki korelasyon katsayılarının 1'e yakın olması yani, $X'X$ matrisinin köşegen dışı birimlerinin mutlak değerinin 1'e yakın olması çoklu bağlantı sorunu olabileceğini ortaya koyar. Ancak çoklu bağlantın belirlenmesi için sadece iki değişken arasındaki korelasyon katsayısına bakmak yeterli değildir.

- Varyans Şişirme Çarpanı (Variance Inflation Factor - VIF) değerlerinin bir veya bir kaçı 10'u aşarsa, bağımsız değişkenler arasında doğrusal ilişki olduğu söylenebilir. VIF değerleri

$$VIF_j = \frac{1}{1-R_j^2}\ , \qquad j = 1, 2, ..., p. \tag{1.5}$$

eşitliği ile elde edilir. Burada,

$R_j^2 : x_j$ 'nin diğer $p-1$ sayıda bağımsız değişken üzerinden modellendiği zaman bulunan belirleme katsayısıdır.

VIF değerleri $X'X$ matrisinin tersinin köşegen değerleridir. $R_j^2 = 0$ ve $VIF_j = 1$ olduğunda j nci değişken diğer bağımsız değişkenlerle doğrusal ilişkili değildir. Ancak $R_j^2 = 1$ iken VIF_j değerinin ∞ a gitmesi j nci değişkenin diğer bağımsız değişkenlerle doğrusal ilişkisinin bulunduğunu kanıtlamaktadır (Joshi, 2012).

- Tolerans Değerlerinin (Tolerance Value - TV) 0.10 ya da daha az olması da çoklu bağlantının bir göstergesidir. TV

$$TV_j = \frac{1}{VIF_j}\ , \qquad j = 1, 2, ..., p. \tag{1.6}$$

eşitliği ile edilir. Eğer TV değeri 0'a eşitse tam çoklu bağlantının varlığından, 1'e eşitse değişkenlerin birbirleriyle ilişkisiz olmasından söz edilebilir.

- Koşul Sayısının (Condition Number - CN) 100'den küçük olması ciddi bir çoklu bağlantı sorunu olmadığını, 100 ile 1000 arasında olması orta dereceli çoklu bağlantıdan güçlü çoklu bağlantıya doğru bir eğilim olduğunu, 1000'den büyük olması ise ciddi derecede çoklu bağlantı sorunun olduğunu gösterir. CN değeri

$$CN = \frac{\lambda_{\max}}{\lambda_{\min}} \tag{1.7}$$

eşitliği ile elde edilir. Burada,

$\lambda_{\max}$: $X'X$ matrisinin en büyük özdeğeri

$\lambda_{\min}$: $X'X$ matrisinin en küçük özdeğeridir (Duran, 2011; Montgomery vd., 2012).

Diğer yandan Koşul İndeksinin (Condition Index – CI) 15 ile 30 arasında olması orta dereceli, 30'dan büyük olması ise yüksek derecede çoklu bağlantının olduğunun bir göstergesidir. CI değeri

$$CI = \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}} \tag{1.8}$$

eşitliği ile elde edilir.

- $X'X$ matrisinin bir veya daha fazla özdeğerinin sıfıra yakın değerler alması da çoklu bağlantı sorunun var olduğunu göstermektedir (Joshi, 2012).

- Eğer özdeğerlerin terslerinin toplamı bağımsız değişken sayısından büyükse çoklu bağlantı sorunu mevcuttur. Çünkü veride çoklu bağlantı yoksa $\sum_{j=1}^{p} \frac{1}{\lambda_j} = p$ eşitliği sağlanır. Diğer bir ifadeyle özdeğerlerin tersleri toplamı bağımsız değişken sayısına eşittir (Joshi, 2012).

Regresyon katsayılarının LS tahmin edicileri en iyi doğrusal yansız tahmin edicilerdir. Burada "en iyi" ifadesi "en küçük varyanslı" anlamındadır. Ancak çoklu bağlantının varlığında tahmin edicinin "en küçük varyanslı" olma özelliği bozulmaktadır. Bu durumda çoklu bağlantı problemini ortadan kaldırmak gerekir. Çoklu bağlantı problemini ortadan kaldırmak için birçok yöntem önerilmiştir. Bu yöntemlerden bir tanesi yanlı tahmin tekniklerini kullanmaktır (Rawlings vd., 1998). Yanlı tahmin teknikleri LS tekniğinin bağımsızlık varsayımı, homojen varyans varsayımı ve doğrusallık varsayımını sağlamaktadır, fakat güven aralığı hesaplanmadığından normallik varsayımı yapılmamaktadır (Albayrak, 2012).

En sık kullanılan yanlı tahmin teknikleri Ridge Regresyon (Ridge Regression - RR), Temel Bileşenler Regresyonu (Principal Component Regression - PCR), Kısmi En Küçük Kareler Regresyonu (Partial Least Squares Regression - PLSR) ve Latent Kök Regresyonudur (Latent Root Regression - LRR).

Belirtmek gerekir ki, bu tez çalışmasında yanlı tahmin tekniklerinden PLSR ve PCR ele alınacaktır.

Veri boyutunu indirgeyerek, bağımlı değişkeni olabildiğince az sayıda bağımsız değişken ile açıklayan PLSR ve PCR teknikleri; istatistik, analitik kimya, tıp, biyoloji ve kemometri gibi alanlarda sıkça kullanılmaktadır (Dinç, 2007). İki teknik de çoğunlukla varyansın yanlılığa baskın olma eğiliminde olduğu çoklu bağlantılı verilere uygulandığından dolayı birçok durumda benzer sonuçlar doğurur ve doğru bir şekilde kullanıldıklarında LS' den daha başarılı tahminler verir (Colbert vd., 2002; Frank ve Friedman, 1993) .

PLSR ve PCR tekniklerinde ilk olarak, değişkenler arasındaki birim farklılıklarını ortadan kaldırmak için X bağımsız değişkenler matrisi standartlaştırılır. Daha sonra standartlaştırılmış bu değişkenlerin lineer kombinasyonları kullanılarak, temel bileşenler ya da bileşenler adı verilen yeni ortogonal değişkenler elde edilir. Son olarak bu değişkenlere LS tekniği uygulanarak regresyon katsayıları tahmin edilir (Vigneau vd., 1996).

PLSR ve PCR tekniği çoklu bağlantı problemini benzer şekilde ele almaktadır. Ancak yapılan simülasyon çalışmaları, PLSR' nin PCR 'den daha az sayıda bileşen ile daha küçük hata kareler ortalamasına (Mean Square Error - MSE) ulaştığını gösterir. Ayrıca PLSR PCR' den daha az hesaplama gerektirir**.** Bu iki teknik, gözlem sayısı bağımsız değişken sayısından fazla olan verilere uygulanabildiği gibi, gözlem sayısı bağımsız değişken sayısından küçük olan verilere de uygulanabilir (Helland, 1988; Vigneau vd., 1997).

1960'larda Herman Wold PLSR'yi ekonometrik bir teknik olarak geliştirmiştir, ancak bu tekniğin asıl savunucuları kimya mühendisleri ve kemometriciler olmuştur. PLSR spektral verideki kimyasal değişkenleri tahmin etmek için bir kalibrasyon metodu olarak Svante Wold ve Harald Martens tarafından geliştirilmiştir. Bu teknik çoklu regresyon ve temel bileşenler analizinin özelliklerini birleştiren ve genelleştiren bir tekniktir. Amaç bağımlı değişkendeki bilginin çoğunu koruyarak bağımlı ve bağımsız değişkenler arasındaki kovaryansı maksimum yapacak şekilde optimum sayıda bileşen elde etmek ve bağımlı değişkeni tahmin etmektir. Ayrıca PLSR tekniği, bağımlı değişken sayısı tek olduğunda kullanılabildiği gibi, birden fazla olduğu zaman da kullanılabilir (Helland, 1988; Vigneau vd., 1997; Abdi, 2003; Li vd., 2002; Garthwaite, 1994; Tobias, 1995).

PLSR ve PCR teknikleri arasındaki temel farklılık; PCR temel bileşenlerin elde edilmesinde sadece bağımsız değişkenler üzerindeki bilgiyi kullanırken, PLSR tekniği hem bağımlı hem de bağımsız değişkenler üzerindeki bilgiyi kullanır (Naes ve Martens, 1985; Garthwaite, 1994).

Bu tez çalışmasının amacı, çoklu bağlantının yani bağımsız değişkenler arasındaki korelasyonun az, orta ve yüksek derecede olduğu, farklı gözlem sayısı, farklı değişken

sayısı ve farklı varyans durumlarında PLSR ve PCR tekniklerinden hangisinin RMSECV kriteri açısından daha üstün tahmin yeteneğine sahip olduğunu belirlemek ve genellemelere ulaşmaktır. Bu teknikler simülasyon çalışması ile çeşitli durumlar için karşılaştırılmıştır.

Tezin ikinci bölümünde PCR ve PLSR teknikleri ile ilgili literatürdeki uygulamalı ve teorik çalışmalar incelenmiş, hangi kritere göre karşılaştırıldıkları ve yapılan çalışmalarda hangi tekniğin ne gibi durumlarda üstünlük sağladığı, ne zaman benzer sonuçlar verdiği incelenmiştir.

Üçüncü bölümde PLSR ve PCR tekniklerinin matematiksel modellerine ve tahmin edicilerinin nasıl elde edildiğine yer verilmiştir. Ayrıca PLSR modelinde kullanılan belli başlı algoritmalara ve PLSR modelinin nasıl yorumladığına da değinilmiştir.

Dördüncü bölümde MATLAB programı kullanılarak normal dağılımdan korelasyonlu veri üretilmiş ve daha sonra farklı gözlem sayıları, farklı korelasyonlar, farklı bağımsız değişken sayıları ve farklı varyans değerleri için PLSR ve PCR teknikleri RMSECV kriteri bakımından Monte Carlo simülasyonu ile karşılaştırılmıştır. Buradaki amaç, bu iki tekniğin veriye uyum başarısı hakkında genelleme yapabilmektir. Ayrıca RMSECV değerinin nasıl hesaplandığına bu bölümde yer verilmiştir.

Beşinci bölümde yapılan simülasyon çalışmasından elde edilen sonuçlar literatürdeki çalışmalarla karşılaştırılarak sonuç ve önerilere yer verilmiştir.

2. LİTERATÜR ARAŞTIRMASI

Literatürde PLSR ve PCR tekniklerini birbirleriyle ve diğer yanlı tahmin teknikleriyle hem gerçek veri seti üzerinde hem de simülasyon çalışması ile karşılaştıran pek çok çalışmaya rastlanmaktadır. Bu çalışmaların bazıları aşağıda verilmiştir:

Naes ve Martens (1985), PCR ve PLSR tekniklerini hem simülasyon ile üretilmiş veri hem de gerçek veri üzerinde MSE bakımından karşılaştırmıştır. PCR ve PLSR'nin benzer optimal tahmin sonuçları verdiği ancak PLSR'nin PCR'den daha az bileşen kullandığı görülmüştür.

Helland (1988), PLSR'nin teorik yapısını incelemiş ve simülasyon çalışması ile PLS'nin PCR'den daha az bileşen kullanarak daha düşük MSE değerine ulaştığını göstermiştir.

Thomas ve Haaland (1990), simülasyon çalışması yaparak PLSR ve PCR tekniklerini karşılaştırmıştır. Bu iki tekniğin çok benzer sonuçlar verdiğini, ancak veri seti hakkında spesifik bilginin bulunmadığı durumlarda optimal ya da optimale çok yakın sonuçlar vermesinden dolayı PLSR tekniğinin kullanılmasını tavsiye etmiştir.

Frank ve Friedman (1993), kemometride yaygın olarak kullanılan PCR ve PLSR metodlarını incelemiş, istatistiksel açıdan LS, RR, PCR, PLSR ve değişken altseti seçme (variable subset selection – VSS) tekniklerini, çeşitli koşullar altında karşılaştırarak bir simülasyon çalışması yapmıştır. Çalışılan tüm durumlarda RR, diğer tekniklere baskın çıkmıştır. PLSR ise PCR'ye göre çok fazla olmasa da üstün bulunmuştur.

Jong (1993), aynı sayıda bileşen kullanıldığında, PLSR tekniğinin belirleme katsayısının en az PCR tekniğinin belirleme katsayısı kadar büyük oluğunu teorik açıdan ele almıştır.

Luinge vd. (1993), eski iki teknik olan Röse-Gottlieb ve Kjeldahl referans değerleriyle, sütün içerisindeki yağ, protein ve laktoz miktarlarını tahmin etmeye

çalışmıştır. Ayrıca yağ, protein ve laktoz oranı MultiSpec kızılötesi filtre cihazını kullanarak tespit edilmiştir. Daha sonra PLSR ve PCR teknikleri veriye uygulanarak sonuçlar karşılaştırılmıştır. Yapılan çalışmada bu tekniklerin tahmin hatasına göre karşılaştırılabilir olduğu, PLSR ile PCR'nin benzer sonuçlar verdiği görülmüştür. Modeli oluşturmak için de aynı sayıda bileşen kullanılmıştır.

Almøy (1996), koşulsuz beklenen hata kareler (unconditional expected squared error) kriterine göre 5 farklı tahmin tekniğini karşılaştırmıştır. Bunlar: özdeğerlerin büyüklüğüne göre temel bileşenler regresyonu (PCR1), t -değerlerinin büyüklüğüne göre temel bileşenler regresyonu (PCR2), PLSR, kısıtlı temel bileşenler regresyonu (RPCR) (restricted principal component regression) ve modified en çok olabilirlik regresyonudur (Modified maximum likelihood - MMLR). Geniş bir simülasyon çalışması yapılmış ve PCR1, PLSR ve RPCR'nin benzer sonuçlar verdiği görülmüştür.

Vigneau vd. (1996), gıda maddelerinin karışımlarının kızılötesine yakın spektrumlarından oluşan veriye uygulanan LRR'nin sınırlarını (limit) belirlemeye çalışmış ve elde ettiği değerleri PCR ve PLSR ile karşılaştırmıştır. Minumum hatalarına göre bu üç teknik karşılaştırılmış ve az bir farkla en küçük hataya LRR tekniği ile ulaşıldığı görülmüştür. Ancak en küçük hatayı elde etmek için LRR'nin PLSR ve PCR'den çok daha fazla sayıda bileşene ihtiyacı oluğu görülmüştür.

Diaz vd. (1997), tarım ilaçlarının üçlü karışımından oluşan veri üzerine PCR, PLSR ve LS tekniklerini uygulamıştır. Aynı sayıda bileşeni kullanarak kestirimin hata kareler ortalamasının karekökü (Root Mean Square Error of Prediction-RMSEP) kriterine göre PCR ve PLSR nin benzer sonuçlar verdiği görülmüştür.

Ni ve Gong (1997), PCR ve PLSR tekniklerinin tahmin yeteneğini, göreli tahmin hatası (relative prediction error) açısından karşılaştırmıştır. Tahminin hassasiyeti (precision of prediction) açısından iki teknik arasında önemli bir fark gözlemlenmemiştir.

Vigneau vd. (1997) spektroskopydeki kalibrasyon problemleri ile ilgili iki veri seti üzerinde RR, PCR, PLSR ve Ridge Temel Bileşenler Regresyonunun (Ridge Principal Component Regression- RPCR) tahmin yeteneklerini modelde yer alan bileşen sayısına

göre incelemiştir. Bu teknikler (Cross Validation-CV) yöntemi kullanılarak karşılaştırılmıştır.

Guiteras vd. (1998), iki senkronize flüoresans spektrumlarının birleşiminden elde edilen çok değişkenli veri üzerine LS, PCR ve PLSR tekniklerini uygulamıştır. Bu üç teknik fark kareler ortalamasının göreli karekökü (relative root mean squared difference) kriterine göre karşılaştırılmıştır. Tahmin performansı bakımından PCR de iyi sonuçlar vermesine rağmen, en iyi tahmin sonuçlarını PLSR tekniği vermiştir.

Yeniay ve Göktaş (2002), Türkiye'nin Kişi Başına Düşen Gayri Safi Yurtiçi Hasılasını tahmin etmek için tespit etiği 26 tane bağımsız değişkene ilişkin Türkiye'nin 80 ilinden derlediği veri seti üzerine LS, RR, PCR ve PLSR tekniklerini uygulamıştır. Dört farklı teknikle elde edilen modellerin tahmin yetenekleri karşılaştırılmış ve sırasıyla PLSR ve PCR'nin en iyi modelleri verdiği görülmüştür.

Wentzell ve Montoto (2003), çok sayıda bileşen içeren kompleks kimyasal karışımlar üzerine simülasyon çalışması yapmıştır. Çalışmada karışım bileşenlerinin sayısı, kalibrasyon örneklemlerinin sayısı, bileşen sayısı gibi simülasyon parametreleri kullanılmış, PCR ve PLSR teknikleri karşılaştırılmıştır. Bu iki tekniğin tahmin yetenekleri arasında önemli bir fark görülmemiştir.

Cheng ve Wu (2006), PLSR algoritmasına dayanarak Uyarlanmış Kısmi En Küçük Kareler Regresyonu (Modified Partial Least Square Regression - MPLSR) tekniğini tanıtmıştır. Hem otomobil sektörüne ait gerçek bir veri üzerinde hem de Monte Carlo simülasyonu ile RR, PCR, PLSR ve MPLSR tekniklerinin performansı karşılaştırılmıştır. Özellikle gözlem sayısı ve bağımsız değişken sayısı arasındaki oran düşük olduğunda MPLSR tekniğinin en iyi olduğu görülmüştür.

Hemmateenejad vd. (2007), dalga boyu seçiminin etkisini göz önünde bulundurarak PCR ve PLSR tekniklerini karşılaştırmıştır. Tüm spektral veri kullanıldığında PLSR tekniğinin PCR tekniğinden bazen daha iyi bazen daha kötü sonuçlar ürettiğini fakat seçilen dalgaboyu bölgesinde PCR ve PLSR tekniklerinin benzer sonuçlar verdiğini

görmüştür. Dolayısıyla bir tekniğin diğer tekniğe göre üstün olup olmadığı hakkında yorum yapılamayacağını söylemiştir.

Li (2010), çalışmasında PCR, PLSR ve RR tekniklerini incelemiş, teorik yönlerini ele almıştır. Ayrıca hem simülasyon çalışması yaparak hem de gerçek bir veri seti üzerinde çalışarak RMSEP kriterine göre, tahmin performanslarını karşılaştırmıştır. RR, PCR ve PLSR tekniklerinin benzer sonuçlar verdiğini görmüştür.

Yaroshchyk vd. (2012), çalışmasında PCR, PLSR, Multi-Blok PLSR (MB-PLSR) ve Serial PLSR (S-PLS) tekniklerini tahminin doğruluğu bakımından karşılaştırmıştır. PLSR ve PCR tekniklerinin benzer sonuçlar verdiği ancak PLSR tekniğinin daha az bileşen kullandığı görülmüştür. Ayrıca diğer teknikler PLSR ve PCR ile kıyaslandığında daha kötü performans vermiştir.

Irfan vd. (2013), gerçek bir veri seti üzerine LS, RR, PCR ve PLSR tekniklerini uygulamış ve bu teknikleri hata kareler ortalamasının karekökü (Root Mean Square Error-RMSE), çapraz geçerliğin hata kareler ortalamasının karekökü (Root Mean Square Error cross Validation – RMSECV), çapraz geçerlik parametresi (Cross Validation Parameter-CVP) ve R^2 kriterlerine göre karşılaştırmıştır. PLSR tekniğinin diğer teknikler ile karşılaştırıldığında daha üstün sonuç verdiği görülmüştür.

Schumann vd. (2013), pelvis şeklinin tahmini için PCR ve PLSR tekniklerini kullanmıştır. Bu iki regresyon tekniğiyle yapılan tahminler arasında $\alpha = 0.01$ anlamlılık düzeyinde fark bulunmamıştır. Ayrıca çalışmada, bileşen sayısının elde edilmesinde birini dışarda bırakarak çapraz geçerlik (Leave One Out Cross Validation-LOOCV) yöntemi kullanılmıştır.

Khajehsharifi vd. (2014), gerçek bir veri üzerinde PLSR ve PCR tekniklerinin tahmin yeteneklerini RMSEP kriterine göre karşılaştırmıştır. Askorbik asit, dopamin ve ürik asit için bileşenlerin optimum sayısına ilişkin RMSEP değerleri elde edilmiştir. PLSR'nin daha iyi tahmin yeteneğine sahip olduğu görülmüştür.

Mahesh vd. (2015), Kanada da buğday yetiştirilen bölgelerden elde edilen toplu örneklerden yararlanarak buğdayın protein içeriği ve sertlik değerlerini tahmin etmek için

PLSR ve PCR tekniklerini kullanmıştır. Tahminin hata kareler ortalaması (Mean Square Error of Prediction-MSEP), Çapraz Geçerliliğin Hata Karesi (Square Error of Cross validation-SECV) ve korelasyon katsayısına göre yapılan karşılaştırmalarda, PLSR modelinin tahmin performansının, PCR modelinin tahmin performansından daha iyi olduğu görülmüştür.

3. MATERYAL VE YÖNTEM

Bu bölümde PLSR ve PCR tekniklerinin matematiksel modeline ve tahmin edicilerinin nasıl elde edildiğine, PLSR modelinin nasıl yorumlandığına ve PLSR için tanıtılan algoritmalara değinilecektir.

3.1. PLSR Modeli

PLSR'nin amacı çoklu bağlantı sorunundan dolayı regresyon yaklaşımını kullanmanın mümkün olmadığı ve X matrisinin tekil olduğu durumlarda, birinci bloğu temsil eden p boyutlu bağımsız değişkenler uzayı $X \subset R^p$ ve ikinci bloğu temsil eden m boyutlu bağımlı değişkenler uzayı $Y \subset R^m$ olmak üzere X ve Y blokları arasındaki yapıyı tanımlayabilmek ve Y 'yi X 'ler aracılığıyla tahmin etmektir (D'Ambra ve Sarnacchiaro, 2010). PLSR skor vektörleri aracılığıyla bu iki blok arasındaki ilişkiyi modellemektedir. PLSR sıfır ortalamalı X ve Y değişkenler matrisini aşağıdaki şekilde ayrıştırır:

$$\begin{aligned} X &= TP' + E \\ Y &= UQ' + F \end{aligned} \tag{3.1}$$

Burada T ve U skor matrislerini, P ve Q yük matrislerini ve E ve F artık matrisini göstermektedir (Rosipal ve Krämer, 2006).

Bu ayrışım T ve U arasındaki kovaryansı maximize etmek için yapılır (Maitra ve Yan, 2008). Bu ayrışımı yapmanın X üzerine doğrudan LS' i kullanarak modelleme yapmaktan daha güvenilir olduğuna inanılır, çünkü bileşenler orijinal veri altındaki yapı ile doğru bir şekilde örtüşmektedir (Zeng vd., 2007).

Skor vektörleri t ve u arasındaki örneklem kovaryansı aşağıdaki eşitlik ile elde edilir (Rosipal ve Krämer, 2006):

$$\left[\operatorname{cov}(t,u)\right]^2 = \left[\operatorname{cov}(Xw, Yc)\right]^2 = \max_{|r|=|s|=1}\left[\operatorname{cov}(Xr, Ys)\right]^2 \tag{3.2}$$

Burada $\text{cov}(t,u) = t'u/n$ dir ve (3.2) nolu denklemdeki w ve c ağırlık vektörleri ileride bahsedilecek doğrusal olmayan iteratif en küçük kareler (Nonlinear Iterative Partial Least Squares – NIPALS) algoritmasına dayanmaktadır (Rosipal ve Krämer2006).

PLSR modeli kurulurken tüm bağımsız değişkenleri kullanmak yerine, daha az sayıda bileşen kullanılır. Bileşen olarak adlandırılan yeni değişkenler X skorları olarak adlandırılır ve $t_a (a = 1,2,...,A)$ ile gösterilir. X skorları hem Y 'nin tahmin edicileridir hem de X 'i modellerler. En azından, bir kısım X ve Y 'nin aynı bileşenler tarafından modellendiği varsayılır. X skorları A sayıda ve ortogonaldir. Bunlar orijinal x_j değişkenlerinin w_{ja}^* $(a = 1,2,...,A)$ ağırlıkları ile çarpımının lineer kombinasyonlarından oluşur. Bazen bu ağırlıklar r_{ja} ile de gösterilir. Formüller aşağıdaki gibi vektör biçiminde ya da parantez içindeki gibi matris biçiminde gösterilmektedir.

$$t_{ia} = \sum_j W_{ja}^* X_{ij}$$

$$(T = XW^*) \qquad (3.3)$$

Burada T skor matrisini, W^* ağırlık matrisini göstermektedir.

PLSR'de ağırlıklar bileşenler ve bağımlı değişken arasındaki kovaryansı maximize ederek belirlenir (Zeng vd.,2007)

(3.3) denkleminde gösterilen X 'in skorları olan t 'ler belli özelliklere sahiptir. Bunlar aşağıda verilmiştir.

(i) Bu skorlar X 'i iyi bir şekilde temsil eden p_{aj} yükleriyle çarpılır. Dolayısıyla X 'e ait e_{ij} artıkları küçüktür.

$$x_{ij} = \sum_{a} t_{ia} p_{aj} + e_{ij}$$

$$(X = TP' + E) \tag{3.4}$$

Y çok değişkenli olduğu zaman, Y skorları olan u_a'lar, Y 'yi iyi bir şekilde temsil eden c_{am} ağırlıkları ile çarpılır. Dolayısıyla g_{im} artıkları küçüktür.

$$y_{im} = \sum_{a} u_{ia} c_{am} + g_{im}$$

$$(Y = UC' + G) \tag{3.5}$$

(ii) X skorları, Y 'nin iyi tahmin edicileridir.

$$y_{im} = \sum_{a} c_{ma} t_{ia} + f_{im}$$

$$(Y = TC' + F) \tag{3.6}$$

Burada gözlenen ve modellenen bağımlı değişkenler arasındaki sapmayı gösteren Y artıkları f_{im} ile gösterilmektedir. (3.3) denklemi (3.6)'da yerine yazıldığında (3.7) denklemi elde edilir.

$$y_{im} = \sum_{a} c_{ma} \sum_{j} w^{*}_{ja} x_{ij} + f_{im} = \sum_{j} b_{mj} x_{ij} + f_{im}$$

$$(Y = XW^{*}C' + F = XB + F) \tag{3.7}$$

Burada b_{mj} PLS- regresyon katsayılarını göstermektedir.

$$b_{mj} = \sum_{a} c_{ma} w^{*}_{ja}$$

$$(B = W^{*} C') \tag{3.8}$$

Burada $m = 1$ olduğunda, yani Y bağımlı değişkeni tek ve $X'X$ matrisi köşegen olduğunda özel bir durum ortaya çıkmaktadır. Bu durumda X matrisi ortogonal tasarımdan meydana geldiğinden ve X matrisinde korelasyon yapısı bulunmadığından dolayı PLSR ile tek bir bileşenin elde edilmesinden sonra LS çözümüne ulaşır. Sonuç olarak PLSR ile LS regresyon katsayıları $w_1 c_1'$'e eşit bulunur (Wold., 2001).

a bileşenlerinin her birinin elde edilmesinden sonra X matrisinden $t_a p_a'$ çıkartılarak X matrisi indirgenebilir ve (3.3) denklemi aşağıdaki şekilde yazılabilir (Wold.,2001).

$$t_{ia} = \sum_{j} w_{ja} e_{ij,a-1}$$

$$(t_a = E_{a-1} W_a) \tag{3.9}$$

$$e_{ij,a-1} = e_{ij,a-2} - t_{i,a} p_{a-1,j}$$

$$(E_{a-1} = E_{a-2} - t_{a-1} p'_{a-1}) \tag{3.10}$$

$$e_{ij,0} = X_{ij}$$

$$(E_0 = X) \tag{3.11}$$

Yukarıdaki denklemlerde PLSR modeli alternatif olarak w ağırlıkları cinsinden yazılmıştır.

Ancak w ağırlıkları X ile doğrudan ilgili olan w^* ağırlıklarına dönüştürülebilir. Bu iki ağırlık arasındaki ilişki aşağıdaki gibidir (Wold vd., 2001):

$$W^* = W(P'W)^{-1} \tag{3.12}$$

Ayrıca PLSR'de $X'YY'X$ varyans-kovaryans matrisinin ilk özvektörü, birinci ağırlık vektörü w_1 ve diğer ağırlık vektörleri ise indirgenmiş $Z_a'YY'Z_a$ matrisinin özvektörleridir. Burada

$$Z_a = Z_{a-1} - T_{a-1}P_{a-1}' \tag{3.13}$$

eşitliği ile elde edilir (Wold vd., 2001).

Aynı şekilde ilk ağırlık vektörü c_1 de; $Y'XX'Y$ matrisinin ilk özvektörüdür. Ayrıca ilk skor vektörleri t_1 ve u_1 vektörleri $XX'YY'$ matrisi ve $YY'XX'$ matrislerinin ilk özvektörleridir Daha sonraki X skor vektörleri olan t_a' lar da $Z_aZ_a'YY'$ matrisinin özvektörleridir (Abdi, 2003; Wold vd., 2001).

u_n , c_n , t_n , w_n vektörlerinin iteratif olarak elde edilişi aşağıdaki şekilde gösterilebilir (Höskuldsson,1988) :

$$\begin{aligned} u_n &= Yc_n / (c_n'c_n) \\ &= YY't_n / (c_n'c_n)(t_n't_n) \\ &= YY'Xw_n / [(c_n'c_n)(t_n't_n)(w_n'w_n)] \\ &= YY'XX'u_{n-1} / [(c_n'c_n)(t_n't_n)(w_n'w_n)(u_{n-1}'u_{n-1})] \end{aligned} \tag{3.14}$$

$$c_n = Y'XXYc_{n-1} / [(t_n't_n)(w_n'w_n)(u_{n-1}'u_{n-1})(c_{n-1}'c_{n-1})]$$
$$t_n = XX'YY't_{n-1} / [(w_n'w_n)(u_{n-1}'u_{n-1})(c_{n-1}'c_{n-1})(t_{n-1}'t_{n-1})]$$
$$w_n = X'YY'Xw_{n-1} / [(u_{n-1}'u_{n-1})(c_{n-1}'c_{n-1})(t_{n-1}'t_{n-1})(w_{n-1}'w_{n-1})] \qquad (3.15)$$

Bu denklemler algoritmanın, bir matrisin en büyük özdeğerini bulmak için kullanılan power metoduna benzer şekilde çalıştığını göstermektedir.

İterasyon sonucunda elde edilen yakınsama aşağıdaki denklemler ile gösterilebilir:

$$YY'XX'u = au$$
$$Y'XX'Yc = ac$$
$$XX'YY't = at$$
$$X'YY'Xw = aw$$

Power metoduna göre a en büyük özdeğerdir. u , c , t ve w ise yukarıdaki matrislerde en büyük özdeğere karşılık gelen özvektörlerdir. Bu vektörlere ilişkin bazı özellikler aşağıda verilmiştir:

- w vektörleri karşılıklı ortogonaldir:

$$\left(w_i, w_j\right) = w_i'w_j = 0 \qquad i \neq j \qquad (3.16)$$

İspat:

İleride de bahsedileceği gibi PLSR algoritmasına dayanarak, artık matrisi X_i, önceki artık matrislerinin hesaplanmasına dayanmaktadır. Yani,

$$\begin{aligned} X_i &= X_{i-1} - t_{i-1}p_{i-1}' \\ &= X_{i-1} - t_{i-1}t_{i-1}'X_{i-1} / \left(t_{i-1}'t_{i-1}\right) \\ &= \left[I - t_{i-1}t_{i-1}' / \left(t_{i-1}'t_{i-1}\right)\right] X_{i-1} \\ &= \left[I - t_{i-1}t_{i-1}' / \left(t_{i-1}'t_{i-1}\right)\right]\left[X_{i-2} - t_{i-2}t_{i-2}'X_{i-2} / \left(t_{i-2}'t_{i-2}\right)\right] \end{aligned} \qquad (3.17)$$

şeklinde hesaplanmaktadır.

$i < j$ olduğu varsayılsın. Dolayısıyla (3.17)'den faydalanarak

$$X_j = Z\left[X_i - t_i t_i' / \left(t_i' t_i\right) X_i\right] \tag{3.18}$$

elde edilir. Burada Z herhangi bir matristir.

$$X_j w_i = 0 \qquad\qquad j > i \tag{3.19}$$

olduğunu gösterelim.

(3.18)'i (3.19)'da yerine yazılırsa

$$\left[X_i - t_i t_i' / \left(t_i' t\right) X_i\right] w_i = t_i - t_i t_i' / \left(t_i' t_i\right) t_i = 0$$

elde edilir. Buradan da

$$w_j' w_i = w_j' X_j' Y_j Y_j' X_j w_i / a_j = 0$$

olduğu görülür.

- t_i vektörleri karşılıklı ortogonaldir.

$$\left(t_i, t_j\right) = t_i' t_j = 0 \qquad\qquad i \neq j \tag{3.20}$$

İspat:

$i < j$ olduğu varsayılsın. X_j (3.17) denkleminde gösterildiği gibi yazılırsa

$$\begin{aligned} X_j &= X_{j-1} - X_{j-1} w_{j-1} t_{j-1}' X_{j-1} / \left(t_{j-1}' t_{j-1}\right) \\ &= X_{j-1}\left[I - w_{j-1} t_{j-1}' X_{j-1} / \left(t_{j-1}' t_{j-1}\right)\right] \\ &= X_{i+1} Z \\ &= \left[X_i - t_i t_i' X_i / \left(t_i' t_i\right)\right] Z \end{aligned}$$

Burada Z herhangi bir matristir. Buradan da

$$t_i' X_j = 0 \qquad i < j \qquad (3.21)$$

olduğu görülür ve böylece

$$t_i' t_j = t_i' X_j w_j = 0 \qquad i < j \qquad (3.22)$$

elde edilir.

- $i < j$ varsayımı altında w_i vektörleri p_j vektörlerine, t_i vektörleri de u_j vektörlerine ortogonaldir.

$$\left(w_i, p_j\right) = w_i' p_j = 0 \text{ ve } \left(t_i, u_j\right) = t_i' u_j = 0 \qquad i < j \qquad (3.23)$$

3.2. PLSR Modelinin Yorumlanması

PLSR modeli x değişkenlerinin lineer kombinasyonları olan yeni değişkenler elde eder ve bu yeni değişkenler bağımlı y değişkeninin tahmin edicileri olur. Tüm t, u, w, p, c parametreleri PLSR algoritması tarafından belirlenir (Wold vd., 2001).

PLSR modelinin yorumlanmasında, t ve u skorları verilen problem ve model konusunda nesneleri ve onların benzerlik ve farklılıkları hakkındaki bilgiyi içerir.

w ve c ağırlıkları X ve Y arasındaki sayısal ilişkiyi oluşturmak için değişkenlerin nasıl bir araya getirileceği hakkında bilgi verir. Böylece ağırlıklar hangi X değişkenlerinin daha fazla öneme sahip olduğunu gösterebilmek açısından önemlidir. Aynı zamanda PLS ağırlıkları X ve Y arasındaki pozitif korelasyonu da açıklamaktadır (Wold vd., 2001)

p yükleri, skorlar ve orijinal değişkenler arasındaki ilişkiyi yorumlamak için kullanılır. Yani veri setindeki farklı bileşenler ile en çok hangi değişkenlerin ilişkili

olduğunu söyler. Eğer veri setinde sınıflama yapılmak istenirse, yüklere ait grafikler hangi değişkenlerin farklı sınıfları belirlediğini göstermek için kullanılabilir (Ziegel, 2012).

X artıkları ise Y 'nin modellenmesinde kullanılmayan kısımdır, ancak X uzayında aykırı değerlerin belirlenebilmesi için yararlıdır.

3.3. PLSR Modelinin Bileşenlere İlişkin Varsayımı

Kısmi en küçük kareleri modellemede araştırılan sistem veya sürecin bileşenlerden etkilendiği varsayılır. Ancak bu bileşenlerin sayısı genellikle bilinmemektedir ve PLSR tekniği ile bileşen sayısı tespit edilmektedir. Eğer bileşen sayısı bağımsız değişken sayısına eşitse o zaman bağımsız değişkenler birbirinden bağımsızdır ve LS ile PLSR teknikleri aynı sonuçları verir. Ancak pratikte bağımsız değişkenler birbirinden bağımsız değildir. Yani X matrisi eksik ranklıdır. Bu durumda PLSR, LS çözümünden istatistiksel olarak daha etkin sonuç verir ve bu nedenle daha iyi tahminler elde edilir. PLSR, X 'in Y ile ilişkili olmayan kısımları olabileceğini varsayar. Böylece LS tahminlerinin aksine X içindeki hatayı tolere eder.

3.4. PLSR Algoritmaları

Literatürde PLSR modeli için çok sayıda algoritma tanıtılmıştır. Ancak temel olarak PLSR, doğrusal olmayan yinelemeli kısmi en küçük kareler (Nonlinear Iterative Partial Least Squares - NIPALS) algoritmasına dayanmaktadır. Bunun dışında literatürde PLS yönteminin istatistiksel olarak esinlenilmiş değişikliğinin basit bir uygulaması (Straightforward Implementation of a Statistically Inspired Modification of the PLS Method-SIMPLS) algortiması, evrensel kısmi en küçük kareler (Universal Partial Least Squares-UNIPALS) algoritması, gözlem-uzaklığı kısmi en küçük kareler (Sample-Distance Partial Least Squares) algoritması tanıtılmıştır. Ancak bu çalışmada NIPALS algoritması ve SIMPLS algoritmasına değinilecektir. Çünkü uygulamada kullanılan Matlab PLS Toolbox programı bu algoritmalara dayanmaktadır.

3.4.1. NIPALS algoritması

1984 'te basit NIPALS algoritması PLSR temel algoritması olarak kabul edilmiştir ve Wold tarafından geliştirilmiştir. Bu algoritmanın adımları aşağıdaki gibidir:

1.adım: Algoritma ölçeklendirilmiş ve merkezileştirilmiş X ve Y matrisleri ile başlar.

2.adım: Başlangıç vektörü u genellikle Y ' nin kolonlarından biridir. Ancak tek bir y değişkeni olduğunda $u = y$ alınır.

3.adım: X değişkenine ilişkin ağırlıklar $w = X'u / (u'u)$ biçiminde hesaplanır. w, $\|w\| = 1$ olacak şekilde ölçeklendirilir.

4.adım: X skorları olan t'ler $t = Xw$ şeklinde hesaplanır.

5.adım: Y değişkenine ilişkin ağırlıklar $c = Y't / (t't)$ biçiminde hesaplanır. c , $\|c\| = 1$ olacak şekilde ölçeklendirilir.

6.adım: Son olarak, y skorlarının güncellenmiş bir kümesi, $u = Yc / c'c$ şeklinde hesaplanır.

7.adım: t ' deki değişimden yararlanılarak, yakınsaklık test edilir. Örneğin, $\|t_{eski} - t_{yeni}\| / \|t_{yeni}\| < \varepsilon$ 'dır. Burada ε , 10^{-6} ya da 10^{-8} arasında küçük bir değerdir. Bu yakınsaklık gerçekleşmezse 3. adıma gidilir, eğer gerçekleşirse 8. Adım ile devam edilir. y değişkeninin tek olması durumunda, yakınsama tek seferde olur ve doğrudan 8.adım ile devam edilir.

8.adım: X ve Y 'den, elde edilen bileşen çıkarılarak indirgenmiş matrisler elde edilir ve bu matrisler bir sonraki bileşenin elde edilmesinde yeni X ve Y matrisleri olarak kullanılır.

X yükleri: $p = X't / (t't)$

Y yükleri: $q = Y'u / (u'u)$

Regresyon (t üzerine u'nun): $b = u't / (t't)$

Artık Matrisleri: $X \to X - tp'$ ve $Y \to Y - btc'$

Algoritma bağımlı değişkenlerdeki değişimin büyük bir kısmı açıklanıncaya kadar devam edilir.

3.4.2. SIMPLS algoritması

SIMPLS algoritması 1993'te Sijmen de Jong tarafından yayınlanmıştır ve diğer algoritmalardan temel farkı elde edilen PLS bileşenlerinin ortogonalizasyonudur. SIMPLS algoritması skorları hesaplarken, indirgenmiş X matrislerini kullanmak yerine orijinal X matrislerinin kombinasyonlarını kullanır. Bu yaklaşım NIPALS ile her zaman aynı sonucu vermese de fark çok küçüktür ve dolayısıyla birçok durumda önemli değildir. Bu küçük farklılığın meydana gelme sebebi $X'Y$ matrisinin NIPALS algoritması ile aynı şekilde indirgenmemesidir. Fakat SIMPLS algoritması veri setinin tüm türleri için oldukça hızlıdır (Lindgren ve Rannar, 1998).

NIPALS algoritmasının aksine bu algoritmada, ilk önce amaç belirlenir, daha sonra optimizasyon kriteri türetilir ve en son algoritma oluşturulur (De Jong, 1993)

Bu algoritmada NIPALS metoduyla aynı çizgide kalabilmek için X 'in ardışık ortogonal vektörleri t_a ve buna karşılık gelen Y 'nin skorları u_a aşağıdaki gibi hesaplanır. Dolayısıyla burada yapılacak olan merkezileştirilmiş veriye doğrudan uygulanabilecek olan r_a ve q_a , $(a = 1,..., A)$ ağırlık vektörlerinin hesaplanmasıdır.

$$t_a = X_0 r_a \qquad a = 1, 2, ..., A \tag{3.24}$$

$$u_a = Y_0 q_a \qquad a = 1, 2, ..., A \tag{3.25}$$

Bu ağırlıklar, skor vektörleri t_a ve u_a arasındaki kovaryansı maximum yapacak şekilde belirlenmelidir. Aşağıda verilen koşullar ağırlıkların elde edilmesindeki kısıtlardır (De Jong, 1993).

- Kovaryansın maximizasyonu: $u_a' t_a = q_a' (Y_0' X_0) r_a = \max!$
- r_a ağırlıklarının normalleştirilmesi: $r_a' r_a = 1$
- q_a ağırlıklarının normalleştirilmesi: $q_a' q_a = 1$
- t skorlarının ortogonalliği: $t_b' t_a = 0 \qquad a > b$

Son kısıt konulmadığı zaman sadece tek bir çözüm vardır: r_1 ve q_1, p bağımsız ve m bağımlı değişken sayısına sahip çapraz çarpım matrisi $S_0 \equiv X_0'Y_0$'ın ilk sol ve sağ tekil vektörleridir. Burada X_0 ve Y_0 matrisleri sırasıyla X ve Y veri matrislerinin merkezileştirilmiş şeklidir. Birden fazla çözüm elde etmek ve X 'in ortogonal faktörlerinin bir kümesini oluşturmak için, yukarıda verilen kısıtlardan son kısıtın da var olması gerekir. Bu yüzden $a > b$ olduğunda aşağıdaki eşitliğin sağlanması gerekir (De Jong, 1993):

$$t_b' t_a = t_b' X_0 r_a = \left(t_b' t_b\right) p_b' r_a = 0 \tag{3.26}$$

Burada p_b, orijinal X değişkenleri ve *b*. PLS faktörü arasındaki ilişkiyi gösteren bir yük vektörüdür. Ancak (3.26) denklemi herhangi bir yeni ağırlık vektörü r_a'nın $(a>1)$ önceki tüm yük vektörlerine, ortogonal olmasını gerektirmektedir.

$P^{\perp}_{a-1}$ gerekli bir dik gösterici (projector) olmak üzere (3.27) denklemi ile ifade edilir ve (3.28) denkleminin sağlanmasını gerektirir.

$$P^{\perp}_{a-1} = I_p - P_{a-1}\left(P_{a-1}' P_{a-1}\right)^{-1} P_{a-1}' \tag{3.27}$$

$$r_a = P^{\perp}_{a-1} r_a \qquad (a>1) \tag{3.28}$$

Burada $P_{a-1} \equiv \left[p_1, p_2, ..., p_{a-1}\right]$ şeklinde tanımlanmaktadır. Yani kısaca (3.27) ve (3.28) denklemleri yukarıda bahsedilen diklik ile ilgili son kısıtı açıklamaktadır. q_a ve r_a için çözüm, P_{a-1}'e ortogonal bir alt uzaya yansıtılan S_0'ın tekil değer ayrışımından bulunan tekil vektörlerin ilk çifti ile elde edilir. S_a çarpım matrisi ise a tane yük vektörü yansıtıldığında aşağıdaki gibi elde edilir (De Jong, 1993)

$$S_a \equiv P_a^{\perp}(X_0'Y_0) = P_a^{\perp} S_0 \qquad a \geq 1 \tag{3.29}$$

S_{a+1}'i ise önceki S_a değerlerinden hesaplamak mümkündür. Bunu yapabilmek için V_a olarak adlandırılan P_a'nın ortonormal tabanına ihtiyaç vardır. V_a , P_a'nın $V_1 = v_1 \propto p_1$ ile başlayan ve (3.30) denklemindeki gibi hesaplanan Gram-Schmidt ortonormalizasyonundan elde edilebilir.

$$v_a \propto p_a - V_{a-1}\left(V'_{a-1}p_a\right) \qquad a = 2,3,4,...,A \tag{3.30}$$

V 'nin ortonormalliğinden yararlanılarak S_a çarpım matrisleri aşağıdaki gibi sürekli olarak indirgenebilir (De Jong, 1993):

$$S_a = S_{a-1} - v_a\left(v'_a S_{a-1}\right) \qquad a > 1 \tag{3.31}$$

SIMPLS ve NIPALS algoritması arasındaki fark, SIMPLS algoritmasında indirgeme işlemi çapraz çarpımlar matrisi S_0'a uygulanırken, daha büyük veri matrisleri X_0 ve Y_0'a uygulanmamaktadır. Her S_a'nın ilk tekil vektör çiftleri iteratif güç metodunu kullanarak hesaplanabilmektedir. Genellikle Y değişkenlerinin sayısı X değişkenlerinin sayısından daha küçüktür. Örneğin Y değişkenlerinin sayısı m iken, X değişkenlerinin sayısı p olsun. Küçük mxm boyutlu simetrik matris $S'_{a-1}S_{a-1}$ 'in baskın özvektörü q_a'nın hesaplanması yeterli olacaktır ve dolayısıyla da r_a aşağıdaki gibi elde edilecektir:

$$r_a \propto S_{a-1}q_a \tag{3.32}$$

Tek değişkenli $Y\left(=y\right)$ için $S_{a-1}\left(=s_{a-1}\right)$, y değişkeni ile X değişkenleri arasındaki kovaryansı gösteren bir vektör olacaktır. Bu sebepten dolayı $S'_{a-1}S_{a-1} = s'_{a-1}s_{a-1}$ skaler, q_a =1 ve $r_a \propto S_{a-1}$'dir. Bu durumda da çözüm, iteratif değildir (De Jong, 1993).

Hem standart NIPALS algoritması hem de SIMPLS algoritmasında X ve Y değişkenlerinin skorlarına ilişkin kovaryans maximize edilir ve ardışık t_a vektörleri $\left(a = 1,2,...\right)$ ortogonal olmak zorundadır. NIPALS-PLS algoritması ile A faktörden sonra

veri matrisi açık bir şekilde $X_A = (I_n - TT')X_0$ 'a, SIMPLS algoritması ile dolaylı bir şekilde $X_0(I_p - VV')$ matrisine indirgenmektedir. Dolayısıyla artık matrisleri farklı olduğundan bu iki algoritma da farklı sonuçlar verir. Aslında çeşitli veri setleri üzerine yapılan uygulama, Y tek değişkenli $(m=1)$ olduğu zaman sonuçların aynı olduğunu, ancak çok değişkenli $(m>1)$ olduğu zaman kısmen de olsa farklılık olduğunu ortaya koymaktadır.

Bu tez çalışmasında tek değişkenli Y ile çalışıldığından dolayı, çok değişkenli Y için algoritmaların nasıl işlediğine burada değinilmeyecektir.

3.5. PCR

PCR, modelin durağan olmayan yapısını gidererek ve regresyon katsayılarının varyanslarını düşürerek çoklu bağlantı problemini ele almak için tanıtılmıştır (Massy, 1965).

PCR tekniğinde ilk olarak temel bileşenler analizi (Principal Component Analysis-PCA) ile bileşenler elde edilir daha sonra temel bileşen skorları bağımsız değişken olarak kullanılarak regresyon analizi yapılır (D'Ambra ve Sarnacchiaro, 2010).

PCR doğru bir şekilde kullanıldığında LS 'den daha iyi tahminler ve daha durağan regresyon katsayıları elde edilir (Ziegel, 2012).

PCR'de değişkenler standartlaştırıldığından $X'X = R$ dir. Burada R bağımsız değişkenler için korelasyon matrisini göstermektedir. PCR analizini göstermek için X 'in tekil değer ayrışımı (SVD-Singular Value Decomposition) aşağıdaki gibi yapılır:

$$X = USP' \tag{3.33}$$

U 'nun sütun vektörleri u 'ların kareleri toplamı bire eşit ve ortogonaldir. Temel bileşenler skor matrisi T ile S arasında $T = US$ bağıntısı yazılabilir. Burada S matrisi, tekil değerlerden oluşan köşegen matristir

U ve y arasındaki regresyon modeli aşağıdaki gibi yazılır:

$$y = \alpha_0 + U\alpha + \varepsilon \tag{3.34}$$

Burada U , X 'in lineer bir dönüşümünü göstermektedir. Alternatif olarak bu denklem, skorlara dayanan regresyon denklemine de dönüştürülebilir ve aşağıdaki gibi gösterilir:

$$y = \alpha_0 + T\gamma + \varepsilon \tag{3.35}$$

Model (3.34) ve (3.35) modelleri aynı uyumu vermektedir, dolayısıyla hata terimleri bu iki modelde de aynıdır (Naes ve Mevik, 2001).

PCR, U veya T 'nin sütunları veya bileşenlerinin bir altseti (bu alt set genellikle büyük özdeğerlere karşılk gelmektedir) üzerine y 'nin regresyonu olarak tanımlanır. Bunu yapmaktaki amaç durağanlığı bozan bileşenleri devreden çıkarmaktır. Yani çok küçük özdeğerler hesaplanması durumunda, çözümü etkileyen çoklu bağlantı problemlerinden kaçınmak için bu özdeğerlere karşılık gelen skor vektörleri atılabilir (Geladi ve Kowalski, 1986).

U_A 'nın $X'X$ 'in A tane en büyük özdeğerine karşılık gelen U 'nun sütunlarından oluşan bir matris olduğu varsayılsın. Bu durumda PCR'ye ilişkin regresyon modeli aşağıdaki gibi yazılır:

$$y = \alpha_0 + U_A\alpha_A + f \tag{3.36}$$

f , yukarıdaki denklemlerde gösterilen ε 'den farklıdır. α_A vektöründeki α 'ların tahminleri LS tekniği ile bulunur. PCR tahmin edicisi $\hat{y}_{PCR}$ aşağıdaki gibi yazılır (Naes ve Mevik, 2001).

$$\hat{y}_{PCR} = \bar{y} + u'_A\hat{\alpha}_A \tag{3.37}$$

u_A 'nın değeri x 'in ilk A tane temel bileşen üzerine yansıtılması ve t skor vektörünün özdeğerlerin karekköküne bölünerek bulunur. Eğer gözlem sayısı değişken sayısından fazla ise PCR'de hesaplanabilecek temel bileşen sayısı maksimum p tanedir ve $A = p$ ise, PCR tahmin edicisi LS tahmin edicisi $\hat{y}$ ile aynı olur. Uygulamada A genellikle çapraz geçerlilik (Cross Validation-CV) ile belirlenir (Naes ve Mevik, 2001).

$\hat{y}_{PCR}$ tahmin edicisinin Varyans (variance), yan (bias) ve Hata Kareler Ortalaması (Mean Square Error-MSE) aşağıdaki gibidir (Tormod ve Harald, 1988):

- $$\text{var}(\hat{y}_{PCR}) = \sigma^2 / N + \sigma^2 \sum_{a=1}^{A} t_a^2 / \lambda_a \tag{3.39}$$

- $$bias(\hat{y}_{PCR}) = -\sum_{a=A+1}^{p} (t_a / \sqrt{\lambda_a})\alpha_a \tag{3.40}$$

- $$MSE(\hat{y}_{PCR}) = \text{var}(\hat{y}_{PCR}) + bias(\hat{y}_{PCR})^2 + \sigma^2$$
$$= \sigma^2 / N + \sigma^2 \sum_{a=1}^{A} t_a^2 / \lambda_a + \left(-\sum_{a=A+1}^{p} (t_a / \sqrt{\lambda_a})\alpha_a \right)^2 + \sigma^2 \tag{3.41}$$

Bu bölümde PLSR ve PCR tekniklerinin teorik yapısından bahsedilmiştir. Bir sonraki bölümde gerçek bir veri seti üzerinde bu iki tekniğin nasıl uygulandığı, hangisinin daha iyi performans gösterdiği incelenecek ve elde edilen sonuç yorumlanacaktır.

4. BULGULAR VE TARTIŞMA

Bu bölümde PLSR ve PCR teknikleri Monte Carlo simülasyonu ile veriye uyum başarıları yönünden karşılaştırılmıştır. Elde edilen sonuçlar değerlendirilirken, RMSECV değerleri ve bunlara karşılık gelen bileşen sayıları dikkate alınmıştır. Ancak simülasyon sonuçlarına ve yorumlarına geçilmeden önce RMSECV hesabı ve RMSECV ile ilgili birkaç kavrama değinilecektir.

Çapraz Geçerlilik (CV) aday altküme regresyon fonksiyonlarını karşılaştırmak için kullanılabilen, istatistiksel analiz sonucunun bağımsız değişken setine nasıl genelleneceğini tespit edebilen bir model geçerliliği yöntemidir.

CV kriterini elde etmek için, *i.* gözlemi hariç tutan bir regresyon fonksiyonundan, *i.* gözlemin tahmin edilen değeri hesaplanır. Yani n gözleme sahip bir örnekte birinci gözlem dışarıda bırakılarak, geriye kalan $n-1$ gözlem üzerinden regresyon modeli tahmin edilir. Dışarıda bırakılan bu gözlem elde edilen regresyon modelinde yerine konularak tahmini değeri hesaplanır. Bu işlem sırasıyla her gözlem için yapılır. Eğer a sayıda gözlemi doğrulamak için geriye kalan $n-a$ tane gözlem model tahmini için kullanılırsa, o zaman yapılan bu işlem a gözlemi hariç tutan CV'yi ifade eder.

Bu çalışmada bir gözlemi hariç tutarak çapraz geçerlilik (LOOCV- Leave-one-out Cross Validation) tekniği kullanılmıştır. LOOCV y_i ile $\hat{y}_{(i)}$ arasındaki bağımsızılığı kullanır. y_i ile $\hat{y}_{(i)}$ arasındaki farkın kareler toplamı da PRESS olarak adlandırılır (Allen, 1974)

$$PRESS = \sum_{i=1}^{n} \varepsilon_{(i)}^2 = \sum_{i=1}^{n} (y_i - \hat{y}_{(i)})^2$$

Burada $\hat{y}_{(i)}$, i'nci gözlem veriden çıkartıldıktan sonra geriye kalan $n-1$ gözlem üzerinden hesaplanır.

PRESS istatistiği, LOOCV yöntemi kullanılarak hesaplanır. RMSECV ise bileşen veya temel bileşen sayısı ile ilişkili olarak hesaplanan ve PRESS istatistiğine dayanan bir kavramdır.

$$RMSECV = \sqrt{\frac{PRESS}{n}}$$

4.1. Simülasyon Çalışması

PLSR ve PCR tekniklerinin her ikisi de veri indirgemesi yaparak çoklu bağlantı problemini ortadan kaldıran tekniklerdir. Dolayısıyla bu tez çalışmasında farklı çoklu bağlantı dereceleri dikkate alınarak, bu iki teknik performansları açısından karşılaştırılmıştır. Çoklu bağlantılı veri üretmek için McDonald ve Galarneau(1975), Gibbons(1981), Kibria (2003) ve diğer araştırmacıların da kullandığı aşağıdaki korelasyonlu veri üretme formulü kullanılmıştır.

$$x_{ij} = (1-\gamma^2)^{1/2} z_{ij} + \gamma z_{ip} \qquad i = 1,2,...,n \qquad j = 1,2,...,p$$

Burada z_{ij} ' ler bağımsız standart normal dağılımdan üretilmiştir. p bağımsız değişkenlerin sayısını, γ ise herhangi iki bağımsız değişken arasındaki korelasyonu göstermektedir. n gözlem için bağımlı değişken değerleri aşağıdaki şekilde belirlenmiştir.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \varepsilon_i \qquad i = 1,2,...,n$$

Burada ε_i ler bağımsız ve sıfır ortalamalı, σ standart sapmalı dağılımdan üretilmiş rassal sayılardır (Ebegil ve Gökpınar, 2012). Yani ilk önce standart normal dağılımdan nxp boyutunda z_{ij} ' ler üretilmiş, daha sonra bunları ve çeşitli korelasyon derecelerini kullanarak x_{ij} 'ler bulunmuştur. Son olarak hata terimleri de eklenerek y değerleri elde edilmiştir. Simülasyon çalışmalarında ortak kısıtlayıcı olarak kullanılan (common restriction) $\sum_{i=1}^{p} \beta_i^2 = 1$ kullanılarak regresyon katsayıları elde edilmiştir.

Bu tez çalışmasında farklı çoklu bağlantı derecelerinin yanı sıra, farklı değişken sayıları, farklı gözlem sayıları ve farklı σ değerleri için de PLSR ve PCR tekniklerinin karşılaştırılması yapılmıştır. Bunun için n=30, 50, 100 ve 150 gözlem ile p=10 ve 20 bağımsız değişken dikkate alınarak simülasyon sonuçları elde edilmiştir. LOOCV yöntemiyle PRESS istatistikleri ve buna bağlı olarak RMSECV değerleri hesaplatılmıştır. Bu işlemler 5000 kez tekrarlanmış ve elde edilen ortalama RMSECV değerleri ile buna karşılık gelen bileşen sayıları tespit edilmiştir.

Çalışmada simülasyon sonuçlarının elde edilebilmesi için Matlab R2012a programı kullanılmıştır.

Aşağıdaki kısımda farklı gözlem sayıları dikkate alınarak elde edilen simülasyon sonuçları farklı çizelgelerde verilmiştir.

Çizelge 5.1. n=30 iken PLSR ve PCR tekniğine ilişkin bileşen sayıları ve RMSECV değerleri

n=30									
		p=10				p=20			
		PLSR		PCR		PLSR		PCR	
σ	γ	Bileşen	RMSECV	Bileşen	RMSECV	Bileşen	RMSECV	Bileşen	RMSECV
0,10	0,50	6	0,1254	10	0,1259	10	0,1857	20	0,1938
	0,70	5	0,1249	10	0,1259	10	0,1829	19	0,1870
	0,85	5	0,1244	9	0,1246	9	0,1783	19	0,1806
	0,95	4	0,1235	9	0,1233	7	0,1697	18	0,1738
0,50	0,50	3	0,6142	3	0,6140	5	0,8069	1	0,7339
	0,70	3	0,6103	9	0,6107	1	0,7300	1	0,6712
	0,85	3	0,6058	8	0,6100	1	0,6262	1	0,6029
	0,95	1	0,5855	1	0,5820	1	0,5550	1	0,5487
1	0,50	3	1,2085	8	1,2067	1	1,2095	1	1,1847
	0,70	3	1,2012	8	1,1934	1	1,1584	1	1,1150
	0,85	1	1,1463	1	1,1244	1	1,1076	1	1,0741
	0,95	1	1,0699	1	1,0620	1	1,0596	1	1,0486

Çizelge 5.1 n=30 gözlem için PLSR ve PCR teknikleriyle indirgenen bileşen sayılarını ve RMSECV değerlerini göstermektedir. Sonuçlar incelendiğinde σ=0.10 ve p=10 iken; düşük, orta ya da yüksek korelasyonlu verilerin RMSECV değerleri hemen hemen aynıdır fakat bu değerlere karşılık gelen bileşen sayıları PLSR'de PCR'den daha

azdır. Ayrıca korelasyon 0.50'den 0'95'e doğru artmaya başlarken RMSECV değerlerinde çok az da olsa azalma görülmektedir. Korelasyonun 0.50 ve 0.70 olduğu durumlarda PCR'nin veride indirgeme yapamadığı ve başlangıçtaki değişken sayısının korunduğu görülmektedir.

Aynı şekilde n=30 ve σ=0.10 iken, değişken sayısı iki katına çıkartıldığında da, yani p=20 değişkenli veri ele alındığında RMSECV değerleri birbirine yakın olmakla birlikte PLSR tekniğinde daha küçük RMSECV değerleri elde edilmiştir. Ayrıca bu durumda bütün korelasyon derecelerinde PLSR tekniği elde edilen bileşen sayısı bakımından PCR'den çok daha iyi sonuç vermiştir. Öyle ki PCR tekniği PLSR'nin iki katı bileşen sayısına sahiptir.

Gözlem sayısı aynı kalıp, p=10 ve σ=0.5 iken PLSR ve PCR'ye ilişkin RMSECV değerleri σ=0.10 iken elde edilen RMSECV değerlerinden daha fazladır, ancak bileşen sayıları özellikle PLSR'de oldukça düşüktür. Ancak σ=0.5 ve korelasyon 0.95 olduğunda PLSR ve PCR veriye uyum başarısı bakımından çok benzer sonuç vermiştir.

n= 30, p=20 ve σ=0.5 iken ise PCR, PLSR'den orta ve yüksek korelasyon durumunda az bir farkla daha iyi sonuç vermiştir. Çünkü RMSECV değerleri daha düşük çıkmıştır. Ancak korelasyonun az olduğu durumda PCR, PLSR'den çok daha iyi sonuç vermiştir. Çünkü PLSR bileşen sayısını 5'e indirgerken, PCR 1'e indirgemiştir.

σ = 1, p=10 ve korelasyonlar yüksek iken PLSR ve PCR tekniğiyle bileşen sayısı açısından aynı sonuçlar elde edilirken PCR'nin RMSECV değerleri daha küçük bulunmuştur. Ancak korelasyon düşmeye başladıkça PLSR daha düşük bileşen sayısı ve benzer RMSECV değeriyle PCR'den çok daha iyi sonuç vermektedir. p=20 iken, yani değişken sayısı arttıkça bütün korelasyon dereceleri için her iki teknik değişken sayısını indirgeme açısından oldukça başarılıdır. Ancak PCR tekniği aynı bileşen sayısında daha düşük RMSECV değeri vermiştir.

Çizelge 5.2. n=50 iken PLSR ve PCR tekniğine ilişkin bileşen sayıları ve RMSECV değerleri

n=50									
p=10						p=20			
		PLSR		PCR		PLSR		PCR	
σ	γ	Bileşen	RMSECV	Bileşen	RMSECV	Bileşen	RMSECV	Bileşen	RMSECV
0,10	0,50	4	0,1127	10	0,1130	6	0,1311	19	0,1308
	0,70	4	0,1123	9	0,1122	6	0,1303	19	0,1299
	0,85	4	0,1124	9	0,1121	5	0,1297	19	0,1299
	0,95	4	0,1121	9	0,1120	5	0,1279	19	0,1299
0,50	0,50	3	0,5566	9	0,5570	4	0,6293	18	0,6323
	0,70	3	0,5553	9	0,5569	4	0,6256	17	0,6291
	0,85	3	0,5565	9	0,5593	4	0,6233	17	0,6169
	0,95	3	0,5550	8	0,5540	1	0,5728	1	0,5691
1	0,50	3	1,1146	9	1,1171	3	1,2409	16	1,2296
	0,70	3	1,1125	8	1,1108	1	1,2163	1	1,1706
	0,85	3	1,1120	8	1,1039	1	1,1217	1	1,0988
	0,95	1	1,0650	1	1,0610	1	1,0528	1	1,0486

Çizelge 5.2'de n=50 gözlem için sonuçlar incelendiğinde σ =0.10 ve p=10 iken PLSR ve PCR'ye ilişkin RMSECV değerleri birbirine çok yakın çıkmakla birlikte incelenen tüm korelasyon derecelerinde bileşen sayısı PLSR'de daha az çıktığından PLSR tekniği çok daha iyi indirgeme yapmıştır. Gözlem sayısı ve σ değiştirilmediğinde, ancak değişken sayısı iki katına (p=20) çıkartıldığında yine çok yakın RMSEV değerleri elde edilmiş, fakat PLSR tekniği bileşen sayısı bakımından PCR tekniğine üstünlük sağlamıştır.

σ =0,50 ve p=10 iken PLSR ve PCR teknikleri ile yakın RMSECV değerine ulaşılmıştır, fakat bileşen sayısına bakıldığında PLSR'nin PCR'den çok daha iyi indirgeme yaptığı görülmektedir. Aynı yorum σ =0,50 ve p=20 iken de yapılabilir fakat korelasyon 0,95 iken bu iki teknik ile aynı bileşen sayısına karşılık aynı çok yakın RMSECV değerleri elde edilmiştir. Böyle bir durumda, bu iki tekniğin aynı performansı gösterdiği söylenebilir.

Standart sapma büyüdükçe yani σ =1 olduğunda p=10 iken bileşen sayısı bakımından PLSR, PCR tekniğinden çok daha üstündür, ancak korelasyon yüksek yani 0,95 iken hem bileşen sayısı hem de RMSECV değeri bakımından aynı sonuçlar elde edilmiştir. Ancak değişken sayısı p=20 olduğunda durum biraz daha farklılaşmaktadır.

Yani çok düşük korelasyonda PLSR yine PCR'den üstünken, korelasyonun artmasıyla birlikte aynı bileşen sayısına karşılık gelen RMSECV değerleri çok yakın, hatta az bir farkla PCR'de PLSR'den daha azdır.

Çizelge 5.3. n=100 iken PLSR ve PCR tekniğine ilişkin bileşen sayıları ve RMSECV değerleri

n=100									
p=10						p=20			
		PLSR		PCR		PLSR		PCR	
σ	γ	Bileşen	RMSECV	Bileşen	RMSECV	Bileşen	RMSECV	Bileşen	RMSECV
0,10	0,50	4	0,1058	9	0,1056	5	0,1120	19	0,1117
	0,70	4	0,1057	9	0,1055	4	0,1119	19	0,1117
	0,85	3	0,1054	9	0,1052	4	0,1116	19	0,1117
	0,95	3	0,1052	9	0,1052	4	0,1113	19	0,1117
0,50	0,50	3	0,5274	9	0,5271	3	0,5529	17	0,5526
	0,70	3	0,5267	9	0,5271	3	0,5529	17	0,5529
	0,85	3	0,5254	9	0,5260	3	0,5517	17	0,5490
	0,95	3	0,5254	8	0,5257	3	0,5510	16	0,5660
1	0,50	3	1,0526	9	1,0519	3	1,1024	17	1,0945
	0,70	3	1,0514	9	1,0521	3	1,1027	16	1,0948
	0,85	3	1,0511	8	1,0494	3	1,1020	16	1,0871
	0,95	2	1,0501	8	1,0459	1	1,0505	1	1,0476

Çizelge 5.3 incelendiğinde p=10 ve 20 iken, ve σ=0.1 olduğunda, tüm korelasyon derecelerinde PLSR çok daha iyi sonuçlar vermiştir. PLSR ve PCR ile çok yakın RMSECV değerleri elde edilmişken buna karşılık gelen bileşen sayısı PLSR tekniğinde çok daha düşüktür.

σ=0,50 olduğunda ve değişken sayıları p=10 ve p=20 iken PLSR tekniği ile elde edilen RMSECV değerleri birbirine oldukça yakındır. Ancak bileşen sayısı PLSR'de PCR'den çok daha azdır.

σ=1 ve p=10 ve 20 iken de benzer yorumlar yapılabilmektedir, ancak korelasyon 0,95 ve p=20 iken aynı bileşen sayısına karşılık RMSECV değerleri her iki teknikte de çok yakındır.

Çizelge 5.4. n=150 iken PLSR ve PCR tekniğine ilişkin bileşen sayıları ve RMSECV değerleri

n=150									
		p=10				p=20			
		PLSR		PCR		PLSR		PCR	
σ	γ	Bileşen	RMSECV	Bileşen	RMSECV	Bileşen	RMSECV	Bileşen	RMSECV
0,10	0,50	3	0,1033	9	0,1032	4	0,1074	19	0,1072
	0,70	3	0,1032	9	0,1032	4	0,1073	19	0,1072
	0,85	3	0,1034	9	0,1034	4	0,1072	19	0,1071
	0,95	3	0,1025	9	0,1025	3	0,1069	18	0,1071
0,50	0,50	3	0,5161	9	0,5156	3	0,5332	18	0,5335
	0,70	3	0,5157	9	0,5156	3	0,5328	17	0,5335
	0,85	3	0,5167	9	0,5170	3	0,5325	17	0,5320
	0,95	3	0,5167	8	0,5168	3	0,5325	16	0,5298
1	0,50	3	1,0348	9	1,0339	3	1,0656	17	1,0620
	0,70	3	1,0339	9	1,0340	3	1,0653	16	1,0610
	0,85	3	1,0337	8	1,0330	3	1,0654	16	1,0580
	0,95	3	1,0337	8	1,0303	1	1,0496	1	1,0479

Çizelge 5.4 incelendiğinde gözlem sayısı n=150 iken PLSR ve PCR teknikleri ile elde edilen RMSECV değerleri ve bunlara karşılık gelen bileşen sayıları görülmektedir.

p=10 olduğunda tüm standart sapma değerleri ve korelasyon derecelerinde PLSR tekniği PCR tekniğinden bileşen sayısı bakımından üstündür yani PLSR tekniği ile PCR tekniğinden daha az bileşen elde edilmiştir. p=20 olduğunda yani değişken sayısı iki katına çıkartıldığında da yine PLSR tekniği PCR tekniğine üstündür.

Fakat σ=1, p=20 ve korelasyon 0,95 olduğunda PLSR ve PCR tekniği ile aynı bileşen sayısına karşılık çok yakın RMSECV değerleri elde edilmiştir.

5. SONUÇLAR VE ÖNERİLER

Bu çalışmada özellikle PLSR tekniği üzerinde durulmuş ve bu teknik PCR tekniği ile karşılaştırılmıştır. Tezde ilk olarak PLSR ve PCR tekniklerinin benzerlik ve farklılıklarına değinilmiştir. Geniş bir şekilde literatürde bu teknikler ile ilgili yapılmış çalışmalara değinilmiştir. Ayrıca tekniklerin matematiksel modelleri ve bu teknikleri kullanarak tahmin edicilerin elde edilişi üzerinde durulmuştur. Son olarak normal dağılımdan veri üretilerek Monte Carlo simülasyonu ile bu iki teknik karşılaştırılmıştır.

Literatürde gerçek veri setleri üzerinde yapılan uygulamalarda PLSR ve PCR teknikleri MSE, RMSE, RMSEP ve göreli tahmin hatası kriterlerine göre karşılaştırılmış ve genelde PLSR tekniğinin daha iyi tahmin sonuçları verdiği görülmüştür. Bu tez çalışmasında sözkonusu iki teknik RMSECV kriteri açısından karşılaştırılmıştır.

Bunun dışında literatürde simülasyon ile yapılan çalışmalar da vardır. Benzer şekilde simülasyon çalışmalarında da yukarıdaki kriterler baz alınarak karşılaştırılmalar yapılmıştır. Elde edilen sonuçlara göre, bazı durumlarda PLSR tekniği PCR tekniğinden daha iyi sonuç verirken, bazı durumlarda bu iki teknik benzer sonuçlar vermiştir. Bu tez çalışmasında normal dağılım kullanılarak korelasyonlu veri üretilmiş ve bu veri üzerinde simülasyon çalışması yapılmıştır. Korelasyonlu veri üretilirken literatürde PLSR ve PCR dışındaki Ridge, Liu vb. yanlı tahmin tekniklerinin ele alındığı simülasyon çalışması yapan makalelerdeki formülasyonlar kullanılmıştır. Daha sonra bu iki tekniği karşılaştırabilmek için gözlem sayıları 30, 50, 100 ve 150; bağımsız değişken sayıları 10 ve 20 alınmıştır. Ayrıca korelasyon dereceleri 0.50, 0.70, 0.85 ve 0.95, standart sapmalar 0.1, 0.5 ve 1 olarak alınmıştır.

Bu iki tekniğin simülasyonla yapılan karşılaştırmalarında MATLAB programı kullanılmıştır. MATLAB kodları yazılırken LOOCV yöntemi kullanılarak PRESS istatistikleri ve bu istatistiklere dayanan RMSECV değerleri elde edilmiştir. Bu işlemler 5000 kez tekrar edilmiştir ve 5000 tekrar sonucunda elde edilen ortalama RMSECV değerleri ve bunlara karşılık gelen bileşen sayıları çizelgelerde gösterilmiştir.

Gözlem sayısı 30 iken standart sapmanın küçük olması durumunda (σ = 0,1), değişken sayısına bağlı olmaksızın PLSR tekniğinin PCR tekniğine üstün olduğu görülmüştür. Ancak yine 30 gözlem için standart sapma büyüdükçe (σ = 1) ve değişken sayısı arttıkça (p=20) bu iki teknik aynı bileşen sayılarında hemen hemen aynı RMSECV değerlerini vermiştir. Dolayısıyla böyle bir durumla karşılaşıldığında bu iki tekniğin birbirine üstün olmadığı yorumu yapılabilir.

Gözlem sayısı arttıkça (n=50, 100 ve 150 için) küçük standart sapma değeri (σ = 0,1) için, bütün korelasyon derecelerinde PCR tekniği değişken sayısını indirgemede başarısızken, PLSR'nin çok iyi bir performans sergilediği görülmüştür. Bu durumda veri boyutunu indirgemede PLSR tekniğinin PCR tekniğine daha üstün olduğu söylenebilir.

Bütün gözlem değerlerinde standart sapma 0,1'den 1'e doğru artma gösterdiğinde RMSECV değerlerinin de arttığı görülmektedir.

Ancak gözlem sayısı ne olursa olsun, standart sapmanın 1'e eşit olduğu, korelasyonun 0,95 ve bağımsız değişken sayısının 20 olduğu durumda PLSR ve PCR tekniğinin aynı sonucu verdiği görülmektedir. Hatta çok az bir farkla PCR'nin PLSR'den daha düşük RMSECV değerine aynı bileşen sayısı ile ulaştığı görülmektedir. Buna göre değişkenliğin ve değişken sayısının fazla, korelasyonun yüksek olduğu verilerle çalışıldığında değişken sayısını indirgeme açısından iki teknik arasında fark olmadığı hatta RMSECV değerleri açısından PCR'nin çok az da olsa bir üstünlüğünün olduğu söylenebilir.

Literatürdeki hem gerçek veri ile hem de simülasyon ile yapılan çalışmaların tümü göz önünde bulundurulduğunda, ve bu tez çalışmasında elde edilen sonuçlar değerlendirildiğinde genel olarak PLSR tekniğinin PCR tekniğine göre daha az bileşen kullandığı, dolayısıyla veri boyutunu daha iyi indirgediği söylenebilir.

Bu simülasyon çalışmasının yapılmasındaki amaç olabildiğince az bileşen sayısı ve düşük RMSECV değeri elde ederek veriyi indirgemektir. Daha iyi genellemelere ulaşabilmek amacıyla standart sapmanın daha büyük ve çok daha küçük değerleri ile çok daha düşük ve yüksek korelasyonlu veriler ele alınabilir. Sadece normal dağılım ile değil,

farklı dağılımlardan da veri üretildiğinde sonuçların ne çıkacağına bakılabilir. Ayrıca LOOCV yönteminin dışında başka CV yöntemleri de kullanılarak sonuçlar karşılaştırılabilir.

KAYNAKLAR DİZİNİ

Abdi, H., 2003, Partial least square regression (PLS regression), Encyclopedia for research methods for the social sciences, 792-795.

Albayrak, A. S., 2012, Çoklu Doğrusal Bağlantı Halinde En Küçük Kareler Tekniğinin Alternatifi Yanlı Tahmin Teknikleri ve Bir Uygulama, Uluslararası Yönetim İktisat ve İşletme Dergisi, 1, 1, 105-126.

Allen, D. M., 1974, The relationship between variable selection and data agumentation and a method for prediction, Technometrics, 16, 1, 125-127.

Almøy, T., 1996, A simulation study on comparison of prediction methods when only a few components are relevant, Computational statistics & data analysis, 21, 1, 87-107.

Bulut, E., 2010, Model Selection Methods For Multivariate Linear Partial Least Squares Regression, Doktora Tezi, Dokuz Eylül Üniversitesi, 57 s.

Cheng, B., Wu, X., 2006, A modified PLSR method in prediction, J. Data Science, 4, 257-274.

Colbert, J. J., Hicks Jr, R. R., Schuckers, M. E., 2002, Coping with multicollinearity: An example on application of principal components regression in dendroecology.

D'Ambra, A., Sarnacchiaro, P., 2010, Some data reduction methods to analyze the dependence with highly collinear variables: A simulation study, Asian J. Math. Stat, 3, 2, 69-81.

De Jong, S., 1993, SIMPLS: an alternative approach to partial least squares regression, Chemometrics and intelligent laboratory systems, 18, 3, 251-263.

Diaz, T. G., Guiberteau, A., Burguillos, J. O., Salinas, F., 1997, Comparison of chemometric methods: derivative ratio spectra and multivariate methods (CLS, PCR and PLS) for the resolution of ternary mixtures of the pesticides carbofuran carbaryl and phenamifos after their extraction into chloroform.Analyst, 122, 6, 513-517.

Dinç, E., 2007, Kemometri Cok Degiskenli Kalibrasyon Yöntemleri, Hacettepe Üniversitesi Eczacılık Fakültesi Dergisi, 27, 61-92.

Duran, E. A., 2011, Regresyon analizinde coklu bağlantı: Parametrik ve semiparametrik tahmin, Doktora Tezi, Gazi Üniversitesi Fen Bilimleri Enstitüsü İstatistik Anabilim Dalı, Ankara.

Ebegil, M., Gökpinar, F., Ekni, M., 2006, A Simulation Study of some Shrinkage Estimators. Hacettepe Journal of Mathematics and Statistics, 35, 2.

Ebegil, M., Gökpınar, F., 2012, A test statistic to choose between Liu-type and least-squares estimator based on mean square error criteria, Journal of Applied Statistics, 39, 10, 2081-2096.

KAYNAKLAR DİZİNİ (devam)

Frank, L. E., Friedman, J. H., 1993, A statistical view of some chemometrics regression tools. Technometrics, 35, 2, 109-135.

Garthwaite, P. H., 1994, An interpretation of partial least squares. Journal of the American Statistical Association, 89, 425, 122-127.

Geladi, P., Kowalski, B. R., 1986, Partial least-squares regression: a tutorial. Analytica chimica acta, 185, 1-17.

Gibbons, D. G., 1981, A simulation study of some ridge estimators, Journal of the American Statistical Association, 76, 131-139.

Guiteras, J., Beltran, J. L., Ferrer, R., 1998, Quantitative multicomponent analysis of polycyclic aromatic hydrocarbons in water samples. Analytica chimica acta, 361, 3, 233-240.

Gujarati, D. N., 2012, Basic econometrics, Tata McGraw-Hill Education.

Helland, I. S., 1988, On the structure of partial least squares regression.Communications in statistics-Simulation and Computation, 17, 2, 581-607.

Hemmateenejad, B., Akhond, M., Samari, F., 2007, A comparative study between PCR and PLS in simultaneous spectrophotometric determination of diphenylamine, aniline, and phenol: effect of wavelength selection.Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 67, 3, 958-965.

Höskuldsson, A., 1988, PLS Regression Methods.Journal of Chemometrics ,2 , 211-228.

Irfan, M., Javed, M., Raza, M. A., 2013, Comparison of shrinkage regression methods for remedy of multicollinearity problem, Middle-East Journal of Scientific Research, 14, 4, 570-579

Jong, S. D., 1993, PLS fits closer than PCR, Journal of chemometrics, 7, 6, 551-557.

Joshi, H., 2012, Multicollinearity Diagnostics in Statistical Modeling and Remedies to deal with it using SAS, http://www.cytel.com/hubfs/0-library-0/pdfs/SP07.pdf, erişim tarihi: 01.12.2015.

Kaşko, Y., 2007, Çoklu Bağlantı Durumunda İkili (Binary) Lojistik Regresyon Modelinde Gerçekleşen I. Tip Hata ve Testin Gücü, Yüksek Lisans Tezi, Ankara Üniversitesi, Zootekni Anabilim Dalı, 48 s.

Khajehsharifi, H., Pourbasheer, E., Tavallali, H., Sarvi, S., Sadeghi, M., 2014, The comparison of partial least squares and principal component regression in simultaneous spectrophotometric determination of ascorbic acid, dopamine and uric acid in real samples, Arabian Journal of Chemistry.

Khalaf, G., 2013, A Comparison between Biased and Unbiased Estimators in Ordinary Least Squares Regression, Journal of Modern Applied Statistical Methods, 12, 2, 17.

KAYNAKLAR DİZİNİ(devam)

Kibria, B. G., 2003, Performance of some new ridge regression estimators, Communications in Statistics-Simulation and Computation, *32*, 2, 419-435.

Li, B., Morris, J., Martin, E. B., 2002, Model selection for partial least squares regression, Chemometrics and Intelligent Laboratory Systems, 64, 1, 79-89

Li, Y., 2010, A Comparison Study of Principle Component Regression, Partial Least Square Regression and Ridge Regression with Application to FTIR Data.

Lindgren, F., Rännar, S., 1998, Alternative partial least-squares (PLS) algorithms, In 3D QSAR in Drug Design, pp. 105-113, Springer Netherlands.

Luinge, H. J., Hop, E., Lutz, E. T. G., Van Hemert, J. A., De Jong, E. A. M., 1993, Determination of the fat, protein and lactose content of milk using Fourier transform infrared spectrometry, Analytica chimica acta, 284, 2, 419-433.

Mahesh, S., Jayas, D. S., Paliwal, J., White, N. D. G., 2015, Comparison of Partial Least Squares Regression (PLSR) and Principal Components Regression (PCR) Methods for Protein and Hardness Predictions using the Near-Infrared (NIR) Hyperspectral Images of Bulk Samples of Canadian Wheat.Food and Bioprocess Technology, 8,1, 31-40.

Maitra, S., Yan, J., 2008, Principle component analysis and partial least squares: Two dimension reduction techniques for regression, Applying Multivariate Statistical Models, 79.

Massy, W. F., 1965, Principal components regression in exploratory statistical research, Journal of the American Statistical Association, 60, 309, 234-256.

McDonald, G. C., Galarneau, D. I., 1975, A monte carlo evaluation of some Ridge-type estimators, Journal of the American Statistical Association, 70, 407-416.

Montgomery, D. C., Peck, E. A., Vining, G. G., 2012, Introduction to linear regression analysis , 821, John Wiley & Sons.

Naes, T., Martens, H., 1985, Comparison of prediction methods for multicollinear data, Communications in Statistics-Simulation and Computation,14, 3, 545-576.

Næs, T., Mevik, B. H., 2001, Understanding the collinearity problem in regression and discriminant analysis, Journal of Chemometrics, 15, 4, 413-426.

Ni, Y., Gong, X., 1997, Simultaneous spectrophotometric determination of mixtures of food colorants, Analytica Chimica Acta, 354, 1, 163-171.

Rawlings, J. O., Pantula, S. G., Dickey, D. A., 1998, Applied regression analysis: a research tool, Springer Science & Business Media.

Rosipal, R., Krämer, N., 2006, Overview and recent advances in partial least squares. In Subspace, latent structure and feature selection, pp. 34-51, Springer Berlin Heidelberg.

Ryan, T. P., 2008, Modern Regression Methods, John Wiley & Sons, pp. 1-10.

Schumann, S., Nolte, L. P., Zheng, G., 2013, Comparison of partial least squares regression and principal component regression for pelvic shape prediction. Journal of biomechanics, 46, 1, 197-199.

KAYNAKLAR DİZİNİ (devam)

Thomas, E. V., Haaland, D. M., 1990, Comparison of multivariate calibration methods for quantitative spectral analysis, Analytical Chemistry, 62, 10, 1091-1099.

Tobias, R. D., 1995, An introduction to partial least squares regression, In Proc. Ann. SAS Users Group Int. Conf., 20th, Orlando, FL ,pp. 2-5.

Tormod, N., Harald, M., 1988, Principal component regression in NIR analysis: viewpoint, background details and selection of components. J. Chemometr, 2, 155-167.

Vigneau, E., Bertrand, D., Qannari, E. M., 1996, Application of latent root regression for calibration in near-infrared spectroscopy, Comparison with principal component regression and partial least squares, Chemometrics and intelligent laboratory systems, 35, 2, 231-238.

Vigneau, E., Devaux, M. F., Qannari, E. M., Robert, P., 1997, Principal component regression, ridge regression and ridge principal component regression in spectroscopy calibration. Journal of chemometrics, 11, 3, 239-249.

Wentzell, P. D., Montoto, L. V., 2003, Comparison of principal components regression and partial least squares regression through generic simulations of complex mixtures, Chemometrics and intelligent laboratory systems, 65, 2, 257-279.

Wichern, D. W., Churchill, G. A., 1978, A comparison of ridge estimators.Technometrics, 20, 3, 301-311.

Wold, S., Ruhe, A., Wold, H., Dunn, III, W. J., 1984, The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. SIAM Journal on Scientific and Statistical Computing, 5, 3, 735-743.

Wold, S., Sjöström, M., Eriksson, L., 2001, PLS-regression: a basic tool of chemometrics, Chemometrics and intelligent laboratory systems, 58, 2, 109-130.

Yaroshchyk, P., Death, D. L., Spencer, S. J., 2012, Comparison of principal components regression, partial least squares regression, multi-block partial least squares regression, and serial partial least squares regression algorithms for the analysis of Fe in iron ore using LIBS, Journal of Analytical Atomic Spectrometry, 27, 1, 92-98.

Yeniay, O., Goktas, A., 2002, A comparison of partial least squares regression with other prediction methods, Hacettepe Journal of Mathematics and Statistics, 31, 99, 99-101

Zeng, X. Q., Li, G. Z., Wu, G. F., Zou, H. X., 2007, On the number of partial least squares components in dimension reduction for tumor classification. In Emerging Technologies in Knowledge Discovery and Data Mining , pp. 206-217, Springer Berlin Heidelberg.

Ziegel, E. R., 2012, A User-Friendly Guide to Multivariate Calibration and Classification. Technometrics.