David J. Olive

High Dimensional Statistics: an Asymptotic Viewpoint

July 31, 2024

Springer

Preface

Many statistics departments offer a one semester graduate course in high dimensional statistics using texts such as Bülmann and van de Geer (2011), Giraud (2022), Lederer (2022), or Wainwright (2019). Statistical learning texts are also used. See Hastie et al. (2009), Hastie et al. (2015), and James et al. (2021). Also see Fujikoshi, Ulyanov, and Shimizu (2010), Koch (2014), Olive (2023e), and Rish and Grabarnik (2015).

High dimensional statistics are used when $n < 5p$ where n is the sample size and p is the number of predictors p. Consider the multiple linear regression model $Y_i = \alpha + \boldsymbol{x}_i^T \boldsymbol{\beta} + e_i = \alpha + x_{i1}\beta_1 + \cdots + x_{ip}\beta_p + e_i$ for $i = 1, ..., n$. Let the full model use all p predictors with $\boldsymbol{\beta} = \boldsymbol{\beta}_F$. In low dimensions where $n \geq 10p$, often $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ is estimated by $\hat{\boldsymbol{\Sigma}} = \hat{\sigma}^2 \hat{\boldsymbol{C}}^{-1}$ where the errors e_i have variance $V(E_i) = \sigma^2$ and where the inverse matrix $\hat{\boldsymbol{C}}^{-1}$ does not exist if $p > n$. Much of the high dimensional literature seeks bounds on the Euclidean norm $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|$. However, if $\hat{\boldsymbol{\beta}}$ is a $\sqrt{n}$ consistent estimator of $\boldsymbol{\beta}_F$, then $\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i$ is proportional to $1/\sqrt{n}$. Hence $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2$ is proportional to p/n which tends to be large when $p >> n$. Similar results hold for estimators $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ for statistical models that depend on a $p \times 1$ vector of parameters $\boldsymbol{\theta}$. Often the high dimensional literature imposes regularity conditions, **that are much too strong**, to force $\|\hat{\boldsymbol{\beta}}_F - \boldsymbol{\beta}_F\|$ to be small as both n and $p \to \infty$.

This text uses large sample theory = asymptotic theory to justify many of the methods used in the test. Several dimension reduction techniques are used. One technique is to use data splitting and variable selection to choose a model I with k predictors where $n \geq 10k$, and then apply the standard low dimensional inference on the resulting model. This changes the high dimensional problem into a low dimensional problem. Sometimes we use the strong assumption that the cases $(\boldsymbol{x}_i, Y_i)^T$ are independent and identically distributed (iid). Then variable selection methods often work because the conditional distribution $Y|\boldsymbol{x}_I^T\boldsymbol{\beta}_I$ has much more information than the marginal distribution for Y.

A second technique is to use large sample theory such that $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ is estimated by $\hat{\boldsymbol{\Sigma}} = \hat{\boldsymbol{C}}$ where the inverse matrix $\hat{\boldsymbol{C}}^{-1}$ is not used. Then tests and confidence intervals for quantities that only use a few of the parameters, such as θ_i or $\theta_i - \theta_k$ can be derived. Hence low dimensional quantities are tested.

A third technique is to replace $\boldsymbol{\theta}$ by the norm $\|\boldsymbol{\theta}\|$ or $\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2$ by the norm $\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|$, reducing the p-dimensional problem of testing $H_0 : \boldsymbol{\theta} = \mathbf{0}$ or $H_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$ to the one-dimensional problem of testing $H_0 : \|\boldsymbol{\theta}\| = 0$ or $H_0 : \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\| = 0$.

The prerequisite for this text is a calculus based course in statistics at the level of Chihara and Hesterberg (2011), Hogg, Tanis, and Zimmerman (2020), Larsen and Marx (2011), Wackerly, Mendenhall and Scheaffer (2008) or Walpole, Myers, Myers and Ye (2016). Linear algebra and one computer programming class are essential. Knowledge of regression would be useful. See Olive (2017a) and Cook and Weisberg (1999). Knowledge of multivariate analysis would be useful. See Olive (2017b) and Johnson and Wichern (2007).

Some highlights of this text follow.

- Prediction intervals are given that can be useful even if $n < p$.
- The response plot is useful for checking the model.
- The large sample theory for the elastic net, lasso, and ridge regression is greatly simplified.
- The large sample theory for some data splitting estimators, variable selection estimators, marginal maximum likelihood estimators, and one component partial least squares will be given. See Olive and Zhang (2023), Olive et al. (2024), and Rathnayake and Olive (2023).

Downloading the book's R functions *slpack.txt* and data files *sldata.txt* into *R*: The commands

```
source("http://parker.ad.siu.edu/Olive/hdpack.txt")
source("http://parker.ad.siu.edu/Olive/hddata.txt")
```

The *R* software is used in this text. See R Core Team (2020). Some packages used in the text include `glmnet` Friedman et al. (2015), `leaps` Lumley (2009), `MASS` Venables and Ripley (2010), and `pls` Mevik et al. (2015).

Acknowledgements

Teaching the material to Math 583 students at Southern Illinois University in 2023 was very useful. Trevor Hastie's website had a lot of useful information. Work by R. Dennis Cook and his coauthors was useful for figuring out OPLS.

Contents

Chapter 1
Introduction

This chapter provides a preview of the book, and some techniques useful for visualizing data in the background of the data are given in Section 1.2. Sections 1.3 and 1.7 review the multivariate normal distribution and multiple linear regression. Section 1.4 suggests methods for outlier detection. Some large sample theory is presented in Section 1.5, and Section 1.6 covers mixture distributions.

1.1 Overview

Statistical Learning could be defined as the statistical analysis of multivariate data. Machine learning, data mining, analytics, business analytics, data analytics, and predictive analytics are synonymous terms. The techniques are useful for Data Science and Statistics, the science of extracting information from data. The *R* software will be used. See R Core Team (2020).

Let $\boldsymbol{z} = (z_1, ..., z_k)^T$ where $z_1, ..., z_k$ are k random variables. Often $\boldsymbol{z} = (Y, \boldsymbol{x}^T)^T$ where $\boldsymbol{x}^T = (x_1, ..., x_p)$ is the vector of predictors and Y is the variable of interest, called a response variable. Predictor variables are also called independent variables, covariates, or features. The response variable is also called the dependent variable. Usually context will be used to decide whether $\boldsymbol{z}$ is a random vector or the observed random vector.

Definition 1.1. A **case** or **observation** consists of k random variables measured for one person or thing. The ith case $\boldsymbol{z}_i = (z_{i1}, ..., z_{ik})^T$. The **training data** consists of $\boldsymbol{z}_1, ..., \boldsymbol{z}_n$. A statistical model or method is fit (trained) on the training data. The **test data** consists of $\boldsymbol{z}_{n+1}, ..., \boldsymbol{z}_{n+m}$, and the test data is often used to evaluate the quality of the fitted model.

Following James et al. (2013, p. 30), the previously unseen test data is not used to train the Statistical Learning method, but interest is in how well the method performs on the test data. If the training data is $(\boldsymbol{x}_1, Y_1), ..., (\boldsymbol{x}_n, Y_n)$, and the previously unseen test data is $(\boldsymbol{x}_f, Y_f)$, then particular interest is in the accuracy of the estimator $\hat{Y}_f$ of Y_f obtained when the Statistical Learning method is applied to the predictor $\boldsymbol{x}_f$. The two Pelawa Watagoda and Olive (2021b) prediction intervals, developed in Section 2.2, will be tools for evaluating Statistical Learning methods for the additive error regression model $Y_i = m(\boldsymbol{x}_i) + e_i = E(Y_i|\boldsymbol{x}_i) + e_i$ for $i = 1, ..., n$ where $E(W)$ is the expected value of the random variable W. The multiple linear regression (MLR) model, $Y_i = \beta_1 + x_2\beta_2 + \cdots + x_p\beta_p + e = \boldsymbol{x}^T\boldsymbol{\beta} + e$, is an important special case. Olive, Rathnayake, and Haile (2022) give prediction intervals for parametric regression models such as generalized linear models (GLMs), generalized additive models (GAMs), and some survival regression models.

The estimator $\hat{Y}_f$ is a *prediction* if the response variable Y_f is continuous, as occurs in regression models. If Y_f is categorical, then $\hat{Y}_f$ is a *classification*. For example, if Y_f can be 0 or 1, then $\boldsymbol{x}_f$ is classified to belong to group i if $\hat{Y}_f = i$ for $i = 0$ or 1.

Following Marden (2006, pp. 5,6), the focus of *supervised learning* is predicting a future value of the response variable Y_f given $\boldsymbol{x}_f$ and the training data $(Y_1, \boldsymbol{x}_1), ..., (Y_n, \boldsymbol{x}_1)$. Hence the focus is not on hypothesis testing, confidence intervals, parameter estimation, or which model fits best, although these four inference topics can be useful for better prediction. The focus of *unsupervised learning* is to group $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ into clusters. *Data mining* is looking for relationships in large data sets.

Notation: Typically lower case boldface letters such as $\boldsymbol{x}$ denote column vectors, while upper case boldface letters such as $\boldsymbol{S}$ or $\boldsymbol{Y}$ are used for matrices or column vectors. If context is not enough to determine whether $\boldsymbol{y}$ is a random vector or an observed random vector, then $\boldsymbol{Y} = (Y_1, ..., Y_p)^T$ may be used for the random vector, and $\boldsymbol{y} = (y_1, ..., y_p)^T$ for the observed value of the random vector. An upper case letter such as Y will usually be a random variable. A lower case letter such as x_1 will also often be a random variable. An exception to this notation is the generic multivariate location and dispersion estimator $(T, \boldsymbol{C})$ where the location estimator T is a $p \times 1$ vector such as $T = \overline{\boldsymbol{x}}$. $\boldsymbol{C}$ is a $p \times p$ dispersion estimator and conforms to the above notation.

The main focus of the first three chapters is developing tools to analyze the multiple linear regression (MLR) model $Y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$ for $i = 1, ..., n$. Classical regression techniques use (ordinary) least squares (OLS) and assume $n >> p$, but Statistical Learning methods often give useful results if $p >> n$. OLS forward selection, lasso, ridge regression, marginal maximum likelihood (MMLE), one component partial least squares (OPLS), the elastic net, partial least squares (PLS), and principal component regression (PCR) will be some of the techniques examined. See Chapter 3.