Outline: (Linear) Models in Large Feature Spaces
Review of Concepts and Notation

(F. Chiaromonte)

Introduction to Statistical Learning

Review (Linear and Generalized Linear Models): Chapters 3, 4

High-dimension, considerations: Chapter 6 Section 4

A QUICK REVIEW OF SOME CONCEPTS AND NOTATION CONCERNING LINEAR MODELS

A classical **linear model** expresses a continuous response Y through an additive function comprising

- an intercept (constant)
- p terms (linear in the slope parameters)
- a random error independent from the random mechanisms underlying the terms.

On a generic observation "i":

$$y_i = \beta_o + \beta' x_i + \varepsilon_i$$

Terms: X = $(X_1 \ldots X_p)$' features (predictors) or specified transformations and functions of the features (e.g., powers, products; observable).

Coefficient parameters: β_o intercept, $\beta = (\beta_1 \ldots \beta_p)$' slopes for each term.

Regression function: the conditional expected value $E(Y|X_1 \ldots X_p) = \beta_o + \beta'X$

Independent random error: added to the regression function and assumed to have mean $E(\varepsilon)=0$ and the same variance parameter $var(\varepsilon)=\sigma^2$ on all observations (homoschedasticity).

In order to develop **inferential procedures**, the random error is also often assumed to be Gaussian:

$$\varepsilon \sim N_1(0,\sigma^2)$$

(T-based confidence intervals and tests for the slope coefficients and the mean response, T-based prediction intervals, F-based ANOVA tests).

The observations are assumed to be iid from the joint distribution of $(Y, X_1 \ldots X_p)$. If $X_1 \ldots X_p$ are viewed as fixed (conditioning; not random), the errors are assumed to be iid across observations. Thus one has, for the vector of errors associated to the n observations in the sample:

$$\underline{\varepsilon}_{(n)} \sim N_n(0,\sigma^2 I)$$

Model in **matrix notation**:

$$\underline{Y}_{(n)} = \underline{1}_{(n)}\beta_o + \underline{X}_{(n,p)}\beta + \underline{\varepsilon}_{(n)}$$

$$\text{for simplicity } \underline{Y}_{(n)} = \underline{X}_{(n,p+1)}\beta + \underline{\varepsilon}_{(n)}$$

$$\underline{X}_{(n,p+1)} = (\underline{1}_{(n)}\ \underline{X}_{(n,p)}) \quad \beta = \begin{pmatrix} \beta_o \\ \beta \end{pmatrix}$$

---

* Intercept = “slope” of the constant term 1 (or assume response is centered, intercept = 0)

Estimating model parameters (fitting):

Because of linearity in the coefficient parameters, whatever the terms represent, fitting can be performed through **least squares** with an explicit, close form solution.

An estimate of the error variance is obtained dividing the minimized sum of squared deviations (Residual Sum of Squares; RSS) by the appropriate number of degrees of freedom.

$$\hat{\beta} = \operatorname{argmin} \| \underline{Y} - \underline{X}\beta \|^2 = (\underline{X}'\underline{X})^{-1}\underline{X}'\underline{Y}$$

$$\hat{\sigma}^2 = \frac{1}{n-(p+1)} \| \underline{Y} - \underline{X}\hat{\beta} \|^2 = \frac{1}{n-(p+1)} RSS$$

Implemented in most statistical software packages, including R. See https://stat.ethz.ch/R-manual/R-devel/library/stats/html/lm.html

If the model is correct (i.e., not misspecified, though some of the coefficients may in fact be 0) the LS coefficient estimators, as well as the LS estimator of the error variance, are **unbiased** for the corresponding parameters.

$$E(\hat{\beta}) = \beta \qquad E(\hat{\sigma}^2) = \sigma^2$$

Moreover, again if the model is not misspecified, the LS coefficient estimators are **BLUE**, i.e., the best (most accurate; minimum sampling variance) among unbiased estimators expressed as linear functions of the response observations – **Gauss-Markov Theorem**.

Finally, if one assumes Gaussian errors, the LS coefficient estimators coincide with the **Maximum Likelihood** (ML) estimators – and the ML estimator for the error variance, which is biased, divides the minimized sum of squares by n instead of the degrees of freedom.

Why? The exponent of the Gaussian likelihood is inversely proportional to the sum of squared deviations:

$$L(\beta \mid \underline{Y}; \underline{X}) \propto \exp\left\{-\frac{1}{2\sigma^2} \| \underline{Y} - \underline{X}\beta \|^2\right\}$$

Residuals diagnostics and **remedial measures:**

Lots of techniques that, based on residuals (observable proxies for the unobservable errors) aim to detect departures from

$$\underline{\varepsilon}_{(n)} \sim N_n(0,\sigma^2 I)$$

i.e., Gaussian iid random errors with 0 mean and shared variance σ^2.

Lots of approaches to manipulate the data, e.g.:

- variance stabilizing transformations of the response to address heteroschedasticity
- identification and removal of outliers/influential observations (*case diagnostics*)

and the model specification, e.g.:

- add/remove terms to address mean patterns in the residuals

to come closer to meeting the assumptions.

Also important, **Multicollinearity**:

Linear dependencies among predictors increase the variance in effect estimates, may even make LS solution unstable.

Diagnose through

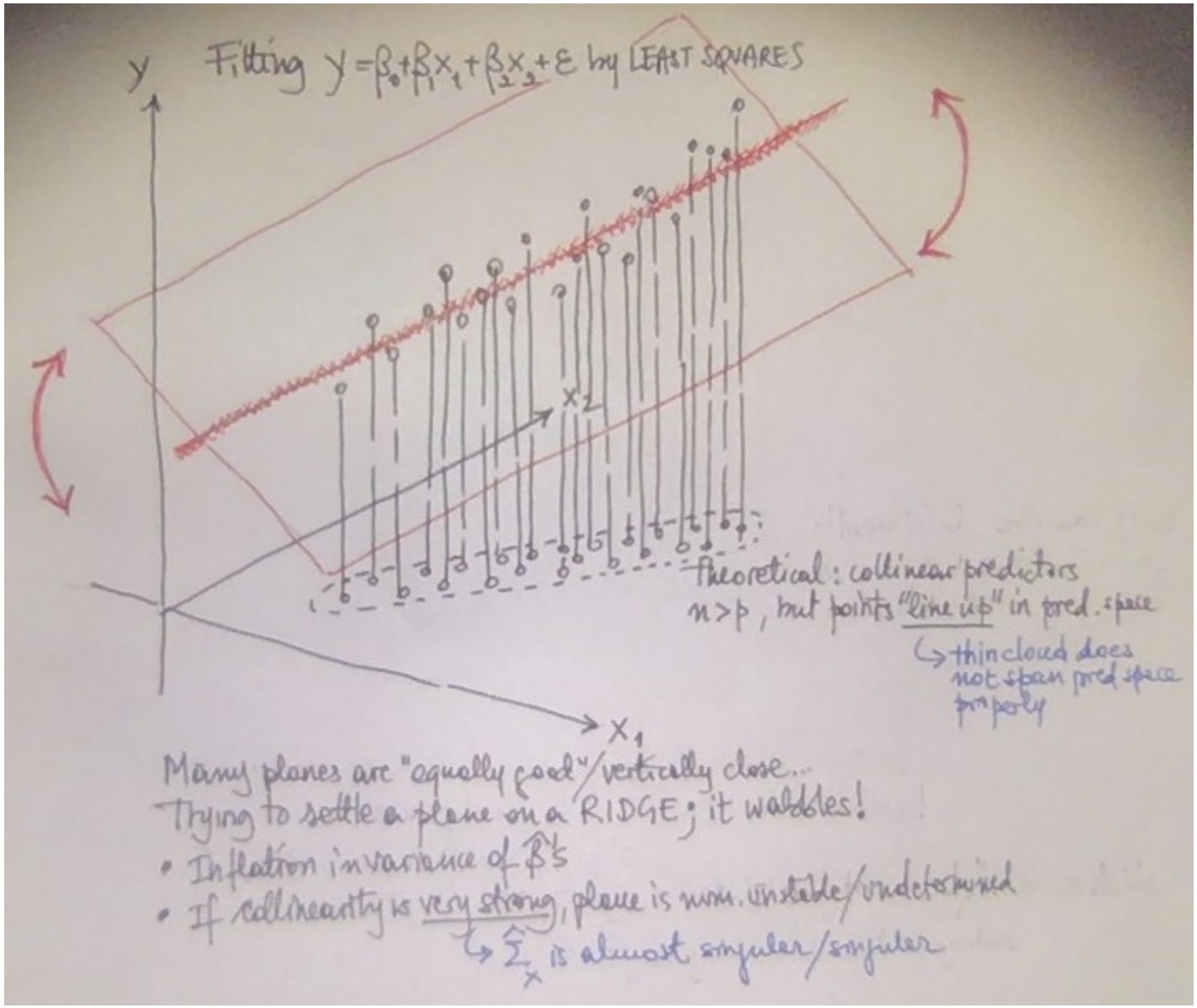
- Pair-wise correlations or scatter-plot matrices
- More complete, **partial R-squares** and **Variance Inflation Factors** (do not depend on the response)

$$R^2_{X_j|\ other\ Xs} \qquad VIF_j = \frac{1}{1 - R^2_{X_j|\ other\ Xs}}$$

Remedies

- stabilize the fit (Ridge)
- eliminate some predictors (LASSO, Best Subsets – feature selection)
- reduce dimension creating "composite" predictors (unsupervised, e.g., Principal Components Analysis; supervised, e.g., Sliced Inverse Regression)

… coming next!


Fitting $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ by LEAST SQUARES
y
x_1
x_2
theoretical: collinear predictors
$n > p$, but points "line up" in pred. space
↳ thin cloud does not span pred. space for poly
Many planes are "equally good"/vertically close...
Trying to settle a plane on a RIDGE; it wobbles!
• Inflation in variance of $\hat{\beta}$'s
• If collinearity is very strong, plane is num. unstable/undetermined
↳ $\hat{\Sigma}_x$ is almost singular/singular

GENERALIZATIONS

- Linear models can be extended to comprise **categorical predictors**, properly encoding (dummies) their main effects and interactions with the continuos predictors.

- In **Generalized Linear Models** link functions are introduced to represent the dependence of a non-continuos response Y on the predictors $X_1 \ldots X_p$, and the stochasticity of $Y \mid X_1 \ldots X_p$ is modeled differently, not through an additive random error:
  - Counts (Poisson regression)
  - Binary labels (Logit or Binomial regression; Probit regression)
  - Multi-class labels (Multinomial regression)

  All (including the standard Normal regression) are implemented in the R package **GLM**
  https://stat.ethz.ch/R-manual/R-devel/library/stats/html/glm.html

  For instance, Logistic regression: $g(x_i) = \ln\left(\frac{p(x_i)}{1-p(x_i)}\right) = \beta_o + \beta' x_i \qquad y_i \sim Bernoulli(p(x_i))$

  Note: $p(X)=E(Y|X_1 \ldots X_p)$. LS estimation is replaced by ML estimation – often implemented numerically; more computation.

- In heteroschedastic cases, and/or when modeling non-iid data (time series, spatial data), special covariance structures for the errors are introduced to represent varying variance and/or correlations among observations: $\underline{\varepsilon}_{(n)} \sim N_n(0, \Sigma(\eta))$.

  LS estimation can be adapted to recover BLUE; **Generalized Least Squares**.
  More sophisticated estimation approaches (based on ML or Bayesian techniques) also exist; more computation necessary.

SOME ADDITIONS: OMITTED VARIABLE BIAS

The OMITTED VARIABLE BIAS (OVB) phenomenon

Important focus for Economics (less so for other fields): Trying to move towards causation, parsing controllable/exogenous and endogenous features. Related emphasis on model misspecification. Suppose an omitted feature Z

(i) affects the response (non-zero coefficient in the true model), and

(ii) correlates with one or more of the X's in the working model (non-zero Cov(X,Z))

then:

$$\underline{Y} = \underline{X}\beta^* + Z\beta_Z + \underline{\varepsilon}^*$$

$$\underline{\varepsilon}^* : \quad E(\underline{\varepsilon}^*) = 0, Cov(\underline{\varepsilon}^*) = \theta^2 I, \text{ indep of X,Z}$$

true model

$$\underline{Y} = \underline{X}\beta + \underline{\varepsilon}$$

$$\underline{\varepsilon} = Z\beta_Z + \underline{\varepsilon}^* : \quad E(\underline{\varepsilon}) \neq 0, Cov(\underline{\varepsilon}) \neq \sigma^2 I, \text{ not indep of X}$$

working model

$$E(\hat{\beta}) = \beta^* + (\underline{X}'\underline{X})^{-1}\underline{X}'Z\beta_Z$$

BIAS (fct of Cov(X,Z))

- The error of the working model includes Z which correlates with the X's; thus, this error is not independent from the random mechanism underlying the X's.
- The LS coefficient estimators for the features in the working model are biased; they subsume parts of the effects of the omitted Z through its correlations with the X's.
- The bias does not vanish with increasing n – causing inconsistency.

IMPORTANT REMARKS:

- Omitting anything that does not carry interdependencies with X's may induce a bigger and/or non-spherical error variability in the model – but does not introduce a bias in the LS estimators of the coefficients of the X's.

- Including more features in the model reduces the risk of OVB – but inflates the variability of the LS coefficients estimators (more below). *Under-specifying vs overfitting; a variance/bias trade-off*.