An Introduction to *An Introduction to Envelopes*[1]

by R. D. Cook

January 15, 2019

The goal of this introduction is to give an historical overview of dimension reduction that leads naturally to envelopes, relatively new dimension reduction methodology that has the potential to produce substantial efficiency gains relative to standard methods, sometimes by amounts equivalent to increasing the sample size many times over. This goal is in contrast to the book, which begins with envelopes and ends in Chapter 9 with connections to other methods. The same ideas are generally available in the book, although not all in one place and not organized to emphasize aspects of the evolution of dimension reduction. Here, details are kept to a minimum, the focus instead being on ideas and philosophy.

Viewed broadly, dimension reduction is a huge area. It touches on all of the applied sciences, is a leitmotiv of statistics and is an active research area today. While some disciplines seem particularly adept at proliferating new dimension reduction algorithms, the focus of the book is on dimension reduction from a statistical perspective.

1 A little Fisherian history

1.1 Fisher, 1922

The modern era of statistics began with Fisher's seminal 1922 paper entitled "On the mathematical foundations of theoretical statistics." The general notion of reducing data was certainly in the statistical lexicon in 1922, but the community then lacked adequate foundations for effectively advancing complex ideas. For instance, Edgeworth (1884) speaks of "reduction of observations" using "facility-curves" and of pursuing either the "most advantageous" or "most probable" value, which are not terms that we readily recognize today. The lack of an adequate foundation is particularly evident in the history of outliers (e.g. Beckman and Cook, 1983).

Fisher (1922) summarized the state of the discipline in his first section, entitled "The Neglect of Theoretical Statistics":

[1] Cook, R. D. (2018). *An Introduction to Envelopes: Dimension Reduction for Efficient Estimation in Multivariate Statistics*. New York: Wiley.

> In spite of the immense amount of fruitful labour which has been expended in its practical applications, the basic principles of this organ of science are still in a state of obscurity, and it cannot be denied that, during the recent rapid development of practical methods, fundamental problems have been ignored and fundamental paradoxes left unresolved.

In this paragraph, Fisher highlights the lack of an adequate theoretical foundation for the development of statistical methodology. It was his intent of course to propose such a foundation, first stating the purpose of statistical methods in his second section entitled "The Purpose of Statistical Methods":

> A quantity of data, which usually by its mere bulk is incapable of entering the mind, is to be replaced by relatively few quantities which shall adequately represent the whole, or which, in other words, shall contain as much as possible, ideally the whole, of the relevant information contained in the original data.

Here, Fisher sets the goal of statistical methodology to be, using contemporary vernacular, dimension reduction without loss of (much) information. The description is quite broad, leaving no hint of how dimension reduction should be pursued, but it does give an important overarching notion that may be used as a guide. Fisher next sketched the context for his theoretical foundation, although today we recognize that this is one of many different ways of implementing his ideas.

> This object is accomplished by constructing a hypothetical infinite population, of which the actual data are regarded as constituting a random sample. The law of distribution of this hypothetical population is specified by relatively few parameters, which are sufficient to describe it exhaustively in respect of all qualities under discussion. Any information given by the sample, which is of use in estimating the values of these parameters, is relevant information. Since the number of independent facts supplied in the data is usually far greater than the number of facts sought, much of the information supplied by any actual sample is irrelevant. It is the object of the statistical processes employed in the reduction of data to exclude this irrelevant information, and to isolate the whole of the relevant information contained in the data.

In Fisher's framework we start by imagining a hypothetical infinite population from which we obtain a random sample with distribution function that is known up to a relatively few parameters θ.

This is now an everyday description, but in 1922 it was novel. For instance, the conceptual starting point of a "hypothetical infinite population" originated in this paper (Burnside, 1926). Also, here the word "parameter" occurred for the first time with its contemporary meaning (Stigler, 1976). Fisher's notion was that we should reduce the data by retaining the part that contains information about θ. The last sentence can be read informatively as a further commentary on reduction generally: we wish to retain the relevant information in the data and exclude the irrelevant information, where "relevant information" can depend on context without necessarily referencing parameters. Fisher goes on to address a variety of constituent issues, including estimation criteria, and consistency, sufficiency.

1.2 Fisher's idea of sufficiency

According to Fisher we should start the statistical process of extracting relevant information from a sample $\mathcal{D} = \{Y_1, \ldots, Y_n\}$ by specifying the underling law $F(y|\theta)$, where $\theta \in \Theta$. Then a statistic $t(\mathcal{D})$ is said to be sufficient for θ if

$$\mathcal{D} \mid (t, \theta = \theta_1) \quad \sim \quad \mathcal{D} \mid (t, \theta = \theta_2) \; \forall \theta_1, \theta_2 \in \Theta.$$

This statement is a formal expressions of the idea that t captures all the information about θ that is contained in $\mathcal{D}$. In particular, if we know t then the conditional distribution of $\mathcal{D}$ given t does not depend on θ. This was seen as a brilliant idea at the time and considerable effort was devoted to the study of sufficiency for the next 40 years. But by the 1970's Fisher's sufficiency had generally fallen out of favor as a paradigm for guiding methodological studies because of its dependence on a known model and the increasingly complex nature of models. Today there is relatively little emphasis on Fisherian sufficiency per se, being mostly relegated to theory courses and philosophical discussions as a fundamental concept. Nevertheless, Fisher's idea, which Cox and Mayo (2010) called "reduction by sufficiency," has a clear legacy in dimension reduction, a legacy that serves as an overarching philosophy in the book.

1.3 Sufficient reductions: Fisher's sufficiency legacy

Consider a pair of vectors, $\mathbf{X} \in \mathbb{R}^p$ and $\mathbf{Y} \in \mathbb{R}^r$, at least one of which is stochastic. We wish to reduce the dimension of $\mathbf{X}$ to facilitate a study of dependence. (This notation is not necessarily

meant to imply regression.) The following is a generalization of Definition 9.1 in the book and of Fisher's notion of sufficiency (see also Cook, 2007).

Definition 1 *A reduction $R : \mathbb{R}^p \to \mathbb{R}^q$ $q \leq p$ of $\mathbf{X}$ is sufficient for $\mathbf{Y}$ if at least one of the following hold*

1. $\mathbf{X} \mid (\mathbf{Y} = \mathbf{y}_1, R(\mathbf{X})) \sim \mathbf{X} \mid (\mathbf{Y} = \mathbf{y}_2, R(\mathbf{X}))$, $\forall \mathbf{y}_1, \mathbf{y}_2$ *in the space of* $\mathbf{Y}$,
2. $\mathbf{Y} \mid \mathbf{X} \sim \mathbf{Y} \mid R(\mathbf{X})$,
3. $\mathbf{Y} \perp\!\!\!\perp \mathbf{X} \mid R(\mathbf{X})$.

Statement 1 requires $\mathbf{X}$ to be random, but not $\mathbf{Y}$. If we think of $\mathbf{X}$ as the total data (previously denoted as $\mathcal{D}$) and $\mathbf{Y}$ as a parameter vector then the statement reduces to the requirement for a sufficient statistic; that is, $R(\mathbf{X})$ is then a sufficient statistic. Statement 2 is the classical regression context where the predictors $\mathbf{X}$ are fixed. Here only $\mathbf{Y}$ need be random. Statement 3 requires a joint distribution. In each statement, $R(\mathbf{X})$ is relevant information in Fisher's general sense. They are equivalent if $\mathbf{X}$ and $\mathbf{Y}$ have a joint distribution and then we have considerable flexibility in pursuing a reduction. Sufficient dimension reduction, which is discussed in Chapter 9, is concerned with regressions as in statement 2, but historically has pursued methodology via statement 1. These three statements and related constructions form a foundation for nearly all of the book.

2 Parallel development of principal component analysis

Fisher's philosophy of reduction was not the only one to emerge in the early 1900's. Suppose we observe independent copies $\mathcal{D} = \{\mathbf{Y}_1, \ldots, \mathbf{Y}_n\}$ of a random vector $\mathbf{Y} \in \mathbb{R}^r$ which has a density. How do we proceed if we wish to reduce the number of variables? One paradigm proposed was to reduce the data linearly from dimension r to dimension $q < r$: $\mathcal{D} \mapsto \mathbf{L}^T\mathcal{D} = \{\mathbf{L}^T\mathbf{Y}_1, \ldots, \mathbf{L}^T\mathbf{Y}_n\}$, where $\mathbf{L} = (\boldsymbol{\ell}_1, \boldsymbol{\ell}_2, \ldots, \boldsymbol{\ell}_q)$ is an $r \times q$ matrix. According to Fisher we should construct $\mathbf{L}$ to extract information generally and specifically about parameters. But there are no parameters here and no specific goals have been stated so far. On what principle do we proceed? Historically, many have been willing to take a leap of faith and reason that

$$\text{Relevant Information} \sim \text{Variation},$$

so we want the linear combinations with the most variation as a means of extracting the relevant information. Let $\mathbf{S_Y}$ be the sample covariance matrix of $\mathbf{Y}$. Then the first linear combination is constructed using

$$\boldsymbol{\ell}_1 = \arg\max_{\boldsymbol{\ell}} \frac{\boldsymbol{\ell}^T \mathbf{S_Y} \boldsymbol{\ell}}{\boldsymbol{\ell}^T \boldsymbol{\ell}},$$

which is just the first eigenvector of $\mathbf{S_Y}$. Continuing this process leads to principal component analysis (PCA), probably the most widely used dimension reduction method in the applied sciences. Let $\boldsymbol{\ell}_1, \ldots, \boldsymbol{\ell}_r$ now denote the eigenvectors of $\mathbf{S_Y}$ ordered according to its eigenvalues, so $\boldsymbol{\ell}_1$ is the eigenvector associated with the largest eigenvalue. Then, assuming that $q < n$, in PCA we select the first q eigenvectors $\mathbf{L} = (\boldsymbol{\ell}_1, \boldsymbol{\ell}_2, \ldots, \boldsymbol{\ell}_q)$ and reduce the data accordingly. It is important to recognize that PCA always takes the first q eigenvectors, justifying the name "principal". The origins of PCA goes back at least to Adcock (1878). It is sometimes credited to Karl Pearson but its popularity received a substantial boost from Hotelling (1933). We see then that PCA was being developed around the time that Fisher introduced sufficiency.

A simple thought experiment may illustrate how PCA can be an effective method when it's clear that Information $\sim$ Variation is an effective basis for reduction. Suppose we need to determine the final scores S for students in a statistics class, the final scores to be based solely on the students' scores S_1 and S_2 on two examinations. We might decide to average the scores $S = (1/2)(S_1 + S_2)$ or to use a weighted average $S = (2/5)S_1 + (3/5)S_2$, reasoning that the second exam should be weighted more heavily. Another option is to base the final score on the first principal component $S = (S_1, S_2)\boldsymbol{\ell}_1 / \mathbf{1}_2^T \boldsymbol{\ell}_1$, reasoning that the direction of maximum variation in the scores gives the best univariate information to distinguish the students' performances.

On the other hand, PCA seems often used automatically without a considered argument that variation should preserve information in the application at hand. An instance of this is the use of PCA to the reduce the predictor dimension in regression.

2.1 Principal components in regression

For a time, using PCA to reduce the predictors in regression was common practice, particularly when $n < p$. Mosteller and Tukey (1977, p. 397) addressed the use of principal components in regression by posing the following question.

> ...how can we find linear combinations of the [predictors] that will be likely, or un-likely,

to pick up regression from some as yet unspecified [response]?

They then justified PCA by answering

> A malicious person who knew our [predictors] and our plan for them could always invent a [response] to make our choices look horrible. But we don't believe nature works that way – more nearly that nature is, as Einstein put it (in German), "tricky, but not downright mean."

In short, using PCA in regression requires a leap of faith – faith in nature or faith in the wisdom of Mosteller and Tukey.

But this practice was eventually called into question. Some authors did case studies where they found important minor (not principal) components (e.g. Hadi and Ling, 1998). As a simple example to help fix ideas, suppose we are contemplating using the first principal component to reduce a pair of predictors in a logistic regression of gender on height. The two plots in Figure 2.1, which are versions of Figures 2.1 and 2.2 in the book, illustrate two regressions in which principal components exhibit different utility. Figure 2.1a allows visualization of the regression of gender on

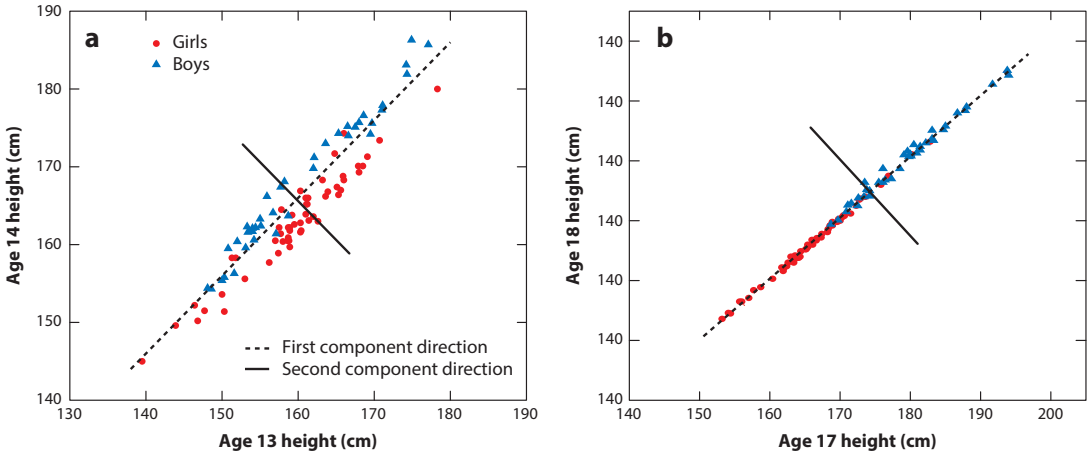

Figure 2.1: Plots of the first two principal components.

heights at ages 13 and 14. Here, projecting the predictors onto their first principal component will lose nearly all of the information on gender. A better procedure in this case would be to neglect the principal component and project the data onto the minor (second) component, but this is not how PCA works. In Figure 2.1b, which describes the regression of gender on heights at ages 17

and 18, projecting the data onto the first principal component will preserve essentially all of the information on gender.

Here we have an interesting historical contrast. Fisher's sufficiency is based on a crisp theoretical argument while principal components are based on the expedient notion that preserving variation somehow preserves information. Many have studied the advantages of principal components using *post hoc* reasoning: we have a method now let's see if it provides something useful. Fisher's approach tells us how we must proceed to achieve a specific theoretical goal. Unrestrained PCA tells us how to proceed but the end result may be equivocal.

2.2 Probabilistic principal components

There was not much Fisherian understanding of principal components offered in the literature until Tipping and Bishop (1999) introduced Probabilistic Principal Components, which are discussed in Section 9.9 of the book. They gave a multivariate model that yields principal components via maximum likelihood estimation. Suppose that $\mathbf{Y}$ can be modeled as $\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{\nu} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\mu} \in \mathbb{R}^r$ and $\boldsymbol{\Gamma} \in \mathbb{R}^{r \times u}$, with $u \leq r$, are non-stochastic and unknown, $\boldsymbol{\nu} \in \mathbb{R}^u \sim N(0, \boldsymbol{\Delta})$ is a vector of latent variables and $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 \mathbf{I}_r)$. The idea here is that the variation in $\mathbf{Y}$ is controlled by two distinct types of variation – extrinsic variation coming from the extrinsic variable $\boldsymbol{\nu}$, which occurs outside of $\mathbf{Y}$ and affects $\mathbf{Y}$ via $\boldsymbol{\Gamma}$, and intrinsic variation that is isotropic noise (e.g. measurement error). The goal is to estimate the part of $\mathbf{Y}$ that is affected by the extrinsic variation. Since $\mathbf{Y} \perp\!\!\!\perp \boldsymbol{\nu} \mid \boldsymbol{\Gamma}^T\mathbf{Y}$ that part is exactly $\boldsymbol{\Gamma}^T\mathbf{Y}$, which corresponds to Fisher's relevant information. Tipping and Bishop showed in effect that the MLE of span($\boldsymbol{\Gamma}$) is the span of the first u eigenvectors of $\mathbf{S}_{\mathbf{Y}}$.

The important point here is that the condition $\mathbf{Y} \perp\!\!\!\perp \boldsymbol{\nu} \mid \boldsymbol{\Gamma}^T\mathbf{Y}$ serves to link PCA with Fisherian reasoning via Definition 1: an appreciation for the benefits and limitations of principal components can come from adopting a Fisherian perspective.

2.3 Discussion

Our use of principal components here is as a representative of dimension reduction methods that are not founded on crisp theoretical principles. The essential point is that firm foundations are essential for good dimension reduction methodology, a seemingly obvious fact that nevertheless is often overlooked, as evidenced by the history of principal components in regression.

The Tipping-Bishop model can be adapted to study predictor reduction in regression by positing

that $\boldsymbol{\nu}$ depends on $\mathbf{Y}$, leading to principal fitted components that arise via the model $\mathbf{X} \mid (\mathbf{Y} = \mathbf{y}) = \boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{\nu}(\mathbf{y}) + \boldsymbol{\varepsilon}$. Principal fitted components are discussed in Chapter 9.

3 Sufficient dimension reduction, SDR

Sufficient dimension reduction, which is discussed in Sections 9.4 to 9.7, was developed primarily for reducing the dimension of the stochastic predictors $\mathbf{X} \in \mathbb{R}^p$ in univariate regression without requiring a parsimonious model, which stands in contrast to Fisher's original setup for sufficient statistics. It is based on Definition 1, restricting the reduction form to be linear, $R(\mathbf{X}) = \boldsymbol{\Gamma}^T\mathbf{X}$. So we pursue the fewest linear combinations $\boldsymbol{\Gamma}^T\mathbf{X}$ of $\mathbf{X}$ so that $Y \perp\!\!\!\perp \mathbf{X} \mid \boldsymbol{\Gamma}^T\mathbf{X}$, without assuming a model for $Y \mid \mathbf{X}$. Of course there are assumptions, but a Fisherian model is not required.

The statement $Y \perp\!\!\!\perp \mathbf{X} \mid \boldsymbol{\Gamma}^T\mathbf{X}$ holds if and only if $Y \perp\!\!\!\perp \mathbf{X} \mid (\boldsymbol{\Gamma}\mathbf{A})^T\mathbf{X}$ for any full rank matrix $\mathbf{A}$. Consequently, we will not be able to estimate $\boldsymbol{\Gamma}$ but only span($\boldsymbol{\Gamma}$). For that reason some prefer to write the condition in the form $Y \perp\!\!\!\perp \mathbf{X}|\mathbf{P}_{\mathcal{S}}\mathbf{X}$, where $\mathcal{S}$ is a subspace of $\mathbb{R}^p$. The intersection of all subspaces for which this holds is called the central subspace and symbolized as $\mathcal{S}_{Y|\mathbf{X}}$. The central subspace has been the target of much theory and methodological inquiry over the past 30-ish years. The first SDR methods were based on the equivalence in Definition 1

$$Y \perp\!\!\!\perp \mathbf{X}|\mathbf{P}_{\mathcal{S}_{Y|\mathbf{X}}}\mathbf{X} \iff \mathbf{X} \mid (Y, \mathbf{P}_{\mathcal{S}_{Y|\mathbf{X}}}\mathbf{X}) \sim \mathbf{X} \mid \mathbf{P}_{\mathcal{S}_{Y|\mathbf{X}}}\mathbf{X}.$$

Li (2018) recently published an excellent monograph on SDR.

SDR does not require a model for $Y \mid \mathbf{X}$ and that can be a distinct advantage. On the other hand, SDR has little to offer when a model is available. For instance, suppose we have a traditional linear regression model, $Y = \alpha + \boldsymbol{\beta}^T\mathbf{X} + \epsilon$, a logistic regression logit$(p) = \alpha + \boldsymbol{\beta}^T\mathbf{X}$ or a Cox model with hazard function $\lambda(t \mid \mathbf{X}) = \lambda_0(t)\exp(\boldsymbol{\beta}^T\mathbf{X})$. In each of these and many other models $Y \perp\!\!\!\perp \mathbf{X} \mid \boldsymbol{\beta}^T\mathbf{X}$ and $\mathcal{S}_{Y|\mathbf{X}} = \text{span}(\boldsymbol{\beta})$. SDR then is of little help because it tells us to do what we would have done anyway – estimate $\boldsymbol{\beta}$.

4 Envelopes

Envelopes are essentially a targeted form of dimension reduction that can be seen as a decedent of SDR. Nearly all of the envelope development so far has been in model-based contexts, but

envelopes apply also in model-free contexts. Studies over the past several years have repeatedly demonstrated that envelope methodology frequently produces substantial efficiency gains relative to standard methods, sometimes by amounts equivalent to increasing the sample size many times over

To motivate envelopes, we return to the driving condition of SDR, $Y \perp\!\!\!\perp \mathbf{X} \mid \mathbf{P}_{\mathcal{S}}\mathbf{X}$. This statement implies that we want to find and keep the relevant information $\mathbf{P}_{\mathcal{S}}\mathbf{X}$ and thus neglect the irrelevant information $\mathbf{Q}_{\mathcal{S}}\mathbf{X}$. But this can be quite difficult in practice if some elements of $\mathbf{X}$ are highly correlated, as often happens. Envelopes are based on the informal premise that we should strive to be really confident that the stuff we discard $\mathbf{Q}_{\mathcal{S}}\mathbf{X}$ is irrelevant. This can be accomplished by adding the requirement $\mathbf{P}_{\mathcal{S}}\mathbf{X} \perp\!\!\!\perp \mathbf{Q}_{\mathcal{S}}\mathbf{X}$, leading to methodology based on the pair of statements

$$Y \perp\!\!\!\perp \mathbf{X} \mid \mathbf{P}_{\mathcal{S}}\mathbf{X} \text{ and } \mathbf{P}_{\mathcal{S}}\mathbf{X} \perp\!\!\!\perp \mathbf{Q}_{\mathcal{S}}\mathbf{X} \iff (Y, \mathbf{P}_{\mathcal{S}}\mathbf{X}) \perp\!\!\!\perp \mathbf{Q}_{\mathcal{S}}\mathbf{X}, \tag{4.1}$$

which implies that the impact of $\mathbf{X}$ on Y is due solely to variation in $\mathbf{P}_{\mathcal{S}}\mathbf{X}$. If $\mathcal{S}$ satisfies these statements then we must necessarily have $\mathcal{S}_{Y|\mathbf{X}} \subseteq \mathcal{S}$, so $\mathcal{S}$ *envelopes* $\mathcal{S}_{Y|\mathbf{X}}$. For the parametric models mentioned at the end of Section 3, this implies that $\boldsymbol{\beta} \in \mathcal{S}$. The intersection of all subspaces $\mathcal{S}$ that satisfy (4.1) is called an *envelope*.

This is where the book begins. Chapter 1 adapts (4.1) for response reduction in multivariate (multi-response) linear regressions and Chapter 4 is on using (4.1) for predictor reduction in multivariate regression. A detailed outline of the book is available in its preface.

References

Adcock, R. J. (1878), 'A problem in least squares', *The Analyst* **5**, 53–54.
URL: *https://archive.org/metadata/jstor-2635758*

Beckman, R. J. and Cook, R. D. (1983), 'Outlier...s', *Technometrics* **25**(2), 119–149.
URL: *https://doi.org/10.1080/00401706.1983.10487840*

Burnside, W. (1926), 'Lvi. on the hypothetical infinite populations of theoretical statistics', *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **1**(3), 670–674.
URL: *https://doi.org/10.1080/14786442608633666*

Cook, R. D. (2007), 'Fisher lecture: dimension reduction in regression.', *Statistical Science* **22**(1), 1–26.
URL: *https://projecteuclid.org/euclid.ss/1185975631*

Cox, D. R. and Mayo, D. G. (2010), II Objectivity and conditionality in frequentist inference, *in* D. G. Mayo and A. Spanos, eds, 'Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability, and the Objectivity and Rationality of Science', Cambridge University Press, Cambridge, pp. 276–304.

Edgeworth, F. Y. (1884), 'On the reduction of observations', *Philosophical Magazine* **17**(104), 135–141.
URL: *http://dx.doi.org/10.1080/14786448408627492*

Fisher, R. A. (1922), 'On the mathematical foundations of theoretical statistics', *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **222**(594-604), 309–368.
URL: *http://rsta.royalsocietypublishing.org/content/222/594-604/309*

Hadi, A. S. and Ling, R. F. (1998), 'Some cautionary notes on the use of principal components regression', *The American Statistician* **52**(1), 15–19.
URL: *https://www.jstor.org/stable/2685559*

Hotelling, H. (1933), 'Analysis of a complex statistical variable into principal components', *Journal of Educational Psychology* **24**(6), 417–441.
URL: *http://dx.doi.org/10.1037/h0071325*

Li, B. (2018), *Sufficient Dimension Reduction: Methods and Applications with R*, Chapman and Hall/CRC press, New York.

Mosteller, F. and Tukey, J. W. (1977), *Data Analysis and Regression: A Second Course in Statistics*, Addison-Wesley, Reading, MA.

Stigler, S. M. (1976), 'Discussion of "On rereading R. A. Fisher," by L. J. Savage', *Annals of Statistics* **4**, 498–500.

Tipping, M. E. and Bishop, C. M. (1999), 'Probabilistic principal component analysis', *Journal of the Royal Statistical Society B* **61**(3), 611–622.
URL: *http://www.jstor.org/stable/2680726*