1992

Data Analytic Applications of Conditional Independence

Jan de Leeuw

Departments of Mathematics and Psychology UCLA

Introduction

In the paper we look at the situation in which the relationship between two vector variables x and y is *mediated* by a third vector variable z. What we have in mind is the following diagram.

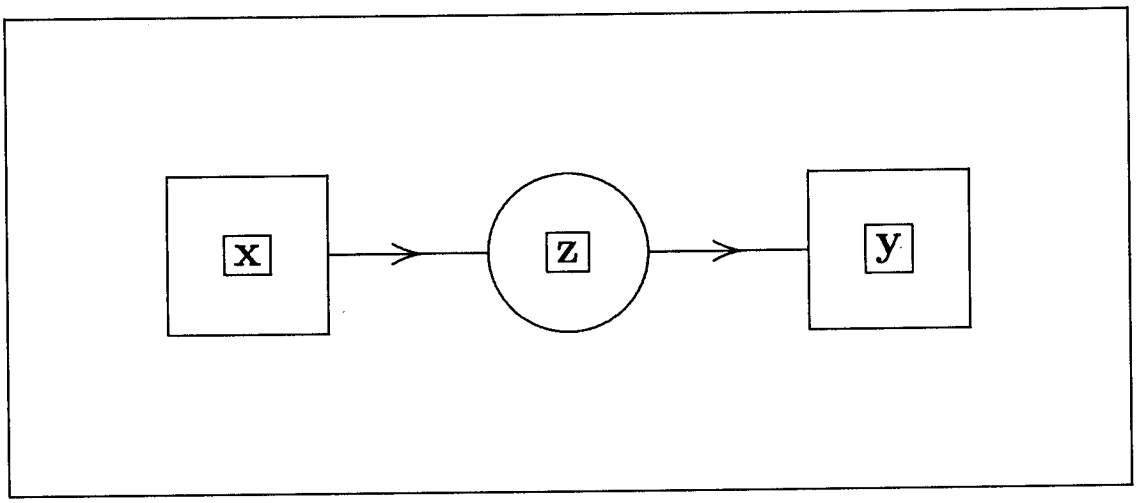

Figure 1: Filtered Regression

There are a couple of basic ideas incorporated in this figure. In the first place the arrows have *direction*, and direction is interpreted loosely as *causality*. Thus z depends on x, and y depends on z. Or: x causes z, and z causes y. The variables in x are *exogeneous* variables, *predictors*, or *input variables* (sometimes, unfortunately, also known as *independent* variables), and the variables in y are *endogeneous*, *outcome*, or *output* variables (again unfortunately sometimes known as *dependent* variables).

Another basic idea incorporated in Figure 1 is that z is unobserved (or *latent*). This is why z is indicated with a circle, while x and y are squares. Variables in z are also known as *mediating* variables, *filters*, *state* variables, *hidden* variables, *constructs*, or *factors*. We are really interested in the dependence of z on x, but unfortunately we cannot observe z directly. We only observe y, which contains error-contamined versions of z, observable indicators for the theoretical construct z, or transformed, censored, truncated, selected, categorized, ranked, incomplete, or otherwise mutilated versions of z. Our hope, of course, is that the dependence of z on x is "really" simple, but since we only observe the dependence of y on x the situation "looks" complicated. The simple relationship $x \rightarrow z$ is contaminated by the transformation $z \rightarrow y$, which can in itself also be simple, but which makes $x \Longrightarrow y$ look complicated. Analysis into two subprocesses hopefully restores the basic simplicity.

The final basic idea is that given z we do not have any dependence of y on x any longer, all dependence of y on x is filtered or channelled through z. There is no arrow which runs directly from x to y. This means that indeed y is just an imperfect or mutilated version of z. In as far as relationship with x is concerned, there is nothing in y that was not already in

z. We shall translated this basic idea into mathematics by using the notion of conditional independence. [Dav79], [Whi90], [Pea88]. Thus our Figure 1 tells us that x and y are conditionally independent given z, which we write as $x \underset{z}{\perp} y$

Some consequences of conditional independence

Throughout the paper we assume that x, y and z have finite means and variances. We shall use the notion of conditional independence to decompose the conditional distribution of y given x into the two supposedly simpler components describing dependence of z on x and of y on z.

The two basic equations are

$$\mathbf{E}(y \mid x) = \mathbf{E}(\mathbf{E}(y \mid z) \mid x), \tag{1}$$
$$\mathbf{V}(y \mid x) = \mathbf{E}(\mathbf{V}(y \mid z) \mid x) + \mathbf{V}(\mathbf{E}(y \mid z) \mid x). \tag{2}$$

If the regression of y on z is linear, then

$$\mathbf{E}(y \mid z) = \mathbf{E}(y) + \mathbf{B}(y \mid z)(z - \mathbf{E}(z)), \tag{3}$$

where $\mathbf{B}(y \mid z) = \mathbf{C}(y, z)\mathbf{V}(z)^{-1}$, and thus

$$\mathbf{E}(y \mid x) = \mathbf{E}(y) + \mathbf{B}(y \mid z)(\mathbf{E}(z \mid x) - \mathbf{E}(z)), \tag{4}$$

With linear regression (2) becomes

$$\mathbf{V}(y \mid x) = \mathbf{E}(\mathbf{V}(y \mid z) \mid x) + \mathbf{B}(y \mid z)\mathbf{V}(z \mid x)\mathbf{B}(y \mid z)'. \tag{5}$$

and with homoscedasticity, in which we have $\mathbf{V}(y \mid z) = \Omega$ independent of z, in addition

$$\mathbf{V}(y \mid x) = \Omega + \mathbf{B}(y \mid z)\mathbf{V}(z \mid x)\mathbf{B}(y \mid z)'. \tag{6}$$

It is useful to emphasize, at this point, that the conditional independence assumption is symmetric in x and y. This means that all formulas in this section remain true if we interchange x and y. We can actually argue from this that conditidional independence, in itself, is not enough to model Figure 1. The direction of the arrows could as well be the other way around, and thus additional structure is needed to give meaning to the particular direction we have chosen. This will be discussed in more detail in later sections.

The Pearson-Aitkin-Lawley Selection Theorem

Suppose $p_0(y,z)$ and $p_1(y,z)$ are two probability density functions, which have different marginals for z, but identical conditional densities for y given z. The idea is that we know $p_1(y,z)$, which is the density *after selection*, and we want to find out about $p_0(y,z)$, the density *before selection*. The results below were discussed first, in the context of natural selection theory, by Karl Pearson [Pe]. The clumsy determinant notation Pearson used was replaced by matrix notation by Aitkin [Ai]. Lawley [La] relaxed the multivariate normality assumptions used by Pearson and Aitkin to linearity of regressions. Birnbaum, Paulsen, and Andrews [Bi] used Lawley's theorem, in a somewhat modernized version, in various psychometric applications. And, finally, Skinner [Sk] discussed a nice geometrical interpretation of the selection theorem, and again gave psychometric applications.

The basic assumption means, in a somewhat different formulation, that there is a variable x, which takes only the values 0 and 1, such that $p(y \mid x \wedge z) = p(y \mid z)$. But this means that y and x are independent given z, and thus the results of the previous section apply. In particular, in the unselected distribution,

$$\mathbf{E}_0(y) = \mathbf{E}_0(\mathbf{E}(y \mid z)), \tag{7}$$

$$\mathbf{V}_0(y) = \mathbf{E}_0(\mathbf{V}(y \mid z)) + \mathbf{V}_0(\mathbf{E}(y \mid z)). \tag{8}$$

If the regression is linear, then these results simplify in the way illustrated in the previous section. Thus we have formulas

The interesting thing is that the unknown parameters $\mathbf{B}(y \mid z)$ and Ω can be computed from the selected population, in which the conditional density is the same as in the unselected one. This leads to the following theorem, which we call the *PAL* theorem.

Theorem 1: In the Pearson-Aitkin-Lawley selection situation, explained above, linearity of regression implies

$$\mathbf{E}_0(y) = \mathbf{E}_1(y) + \mathbf{B}_1(y \mid z)\big[\mathbf{E}_0(z) - \mathbf{E}_1(z)\big], \tag{9}$$

If we assume, in addition, that there is homoscedasticity then

$$\mathbf{V}_0(y) = \mathbf{V}_1(y) + \mathbf{B}_1(y \mid z)\big[\mathbf{V}_0(z) - \mathbf{V}_1(z)\big]\mathbf{B}_1(y \mid z)'. \tag{10}$$

Proof: Simple computation. **Q.E.D.**

Let us give a simple example. We have collected information about the distribution of school achievement tests y in Los Angeles, and we want to know the distribution for the whole country. One way to find out what this distribution is, is to collect information for the whole country. This may be infeasible, and, as usual, in that case we have use assumptions to make up for our lack of empirical information. In this case the assumption

is that we have also collected, in our Los Angeles sample, information on a number of background variables z, such as ethnicity, gender, status, or age, and that $p(y \mid z)$ is the same in Los Angeles as in the rest of the country. If we have a sufficient number of background variables, which we know are important in determining school results, then this assumptions can be quite plausible. If the assumption is true then y and x, which takes the values $\{LosAngeles, USA\}$, are independent given z, and the PAL theorem applies.

The theorem tells us we can infer mean and dispersion of our tests in the whole country from that in Los Angeles, provided we also know $\mathbf{E}_0(z)$ and $\mathbf{V}_0(z)$, i.e. the mean and dispersion of the background variables in the whole country. This information may be available from other sources, such as the census. Of course we still have to assume linearity and homoscedasticity in order to apply the theorem, but this assumption we can check in the Los Angeles data. The critical assumption is conditional independence, and this can, by definition, not be checked. This we have to believe. The usual trade-off also operates here. If we take many covariates in z the assumption will become trivially true, but it can no longer be used in any practical sense, because the conditional distribution cannot be estimated any more. In the social sciences we need many covariates to be plausible, and thus the stability of plausible results will be very poor. If we have only a few covariates we can find stable results, but they will not be plausible.

Selection and Censoring

Now consider the following situation. We have a dependent variable y, which is a transformed, or *censored*, version of a true but unobserved dependent variable z. Thus z depends on x, and y depends on z. The idea is that given z we do not have dependence of y on x any longer, all dependence of y on x is *filtered* through z. Again this can be formalized by using the notion of conditional independence, and by assuming $y \underset{z}{\perp} x$.

The same notion can also be formulated in a way which is closer to classical regression models. Suppose $y = g(z, \epsilon)$, where $\epsilon \perp x$, i.e. ϵ is independent of x. Then obviously $y \mid z \perp x$, and thus $y \underset{z}{\perp} x$.

Selection and Rubin-ignorability

Linear Reduced Rank Regression

Suppose (x, y, z) is a triple of centered random vectors. We think of x as the *predictors*, of y as the *criterium*, and of z as a vector of *unobserved* or *latent* variables, that mediate the dependence of y on x. We suppose x has dimension m, y has dimension n, and z has dimension $p \leq \min(n, m)$. Throughout the paper we assume that both x and z are nondegenerate, in the sense that they are not concentrated on lower-dimensional subspaces. Thus the dispersions $\Sigma_{xx} = \mathbf{V}(x)$ and $\Sigma_{zz} = \mathbf{V}(z)$ are nonsingular.

Reduced rank regression models assume that x and y are conditionally independent given z. We write this as

$$x \underset{z}{\perp} y. \tag{1}$$

It implies that

$$\mathbf{E}(y \mid x) = \mathbf{E}(\mathbf{E}(y \mid z) \mid x), \tag{2a}$$
$$\mathbf{V}(y \mid x) = \mathbf{E}(\mathbf{V}(y \mid z) \mid x) + \mathbf{V}(\mathbf{E}(y \mid z) \mid x). \tag{2b}$$

In addition, *linearity of the regressions* is often assumed. This is

$$\mathbf{E}(y \mid z) = Uz \text{ and } \mathbf{E}(z \mid x) = B'x. \tag{3}$$

As a next step we assume *homoscedasticity*, which is

$$\mathbf{V}(y \mid z) = \Omega \text{ and } \mathbf{V}(z \mid x) = \Theta. \tag{4}$$

Theorem 1: If (2) and (3), then $\mathbf{E}(y \mid x) = UB'x$. If also (4), then $\mathbf{V}(y \mid x) = \Omega + U\Theta U'$.

Proof: Simple computation. **Q.E.D.**

For likelihood inference joint normality of (x, y, z) is assumed, and a set of repeated independent trials (x_i, y_i) is available. Theorem 1 is used to set up the likelihood function, and to estimate the parameters. A diagram illustrating the linear reduced rank regression model is given in Figure 1.

Nonlinear Generalizations: SIR

If we want to generalize the basic structure in the previous section to nonlinear dependence of y on x we can go in a number of directions. We must maintain conditional independence

(1), because it is the very essence of the model. We have to relax some of the linearity and normality assumptions. In the first approach, due to Li [Li1] we make no assumptions about the relationship between y and z. We strengthen $\mathbf{E}(z \mid x) = B'x$ to $z = B'x$, i.e. in terms of (4) we assume $\mathbf{V}(z \mid x) = 0$. Observe that (2a) and $z = B'x$ taken together imply that $\mathbf{E}(y \mid x) = g(B'x)$ for some real g. The work of Li shows that it is possible to estimate B without actually specifying or estimating g. It generalizes earlier work by Goldberger [Go], Brillinger [Br], ...

Theorem 2: Suppose the joint probability density of (x, y, z) satisfies the two *structural assumptions*

$$x \underset{z}{\perp} y, \tag{5a}$$

$$z = B'x \text{ where } B \text{ is } m \times p \text{ of rank } p. \tag{5b}$$

and the *design assumption*

$$\mathbf{E}(x \mid z) = Az. \tag{5c}$$

Then

$$\mathbf{E}(x \mid y) = \Sigma_{xx}B(B'\Sigma_{xx}B)^{-1}B'\mathbf{E}(x \mid y). \tag{6}$$

Proof: We first use (5a). This gives $\mathbf{E}(x \mid y) = \mathbf{E}(\mathbf{E}(x \mid z) \mid y)$. From (5c) we find $\mathbf{E}(x \mid y) = A\mathbf{E}(z \mid y)$, and using (5b) gives $\mathbf{E}(x \mid y) = AB'\mathbf{E}(x \mid y)$. Now, from (5c), $A = \Sigma_{xz}\Sigma_{zz}^{-1}$. Using (5b) finally gives $A = \Sigma_{xx}B(B'\Sigma_{xx}B)^{-1}$. **Q.E.D.**

The same results can be stated somewhat more compactly in terms of normalized scores. Define $\tilde{x} = \Sigma_{xx}^{-1/2}x$ and $\tilde{B} = \Sigma_{xx}^{1/2}B$. Then (6) can also be written as

$$\mathbf{E}(\tilde{x} \mid y) = \tilde{B}(\tilde{B}'\tilde{B})^{-1}\tilde{B}'\mathbf{E}(\tilde{x} \mid y). \tag{7}$$

Observe that $\Pi =_{def} \tilde{B}(\tilde{B}'\tilde{B})^{-1}\tilde{B}'$ is an orthogonal projector, i.e. it is symmetric and idempotent. Let $\tilde{B} = K\Lambda L'$ be the singular value decomposition of $\tilde{B}$, and let $K_\perp$ be the orthogonal complement of K. Thus K is $m \times p$, and $K_\perp$ is $m \times (m-p)$. Also $\Pi = KK'$ and $I - \Pi = K_\perp K_\perp'$.

Theorem 3: Under the conditions of Theorem 2

$$\mathbf{V}(\mathbf{E}(\tilde{x} \mid y))K_\perp = 0, \tag{8a}$$

$$\mathbf{E}(\mathbf{V}(\tilde{x} \mid y))K_\perp = K_\perp. \tag{8b}$$

Proof: From formula (7) it follows that $\mathbf{E}(\tilde{x} \mid y) = KK'\mathbf{E}(\tilde{x} \mid y)$, or $K_\perp'\mathbf{E}(\tilde{x} \mid y) = 0$. This implies $K_\perp'\mathbf{V}(\mathbf{E}(\tilde{x} \mid y)) = 0$. But

$$I = \mathbf{V}(\tilde{x}) = \mathbf{E}(\mathbf{V}(\tilde{x} \mid y)) + \mathbf{V}(\mathbf{E}(\tilde{x} \mid y)). \tag{9}$$

Postmultiplying (9) with $K_\perp$ gives

$$K_\perp = \mathbf{E}(\mathbf{V}(\tilde{x} \mid y))K_\perp.$$

Q.E.D.

The SIR-I algorithm proposed by Li [Li1] is quite simple to understand from Theorem 3. We first estimate $\mathbf{E}(\tilde{x} \mid y)$ by partitioning the range of y into a finite number of intervals, and by taking the averages of the $\tilde{x}$ in each of the intervals. If we use the notation of Gifi [Gf], then the discretization of y gives an indicator matrix G. The diagonal matrix $D = G'G$ gives the sizes of the subgroups. We are interested in the matrix of means $M = \tilde{X}'GD^{-1}$, and we find the column space of $\tilde{B}$ by computing eigenvectors corresponding with the p *largest* eigenvalues of $MDM' = \tilde{X}'GD^{-1}G'\tilde{X}$. These directions are the same as the directions computed in canonical discriminant analysis, i.e. they are in the directions in which the between-group variation is largest with respect to the within-group variation. Alternatively we can use the second part of Theorem 3. This means computing the within-group dispersion matrix in each of the slices, and then by averaging over slices (using weights for the size of the slices). This gives, say, $\tilde{X}'(I - GD^{-1}G')\tilde{X}$, and we use the eigenvectors corresponding with the p *smallest* eigenvalues of this matrix as estimates of the column space of B. Of course the two solutions are identical.

Li's [Li1] SIR-II algorithm takes a different approach. It is based on the following generalization of (8b).

Theorem 4: Under the conditions of Theorem 2

$$K'\mathbf{V}(\tilde{x} \mid y)K_\perp = 0. \tag{10}$$

Proof: Conditional independence (1) is symmetric in x and y. Thus we can interchange x and y in (2b). This gives

$$\mathbf{V}(\tilde{x} \mid y) = \mathbf{E}(\mathbf{V}(\tilde{x} \mid z) \mid y) + \mathbf{V}(\mathbf{E}(\tilde{x} \mid z) \mid y). \tag{11}$$

Now, from the computations in Theorem 2,

$$\mathbf{V}(\mathbf{E}(\tilde{x} \mid z) \mid y) = \mathbf{V}(\Pi\tilde{x} \mid y) = \Pi\mathbf{V}(\tilde{x} \mid y)\Pi.$$

The second part of (11) is somewhat more complicated to handle.

$$\mathbf{V}(\tilde{x} \mid z) = \mathbf{V}(\Pi\tilde{x} + (I - \Pi)\tilde{x} \mid z) = \mathbf{V}((I - \Pi)\tilde{x} \mid z) = (I - \Pi)\mathbf{V}(\tilde{x} \mid z)(I - \Pi).$$

Thus

$$\mathbf{V}(\tilde{x} \mid y) = (I - \Pi)\mathbf{E}(\mathbf{V}(\tilde{x} \mid z) \mid y)(I - \Pi) + \Pi\mathbf{V}(\tilde{x} \mid y)\Pi. \tag{12}$$

Premultiplying (12) by K, and postmultiplying by $K_\perp$, gives (10). **Q.E.D.**

Postmultiplying (9) with $K_\perp$ gives

$$K_\perp = \mathbf{E}(\mathbf{V}(\tilde{x} \mid y))K_\perp.$$

Q.E.D.

The SIR-I algorithm proposed by Li [Li91] is quite simple to understand from Theorem 3. We first estimate $\mathbf{E}(\tilde{x} \mid y)$ by partitioning the range of y into a finite number of intervals, and by taking the averages of the $\tilde{x}$ in each of the intervals. If we use the notation of Gifi [Gif90], then the discretization of y gives an indicator matrix G. The diagonal matrix $D = G'G$ gives the sizes of the subgroups. We are interested in the matrix of means $M = \tilde{X}'GD^{-1}$, and we find the column space of $\tilde{B}$ by computing eigenvectors corresponding with the p *largest* eigenvalues of $MDM' = \tilde{X}'GD^{-1}G'\tilde{X}$. These directions are the same as the directions computed in canonical discriminant analysis, i.e. they are in the directions in which the between-group variation is largest with respect to the within-group variation. Alternatively we can use the second part of Theorem 3. This means computing the within-group dispersion matrix in each of the slices, and then by averaging over slices (using weights for the size of the slices). This gives, say, $\tilde{X}'(I - GD^{-1}G')\tilde{X}$, and we use the eigenvectors corresponding with the p *smallest* eigenvalues of this matrix as estimates of the column space of B. Of course the two solutions are identical.

Li's [Li91] SIR-II algorithm takes a different approach. It is based on the following generalization of (8b).

Theorem 4: Under the conditions of Theorem 2

$$K'\mathbf{V}(\tilde{x} \mid y)K_\perp = 0. \tag{10}$$

Proof: Conditional independence (1) is symmetric in x and y. Thus we can interchange x and y in (2b). This gives

$$\mathbf{V}(\tilde{x} \mid y) = \mathbf{E}(\mathbf{V}(\tilde{x} \mid z) \mid y) + \mathbf{V}(\mathbf{E}(\tilde{x} \mid z) \mid y). \tag{11}$$

Now, from the computations in Theorem 2,

$$\mathbf{V}(\mathbf{E}(\tilde{x} \mid z) \mid y) = \mathbf{V}(\Pi\tilde{x} \mid y) = \Pi\mathbf{V}(\tilde{x} \mid y)\Pi.$$

The second part of (11) is somewhat more complicated to handle.

$$\mathbf{V}(\tilde{x} \mid z) = \mathbf{V}(\Pi\tilde{x} + (I - \Pi)\tilde{x} \mid z) = \mathbf{V}((I - \Pi)\tilde{x} \mid z) = (I - \Pi)\mathbf{V}(\tilde{x} \mid z)(I - \Pi).$$

Thus

$$\mathbf{V}(\tilde{x} \mid y) = (I - \Pi)\mathbf{E}(\mathbf{V}(\tilde{x} \mid z) \mid y)(I - \Pi) + \Pi\mathbf{V}(\tilde{x} \mid y)\Pi. \tag{12}$$

Premultiplying (12) by K, and postmultiplying by $K_\perp$, gives (10). **Q.E.D.**

$$= (K, K_\perp) \begin{pmatrix} K'\tilde{E} - K & 0 \\ 0 & K_\perp' V(\tilde{x}|y) K_\perp \end{pmatrix} \begin{pmatrix} K' \\ K_\perp' \end{pmatrix}$$

Theorem 4 means that for each y there exists a rotation matrix $M(y)$ such that $KM(y)$ are eigenvectors of $\mathbf{V}(\tilde{x} \mid y)$, and there exists another rotation matrix $N(y)$ such that $K_{\perp}N(y)$ are eigenvectors of $\mathbf{V}(\tilde{x} \mid y)$ as well. Computational implications of Theorem 4 are discussed by Li [Li91],[Li90c],[Li90a],[Li90b]. The most interesting possibility seems to be to use a slight variation of the Jacobi-like simultaneous diagonalization algorithm proposed first by De Leeuw and Pruzansky [dLP73].

References

[Ait34] Alexander C. Aitkin. Note on selection from a multivariate normal population. *Proceedings Edinburgh Mathematical Society*, 4:106–100, 1934.

[BPA50] Z.W. Birnbaum, E. Paulson, and F.C. Andrews. On the effect of selection perfomed on some coordinates of a multi-dimensional population. *Psychometrika*, 15:191–204, 1950.

[Bri83] David R. Brillinger. A generalized linear model with gaussian regressor variables. In Peter J. Bickel, Kjell Doksum, and Jr. J.L. Hodges, editors, *A Festschrift for Erich L. Lehmann.* Wadsworth, 1983.

[Dav] A.P. David.

[dLP73] J. de Leeuw and S. Pruzansky. *Psychometrika*, 1973.

[Gif90] Albert Gifi. *Nonlinear Multivariate Analysis.* Wiley, 1990.

[Gol81] Arthur S. Goldberger. Linear regression after selection. *Journal of Econometrics*, 15:357–366, 1981.

[Law44] D.N. Lawley. A note on karl pearson's selection formulae. *Proceedings of the Royal Society of Edinburgh (Mathematics and Physics Section)*, 62:28–30, 1943-1944.

[Li90a] Ker-Chau Li. On principal hessian directions for data visualization and dimension reduction: another application of stein's lemma. Technical report, Department of Statistics UCLA, 1990.

[Li90b] Ker-Chau Li. Sight-seeing with sir: a transformation-based projection pursuit method. Technical report, Department of Statistics UCLA, 1990.

[Li90c] Ker-Chau Li. Uncertainty analysis for mathermatical models with sir. Technical report, Department of Statistics UCLA, 1990.

[Li91] Ker-Chau Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 1991.

[Pea50] Karl Pearson. On the influence of natural selection on the variability and correlation of organs. *Philosophical Transactions (series A)*, 200:1–66, 1950.

[pr]

[Ski84] Chris J. Skinner. The geometric approach to multivariate selection. *Psychometrika*, 49:383–390, 1984.

[wi] J. whittaker.