Analysis of variance using orthogonal projections

Rasmus Waagepetersen

Abstract

The purpose of this note is to show how statistical theory for inference in balanced ANOVA models can be conveniently developed using orthogonal projections. The approach is exemplified for one- and two-way ANOVA and finally some hints are given regarding extension to three- or higher-way ANOVA.

1 Prerequisites on factors and projections

Analysis of variance (ANOVA) models are specified in terms of grouping variables or factors. Suppose observations y_i are indexed by a set I of cardinality n. A factor F is a function that assigns a grouping label among a finite set of labels to each observation. E.g. $F(i) = q$ means that observation y_i (or index i) is assigned to group/level q for the factor F. Suppose F generates k groups. The design matrix Z_F corresponding to F is then $n \times k$ and the iqth entry of Z_F is 1 if i is assigned to group q and 0 otherwise. Note that in many applications, i is a multi-index of the form $i = i_1 i_2 \dots i_p$ for $p \geq 1$.

For any factor F we denote by L_F the column space of Z_F. The orthogonal projection on L_F is denoted P_F. The result of applying P_F to a vector $(y_i)_{i \in I}$ is that y_i is replaced by the average of the observations in the group that i belongs to (i.e. the average of those y_l for which $F(l) = F(i)$).

Two factors play a special role. The unit factor I has a unique level for each observation so $L_I = \mathbb{R}^n$ and $P_I = I$ (with an abuse of notation I is used both for the unit factor and identity matrix). The factor 0 assigns all observations to the same group so $L_0 = \text{span}\{1_n\}$ and $P_0 = 1_n 1_n^\mathsf{T}/n$. Note that $P_0 y$ is simply the vector where each component is given by the average $\bar{y} = \sum_{i \in I} y_i/n$. For any factor F, $L_0 \subseteq L_F \subseteq L_I$.

A factor F is said to be balanced if there is a common number m of observations at each of the k levels (whereby $n = mk$). In this case, the orthogonal projection P_F on L_F is

$$P_F = \frac{1}{m} Z_F Z_F^\mathsf{T}. \tag{1}$$

This result is crucial in the following and the reason why we focus on balanced factors.

2 One-way ANOVA

Consider the model

$$Y_{ij} = \xi + U_i + \epsilon_{ij}, \quad i = 1, \dots, k, \ j = 1, \dots, n_i,$$

where $\xi \in \mathbb{R}$, $U_i \sim N(0, \tau^2)$, $\epsilon_{ij} \sim N(0, \sigma^2)$, $\tau^2, \sigma^2 \geq 0$, and the variables U_i and ϵ_{ij}, are independent $i = 1, \dots, k$ and $j = 1, \dots, n_i$.

Let I consist of indices ij for $i = 1, \dots, k$ and $j = 1, \dots, n_i$, and define the factor F by $F(ij) = i$. Then stacking variables on top of each other we can write the model in vector form as

$$Y = 1_n \xi + Z_F U + \epsilon$$

where n is the total number of observations. As noted in the previous section, Z_F is the design matrix corresponding to F: the ij, qth entry of Z_F is 1 if Y_{ij} belongs to the qth group and zero otherwise. We will assume that F is balanced so that $n_i = m$ for all i.

2.1 Orthogonal decomposition

We now obtain an orthogonal decomposition of $\mathbb{R}^n$:

$$\mathbb{R}^n = V_0 \oplus V_F \oplus V_I$$

where $V_0 = L_0 = \operatorname{span}(1_n)$, $V_F = L_F \ominus V_0$ and $V_I = \mathbb{R}^n \ominus L_F$. Here $\oplus$ denotes sum of orthogonal subspaces:

$$L_1 \oplus L_2 = \{x + y | x \in L_1, y \in L_2\}$$

for orthogonal subspaces L_1 and L_2, and $\ominus$ denotes orthogonal complement:

$$L_2 \ominus L_1 = \{x \in L_2 | x^\mathsf{T} y = 0 \, \forall y \in L_1\}$$

for $L_1 \subseteq L_2$. The dimensions of V_0, V_F and V_I are 1, $k-1$ and $n-k$, and the orthogonal projections on V_0, V_F and V_I are $Q_0 = P_0$, $Q_F = P_F - P_0$ and $Q_I = I - P_F$(see exercise 1).

Using the orthogonal projections we also obtain an orthogonal decomposition of the data vector:

$$Y = Q_0 Y + Q_F Y + Q_I Y$$

into components falling in V_0, V_F and V_I. The covariance matrix is decomposed as:

$$\mathbb{C}\text{ov}Y = \mathbb{C}\text{ov}Z_F U + \mathbb{C}\text{ov}\epsilon = m\tau^2 P_F + \sigma^2 I = \lambda P_F + \sigma^2 Q_I$$

where $\lambda = m\tau^2 + \sigma^2$. Here we used (1) which gives $\mathbb{C}\text{ov}Z_F U = \tau^2 Z_F Z_F^\mathsf{T} = \tau^2 m P_F$ and $I = P_F + Q_I$. Note that there is a one-to-one correspondence between the pairs (λ, σ^2) and (τ^2, σ^2).

The components $Q_0 Y$, $Q_F Y$ and $Q_I Y$ are independent since their covariances are zero. For example

$$\mathbb{C}\text{ov}(Q_0 Y, Q_F Y) = Q_0 \Sigma Q_F = Q_0(\lambda P_F + \sigma^2 Q_I) Q_F$$

which is zero since $P_F = Q_0 + Q_F$ and all products $Q_0 Q_F = Q_0 Q_I = Q_F Q_I = 0$.

Since

$$\mathbb{C}\text{ov}Y = \lambda P_F + \sigma^2 Q_I = \mathbb{C}\text{ov}P_F Y + \mathbb{C}\text{ov}Q_I Y$$

we also consider in the next section the coarser decomposition

$$Y = P_F Y + Q_I Y$$

of Y into independent components $P_F Y$ and $Q_I Y$ falling in L_F and V_I.

2.2 Estimation using factorization of likelihood

Note $P_F 1_n = Q_0 1_n = 1_n$ and $Q_I 1_n = 0$. Moreover $Q_I P_F = 0$. Hence

$$\begin{bmatrix} P_F \\ Q_I \end{bmatrix} Y \sim N\left(\begin{pmatrix} 1_n \xi \\ 0_n \end{pmatrix}, \begin{bmatrix} \lambda P_F & 0 \\ 0 & \sigma^2 Q_I \end{bmatrix} \right).$$

We can thus base maximum likelihood estimation of (ξ, λ) on $P_F Y$ and maximum likelihood estimation of σ^2 on $Q_I Y$.

More precisely,

$$|\Sigma|^{-1/2} \exp(-\frac{1}{2}(Y - 1_n\xi)^{\mathsf{T}}\Sigma^{-1}(Y - 1_n\xi)) =$$
$$\lambda^{-k/2} \exp(-\frac{1}{2\lambda}\|P_F Y - 1_n\xi\|^2) \times (\sigma^2)^{-k(m-1)/2} \exp(-\frac{1}{2\sigma^2}\|Q_I Y\|^2)$$

where we used $\Sigma^{-1} = \sigma^{-2}Q_I + \lambda^{-1}P_F$ and $|\Sigma| = \lambda^k(\sigma^2)^{mk-k}$ (see exercises 2,3,5). The two factors in the above likelihood are in fact 'generalized' densities of the 'degenerate' normal vectors $P_F Y$ and $Q_I Y$ (i.e. these vectors are n dimensional but their distributions are concentrated on lower-dimensional subspaces of dimension k and $n - k$).

Consider e.g. the factor $\lambda^{-k/2}\exp(-\frac{1}{2}\|P_F Y - 1_n\xi\|^2)$ involving the parameters λ and ξ. We can maximize this with respect to λ and ξ in exactly the same way as when we previously considered the likelihood of a linear normal model $N_n(\mu, \sigma^2 I)$. In our case the data vector is $P_F Y$ and $\mu = 1_n\xi \in L_0 = \text{span}\{1_n\}$. We thus obtain

$$\widehat{1_n\xi} = P_0 P_F Y = P_0 Y \text{ and } \hat{\lambda} = \|P_F Y - P_0 Y\|^2/k = SSB/k.$$

Proceeding in the same way for the second factor (where there is no mean parameter), we obtain

$$\hat{\sigma}^2 = \|Q_I Y\|^2/(k(m-1)) = SSE/(k(m-1)).$$

Note that $Q_F Y \sim N_n(0, \lambda Q_F)$. By exercise 6, $\|P_F Y - P_0 Y\|^2 = \|Q_F Y\|^2 \sim \lambda\chi^2(k-1)$ which has mean $\lambda(k-1)$. Thus $\hat{\lambda}$ is biased. An unbiased (REML) estimate is given by

$$\tilde{\lambda} = \|P_F Y - P_0 Y\|^2/(k-1) = SSB/(k-1).$$

3 Two-way analysis of variance

Consider now a model with two factors T (treatment) and P (plot) with numbers of levels d_T and d_P. Moreover let $P \times T$ be the cross-factor (has $d_P d_T$ levels - one for each combination of levels of P and T). More specifically, I consists of indices ptr and the factor mappings are $P(ptr) = p$, $T(ptr) = t$, and $P \times T(ptr) = pt$. Assume $P \times T$ is balanced with $n_{P\times T} = m$ observations for each level. Then P and T are balanced too with numbers of observations $n_P = md_T$ and $n_T = md_P$ for each level.

A two-way ANOVA model with fixed T effects and random P and $P \times T$ effects is

$$Y_{ptr} = \xi + \beta_t + U_p + U_{pt} + \epsilon_{ptr} \quad p = 1, \ldots, d_P,\ t = 1, \ldots, d_T,\ r = 1, \ldots, m,$$

where the $U_p \sim N(0, \sigma_P^2)$, $U_{pt} \sim N(0, \sigma_{P\times T}^2)$ and $\epsilon_{ptr} \sim N(0, \sigma^2)$ are independent. With a convenient abuse of terminology we refer to P and $P \times T$ as random factors. The random effects U_p and U_{pt} can e.g. serve to model soil variation between plots in a field in case of an agricultural experiment. In vector form the model is

$$Y = \mu + Z_P U_P + Z_{P\times T} U_{P\times T} + \epsilon$$

where $\mu \in L_T$. The vectors Y, U_P and $U_{P\times T}$ are again obtained by stacking, e.g. $U_P = (U_1, \ldots, U_{d_P})^{\mathsf{T}}$.

3.1 Orthogonal decomposition

Similar to the one-way ANOVA we define:

$$V_0 = L_0 \quad V_T = L_T \ominus V_0 \quad V_P = L_P \ominus V_0.$$

Since $P \times T$ is balanced it follows that $P_T P_P = P_P P_T = P_0$ (exercise 7) which implies $Q_T Q_P = 0$ where $Q_T = P_T - P_0$ and $Q_P = P_P - P_0$. Hence V_T and V_P are orthogonal and we obtain an orthogonal decomposition of the sum of L_P and L_T:

$$L_P + L_T = \{v + w | v \in L_P, w \in L_T\} = V_0 \oplus V_P \oplus V_T.$$

We further define

$$V_{P\times T} = L_{P\times T} \ominus (L_P + L_T), \quad V_I = \mathbb{R}^n \ominus L_{P\times T}$$

and obtain the orthogonal decomposition:

$$\mathbb{R}^n = V_0 \oplus V_P \oplus V_T \oplus V_{P\times T} \oplus V_I.$$

The dimensions of the 'V' spaces are $f_0 = 1$, $f_P = d_P - 1$, $f_T = d_T - 1$, $f_{P\times T} = d_P d_T - d_P - d_T + 1 = (d_P - 1)(d_T - 1)$, $f_I = n - d_P d_T$. The orthogonal projections on the 'V' spaces are : $Q_0 = P_0$, $Q_P = P_P - Q_0$, $Q_T = P_T - Q_0$, $Q_{P\times T} = P_{P\times T} - Q_P - Q_T - Q_0$ and $Q_I = I - P_{P\times T}$.

In line with the orthogonal decomposition of $\mathbb{R}^n$ we also obtain two decompositions of Y into orthogonal components:

$$\begin{aligned} Y &= Q_0Y + Q_PY + Q_TY + Q_{P\times T}Y + Q_IY \\ &= \tilde{Q}_PY + \tilde{Q}_{P\times T}Y + \tilde{Q}_IY \end{aligned}$$

where $\tilde{Q}_P = Q_0 + Q_P = P_P$ $\tilde{Q}_{P\times T} = Q_{P\times T} + Q_T$ and $\tilde{Q}_I = Q_I$. The second decomposition corresponds to a coarser decomposition

$$\mathbb{R}^n = \tilde{V}_P \oplus \tilde{V}_{P\times T} \oplus \tilde{V}_I$$

into orthogonal subspaces $\tilde{V}_P = V_0 \oplus V_P = L_P$, $\tilde{V}_{P\times T} = V_T \oplus V_{P\times T}$ and $\tilde{V}_I = V_I$ associated with each random factor P, $P \times T$ and I.

The coarser decomposition of $\mathbb{R}^n$ corresponds to a decomposition of the covariance matrix using the $\tilde{Q}$ projection matrices:

$$\begin{aligned} \mathbb{C}\mathrm{ov}Y &= \sigma_P^2 n_P P_P + \sigma_{P\times T}^2 n_{P\times T} P_{P\times T} + \sigma^2 I \\ &= \lambda_P \tilde{Q}_P + \lambda_{P\times T}\tilde{Q}_{P\times T} + \lambda_I \tilde{Q}_I \end{aligned}$$

where

$$\lambda_P = \sigma^2 + n_{P\times T}\sigma_{P\times T}^2 + n_P\sigma_P^2 \tag{2}$$

$$\lambda_{P\times T} = \sigma^2 + n_{P\times T}\sigma_{P\times T}^2 \tag{3}$$

$$\lambda_I = \sigma^2.$$

The mean vector $\mu \in L_T = V_0 \oplus V_T$ is decomposed as:

$$\mu = \tilde{Q}_P\mu + \tilde{Q}_{P\times T}\mu + \tilde{Q}_I\mu = Q_0\mu + Q_T\mu + 0 = \mu_0 + \mu_T$$

where $\mu_0 \in L_0$ and $\mu_T \in V_T$. Note the one-to-one correspondences between the 'σ^2's and the 'λ's and between μ and (μ_0, μ_T).

3.2 Relation to sum-to-zero constraint

In the model for the mean vector μ, the parameters ξ and $\beta_1, \ldots, \beta_{d_T}$ are not identifiable. This is because

$$\mu_{ptr} = \xi + \beta_t = (\xi + k) + (\beta_t - k) = \tilde{\xi} + \tilde{\beta}_t.$$

Hence if we add k to ξ we obtain the same mean vector by just subtracting k from the β_t's. One way to enforce identifiability is to require that one β_t is zero (i.e. choosing a reference group). Another option is to introduce the sum-to-zero constraint: $\sum_{t=1}^{d_T} \beta_t = 0$. Note that vectors $v \in V_T$ are of the form $v_{ptr} = \beta_t$ since they lie in L_T and they further satisfy the sum-to-zero constraint $1_n^\mathsf{T} v = d_P m \sum_{t=1}^{d_T} \beta_t = 0$ due to the orthogonality of V_0 and V_T. In other words, enforcing the sum-to-zero constraint on the β_t parameters corresponds to the decomposition of μ into a constant vector $\xi 1_n$ falling in V_0 and a vector $v = (\beta_t)_{ptr} \in V_T$.

3.3 Estimation using factorization of likelihood

As for one-way ANOVA, Y is decomposed into independent normal vectors

$$\tilde{Q}_P Y \sim N(\mu_0, \lambda_P \tilde{Q}_P) \quad \tilde{Q}_{P\times T} Y \sim N(\mu_T, \lambda_{P\times T} \tilde{Q}_{P\times T}) \quad \tilde{Q}_I Y \sim N(0, \lambda_I \tilde{Q}_I).$$

Accordingly, the density for Y factorizes into a product of (generalized) densities for $\tilde{Q}_P Y$, $\tilde{Q}_{P\times T} Y$ and $\tilde{Q}_I Y$ similar to the case of the one-way ANOVA. For instance, the density for $\tilde{Q}_P Y$ is proportional to

$$\lambda_P^{-d_P/2} \exp(-\frac{1}{2\lambda_P} \|\tilde{Q}_P Y - \mu_0\|^2).$$

Thus in analogy with estimation in the usual linear model $Y \sim N_n(\mu, \sigma^2 I)$ we obtain

$$\hat{\mu}_0 = P_0 \tilde{Q}_P Y = P_0 Y = 1_n \bar{Y}_{\cdot\cdot}$$

and

$$\hat{\lambda}_P = \|\tilde{Q}_P Y - P_0 Y\|^2 / d_P = \|Q_P Y\|^2 / d_P.$$

Similarly,

$$\hat{\mu}_T = Q_T \tilde{Q}_{P\times T} Y = Q_T Y,$$

$$\hat{\mu} = \hat{\mu}_0 + \hat{\mu}_T = P_T Y,$$

$$\hat{\lambda}_{P\times T} = \|\tilde{Q}_{P\times T} Y - Q_T Y\|^2 / (d_P d_T - d_P) = \|Q_{P\times T} Y\|^2 / (d_P d_T - d_P),$$

and

$$\hat{\lambda}_I = \hat{\sigma}^2 = \|Q_I Y\|^2 / (n - d_P d_T).$$

Since V_P and $V_{P\times T}$ have dimensions $f_P = d_P - 1$ and $f_{P\times T} = (d_P - 1)(d_T - 1)$ we often use these (cf. exercise 6) denominators instead of d_P and $d_P d_T - d_P$ for λ_P and $\lambda_{P\times T}$. Then unbiased estimates

$$\tilde{\lambda}_P = \|Q_P Y\|^2/f_P \quad \tilde{\lambda}_{P\times T} = \|Q_{P\times T} Y\|^2/f_{P\times T}$$

are obtained for λ_P and $\lambda_{P\times T}$. Note that the MLE for μ is in fact identical to the MLE in the linear model with $\mu \in L_T$ and no random effects.

3.3.1 Estimation of original variance components

Estimates of the original variances σ_P^2 and $\sigma_{P\times T}^2$ can be obtained by simply solving (2) and (3) with respect to these parameters. That is,

$$\tilde{\sigma}_P^2 = (\tilde{\lambda}_P - \tilde{\lambda}_{P\times T})/n_P \quad \tilde{\sigma}_{P\times T}^2 = (\tilde{\lambda}_{P\times T} - \tilde{\sigma}^2)/n_{P\times T}.$$

Note that there is no guarantee that these estimates are non-negative. In fact there could be several reasons for obtaining negative variance estimates:

- if e.g. the true σ_P^2 is zero or close to zero it is not unlikely to encounter $\tilde{\lambda}_P < \tilde{\lambda}_{P\times T}$ due to sampling variation.
- observations within a group could be negatively correlated which could be reflected by a negative variance estimate (remember definition of intra-class correlation).
- data deviates in some way (e.g. outliers) from the assumed model.

In general, if a negative variance estimate is obtained, it is recommended (as always) to check carefully whether model assumptions seem valid. If data does not seem to contradict the model, one may consider refitting the model without the random factor for which a negative variance estimate was obtained. If negative correlation within groups of the data is plausible one may look into an alternative model to the assumed linear mixed model.

4 Strata

A crucial common point for the one- and two-way ANOVA models in the previous sections is that the factorization of the likelihood is based on an orthogonal decomposition induced by the random factors only. This is due

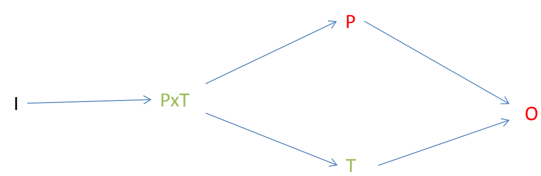

Figure 1: Structure diagram for two-way ANOVA with random effects.

to the decomposition of the covariance matrix into terms given by scaled projection matrices for the random factors.

The fixed effects are decomposed into components according to 'V' spaces for the fixed factors. These components are then assigned to *strata* corresponding to '$\tilde{V}$' spaces for the random factors. We say that a factor F is finer than G (or G coarser than F) if levels of G can be obtained by merging levels of F. We then write $G \preceq F$. The rule for allocating fixed effects to strata is then that F belongs to B strata (where B is a random factor) if B is the coarsest random factor which is finer than F. This is of course assuming that this rule makes sense, i.e. we work with an experimental design where this rule gives a unique allocation of each fixed factor to one random factor. It is for instance required that I is a random factor. Also, for fixed F and random B_1, B_2, we can not have both $F \preceq B_1$ and $F \preceq B_2$ unless either $B_1 \preceq B_2$ or $B_2 \preceq B_1$.

To get an overview of allocation to strata it is useful to draw a structure diagram as shown for the two-way ANOVA in Figure 1. In the structure diagram there is an arrow from F to G if G is coarser than F and there is no intermediate factor which is coarser than F and finer than G. In Figure 1 there are three strata (black, green and red) corresponding to the ‘random’ factors I, $P \times T$ and P. Note that here we can not have both P and T random unless O is random too (if O was fixed in this case there would not be a unique allocation of O to a random factor stratum). We also want to respect the hierarchical principle for the mean structure so $P \times T$ can not be fixed if P is random.

A more general discussion of allocation to strata is given in Section 7.

5 Hypothesis tests and confidence intervals

We here exemplify how hypothesis tests are conducted in balanced ANOVAs by considering one- and two-way ANOVAs.

5.1 One-way ANOVA

For the one-way ANOVA with random effects consider the F-statistic

$$F = \frac{SSB/(k-1)}{SSE/(k(m-1))} = \frac{\tilde{\lambda}}{\hat{\sigma}^2} \sim \frac{\sigma^2 + m\tau^2}{\sigma^2} F(k-1, k(m-1)) = (1+m\gamma)F(k-1,k(m-1))$$

where $\gamma = \tau^2/\sigma^2$ is the signal to noise ratio. Thus with q_L and q_U e.g. 2.5% and 97.5% quantiles for $F(k-1, k(m-1))$, we have

$$P(q_L \le F/(1+m\gamma) \le q_U) = 95\% \Leftrightarrow P((F/q_U - 1)/m \le \gamma \le (F/q_L - 1)/m) = 95\%$$

so that $[(F/q_U - 1)/m, (F/q_L - 1)/m)]$ is a 95% confidence interval for γ, see also Remark 5.10 in [1].

One hypothesis is of particular interest:

$$H_0 : \tau^2 = 0,$$

i.e. there is no variation between groups. A simple F-test for this is based on the observation

$$\tau^2 = 0 \Leftrightarrow \gamma = 0 \Leftrightarrow \sigma^2 = \lambda.$$

Thus large values of the F statistic above is critical for H_0 and under H_0, F has a $F(k-1, k(m-1))$ distribution.

Note that the F statistic coincides with the F-test for no group effects in a one-way ANOVA with fixed group effects (which is equivalent to the likelihood ratio test for no group effects in the fixed effects one-way ANOVA).

5.2 Tests for fixed effects in two-way ANOVA

For the two-way ANOVA considered in Section 3 the interesting hypothesis regarding fixed effects is

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_{d_T}$$

i.e. no fixed group effects. This is equivalent to that $\mu_T = 0$ (referring to the notation of Section 3.1). Thus for constructing the likelihood ratio test we only need to consider the factor corresponding to the term $\tilde{Q}_{P\times T}Y \sim N(\mu_T, \lambda_{P\times T}\tilde{Q}_{P\times T})$ which has density

$$\lambda_{P\times T}^{(d_P d_T - d_P)/2} \exp\left(-\frac{1}{2\lambda_{P\times T}}\|\tilde{Q}_{P\times T}Y - \mu_T\|^2\right).$$

Thus the situation is equivalent to testing $\mu = 0$ for a linear normal model $N_{d_P d_T - d_P}(\mu, \lambda_{P\times T}I)$. We can thus proceed as for a usual linear normal model (without random effects) and obtain the F-test

$$F = \frac{\|Q_T Y\|^2}{\tilde{\lambda}_{P\times T}} \sim F(d_T - 1, (d_P - 1)(d_T - 1)).$$

5.3 Test for variance components in two-way ANOVA

Recall $\lambda_I = \sigma^2$, $\lambda_{P\times T} = \sigma^2 + n_{P\times T}\sigma^2_{P\times T}$ and $\lambda_P = \sigma^2 + n_{P\times T}\sigma^2_{P\times T} + n_P\sigma^2_P$. Hence e.g. $\sigma^2_{P\times T} = 0 \Leftrightarrow \lambda_I = \lambda_{P\times T}$. Thus a natural statistic (but not likelihood-ratio statistic) for testing $\sigma^2_{P\times T} = 0$ is

$$F = \frac{\tilde{\lambda}_{P\times T}}{\hat{\sigma}^2}$$

which has an $F((d_P-1)(d_T-1), n-d_{P\times T})$ distribution if $\sigma^2_{P\times T} = 0$. Big values of F are critical for $\sigma^2_{P\times T} = 0$. Note $\tilde{\lambda}_{P\times T} = \|Q_{P\times T}Y\|^2/((d_P - 1)(d_T - 1))$ so F is identical to the F-statistic for testing the hypothesis of no (fixed) effect of the factor $P \times T$ in a linear normal model without random effects.

5.4 Confidence intervals for parameters in two-way ANOVA

Regarding the mean vector $\mu = 1_n\xi + Z_T\beta$ where $\beta = (\beta_1, \ldots, \beta_{d_T})$, we can consider confidence intervals for components of $\xi + \beta_i$ of μ or for contrasts $\delta_{ij} = \beta_i - \beta_j$ corresponding to differences between group means. Here it is easy to derive confidence intervals for the contrasts. Employing the orthogonal decomposition $\mu = \mu_0 + \mu_T$, the contrasts only involve μ_T in the $P \times T$ stratum. E.g. if c is a vector with $c_{1i1} = 1$, $c_{1j1} = -1$ and zeroes elsewhere then

$$\delta_{ij} = \beta_i - \beta_j = c^\mathsf{T}\mu = c^\mathsf{T}\mu_T.$$

Thus δ_{ij} can be estimated as

$$\hat{\delta}_{ij} = c^\mathsf{T}\hat{\mu}_T = c^\mathsf{T}Q_T\tilde{Q}_{P\times T}Y \sim N(\beta_i - \beta_j, \lambda_{P\times T}c^\mathsf{T}Q_Tc)$$

and we can thus follow the usual route to constructing confidence intervals based on the t-distribution combining the above normal distribution for $\hat{\delta}_{ij}$ and the scaled χ^2 distribution of $\tilde{\lambda}_{P\times T}$ with degrees of freedom $(d_P - 1)(d_T - 1)$. Note

$$c^\mathsf{T}Q_T\tilde{Q}_{P\times T}Y = c^\mathsf{T}Q_TY = \bar{y}_{\cdot i\cdot} - \bar{y}_{\cdot j\cdot}.$$

and

$$c^\mathsf{T}Q_Tc = c^\mathsf{T}P_Tc - c^\mathsf{T}P_0c = c^\mathsf{T}P_Tc = (1, -1)(1/n_T, -1/n_T)^\mathsf{T} = \frac{2}{n_T}.$$

Hence the t-statistic becomes

$$\frac{\bar{y}_{\cdot i\cdot} - \bar{y}_{\cdot j\cdot}}{\sqrt{2\tilde{\lambda}_{P\times T}/n_T}}$$

which has a $t((d_P - 1)(d_T - 1))$ distribution under the null hypothesis $\beta_i = \beta_j$. You can also get the above results by simply observing that

$$\bar{y}_{\cdot i\cdot} - \bar{y}_{\cdot j\cdot} = \beta_i - \beta_j + \frac{m}{n_T}U_{\cdot i} - \frac{m}{n_T}U_{\cdot j} + \frac{1}{n_T}\epsilon_{\cdot i\cdot} - \frac{1}{n_T}\epsilon_{\cdot j\cdot} \sim N\left(\beta_i - \beta_j, \frac{2}{n_T}(m\sigma^2_{P\times T} + \sigma^2)\right).$$

Note here $m = n_{P\times T}$ so $m\sigma_{P\times T} + \sigma^2 = \lambda_{P\times T}$.

Confidence intervals for the 'λ' variance parameters (including $\sigma^2 = \lambda_I$) are straightforward due to the exact scaled χ^2 distributions of their estimates. Confidence intervals for the original variance parameters (other than σ^2) are less straightforward since estimates for these are distributed as differences of scaled χ^2 distributions).

6 Orthogonal decomposition for a K-way ANOVA

In this section we first construct an orthogonal decomposition for a balanced three-way ANOVA and then generalize the approach to a K-way balanced ANOVA for arbitrary $K \geq 1$.

We know by now that for a two-sided ANOVA corresponding to factors F_1 and F_2 with $F_1 \times F_2$ balanced, there exists a decomposition of $\mathbb{R}^n$ into orthogonal subspaces V_I, $V_{F_1\times F_2}$, V_{F_1}, V_{F_2} and V_0. Suppose we introduce a third factor F_3 so that $F_1 \times F_2 \times F_3$ is balanced too. Then also $\{V_I, V_{F_1\times F_3}, V_{F_1}, V_{F_3}, V_0\}$, $\{V_I, V_{F_1\times F_2}, V_{F_1}, V_{F_2}, V_0\}$ and $\{V_I, V_{F_2\times F_3}, V_{F_2}, V_{F_3}, V_0\}$ are sets of orthogonal subspaces. By analogy with two-way ANOVA with factors $F = F_1 \times F_2$ and $G = F_3$ it also follows that $V_{F_1\times F_2}$ and V_{F_3} are orthogonal and similarly for the pairs $V_{F_1\times F_3}$, V_{F_2} and $V_{F_2\times F_3}$, V_{F_1}. It is also easy to see that

$$P_{F_1\times F_2} P_{F_2\times F_3} = P_{F_2}. \tag{4}$$

This implies $V_{F_1\times F_2}$ and $V_{F_2\times F_3}$ are orthogonal. To see this note that the corresponding projections are $Q_{F_1\times F_2} = P_{F_1\times F_2} - P_{F_2} - Q_{F_1}$ and $Q_{F_2\times F_3} = P_{F_2\times F_3} - P_{F_2} - Q_{F_3}$. We now check that $Q_{F_1\times F_2} Q_{F_2\times F_3} = 0$ from which the required orthogonality follows:

$$\begin{aligned} &Q_{F_1\times F_2} Q_{F_2\times F_3} = (P_{F_1\times F_2} - P_{F_2} - Q_{F_1})(P_{F_2\times F_3} - P_{F_2} - Q_{F_3}) = \\ &P_{F_2} - P_{F_2} - P_{F_1\times F_2} Q_{F_3} - P_{F_2} + P_{F_2} - 0 - Q_{F_1} P_{F_2\times F_3} + 0 + 0 = 0. \end{aligned}$$

Here, $P_{F_1\times F_2} Q_{F_3} = 0$ in analogy with a two-way ANOVA with factors $F = F_1 \times F_2$ and $G = F_3$, and $Q_{F_1} P_{F_2\times F_3} = 0$ by the same argument. Thus $V_{F_1\times F_2}$, $V_{F_1\times F_3}$ and $V_{F_2\times F_3}$ are orthogonal too.

Defining

$$\begin{aligned} V_{F_1\times F_2\times F_3} &= L_{F_1\times F_2\times F_3} \ominus (V_0 \oplus V_{F_1} \oplus V_{F_2} \oplus V_{F_3} \oplus V_{F_1\times F_2} \oplus V_{F_1\times F_3} \oplus V_{F_2\times F_3}) \\ V_I &= \mathbb{R}^n \ominus V_{F_1\times F_2\times F_3} \end{aligned}$$

we arrive at the orthogonal decomposition

$$\mathbb{R}^n = V_I \oplus_{A\subseteq\{1,2,3\}} V_{\times_{l\in A} F_l}$$

letting $V_{\times_{l\in\emptyset} F_l} = V_0$.

The approach for the three-way ANOVA can be generalized to obtain the following main result:

Theorem 1. *Suppose $F_1 \times \cdots \times F_K$ is balanced, $K \geq 1$. Define, recursively, for $A \subseteq \{1, \ldots, K\}$,*

$$V_{\times_{l \in A} F_l} = L_{\times_{l \in A} F_l} \ominus \sum_{B \subset A} V_{\times_{l \in B} F_l}$$

with $V_{\times_{l \in \emptyset} F_l} = V_0 = L_0$ and let

$$V_I = \mathbb{R}^n \ominus L_{\times_{l=1}^{K} F_l}.$$

Then we have the orthogonal decomposition

$$\mathbb{R}^n = V_I \oplus_{A \subseteq \{1,\ldots,K\}} V_{\times_{l \in A} F_l}.$$

Proof. We need to show that (*) $V_{\times_{l \in A} F_l}$ and $V_{\times_{l \in B} F_l}$ are orthogonal whenever $A, B \subseteq \{1, \ldots, K\}$, $A \neq B$ (this is already shown for $K = 1, 2, 3$). If $A \cap B = \emptyset$ then (*) follows by analogy to a two-way ANOVA with factors $F = \times_{l \in A} F_l$ and $G = \times_{l \in B} F_l$. If $B \subset A$ we define $F = \times_{l \in B} F_l$ and $G = \times_{l \in A \setminus B} F_l$. Then $V_{\times_{l \in A} F_l} = V_{F \times G}$ and the result again follows by analogy to a two-way ANOVA (V_F and $V_{F \times G}$ orthogonal). The case $A \subset B$ is handled similarly. Finally if $C = A \cap B$ is non-empty and neither $A \subset B$ nor $B \subset A$, let $F = \times_{l \in A \setminus C} F_l$, $G = \times_{l \in B \setminus C} F_l$, and $H = \times_{l \in C} F_l$. Then (*) holds by analogy to a three-way ANOVA. □

The dimensions of the spaces $V_{\times_{l \in A} F_l}$ are easy to obtain according to the following theorem.

Theorem 2. *Let the situation be as in the previous theorem. Then the dimension of $V_{\times_{l \in A} F_l}$ is $\prod_{l \in A} f_l$ where $f_l = d_l - 1$ and d_l is the dimension of L_{F_l}.*

Proof. Let wlog $A = 1, \ldots, k$, $1 \leq k \leq K$. The proof is by induction in k. By definition the dimension of V_{F_i} is $d_i - 1 = f_i$ so the result holds for $k = 1$. We have

$$V_{\times_{i=1}^{k} F_i} = L_{\times_{i=1}^{k} F_i} \ominus \left(L_{\times_{i=1}^{k-1} F_i} \oplus \oplus_{B \subset \{1,\ldots,k-1\}} V_{F_k \times \times_{l \in B} F_l} \right).$$

Thus the dimension of $V_{\times_{i=1}^{k} F_i}$ is

$$\begin{aligned}
&\prod_{i=1}^{k}(f_i+1)-\prod_{i=1}^{k-1}(f_i+1)-\sum_{B\subset\{1,\dots,k-1\}} f_k\prod_{l\in B} f_l \\
=&f_k\prod_{i=1}^{k-1}(f_i+1)-\left[f_k\sum_{B\subseteq\{1,\dots,k-1\}}\prod_{l\in B} f_l - f_k\prod_{l=1}^{k-1} f_l\right] \\
=&\prod_{l=1}^{k} f_l
\end{aligned}$$

where the second equality is by induction and the second follows from $\prod_{i=1}^{k-1}(f_i+1)=\sum_{B\subseteq\{1,\dots,k-1\}}\prod_{l\in B} f_l\prod_{l\in\{1,\dots,k-1\}\setminus B} 1$. □

7 Decomposition into strata

In this section we discuss a general decomposition into strata corresponding to factors with random effects. Let $\mathcal{D}$ be a set of factors and let $\mathcal{B}\subseteq\mathcal{D}$ denote the set of factors with random effects. We assume

1. the factors in $\mathcal{D}$ are ordered in the sense that we can not have both $F\preceq F'$ and $F'\preceq F$ for different $F,F'\in\mathcal{D}$ (otherwise F and F' would induce the same grouping of the observations).

2. there is an orthogonal decomposition $\mathbb{R}^n=\oplus_{F\in\mathcal{D}}V_F$ of $\mathbb{R}^n$ into a sum of subspaces V_F, $F\in\mathcal{D}$, with associated orthogonal projection matrices Q_F.

3. the covariance matrix of the data vector is decomposed as

$$\Sigma=\sum_{B\in\mathcal{B}}\sigma_B^2 n_B P_B$$

   where for $B\in\mathcal{B}$,

$$P_B=\sum_{F\in\mathcal{D}:F\preceq B} Q_F. \tag{5}$$

We want $\tilde{Q}_{B'}$'s, $B' \in \mathcal{B}$, so that

$$P_B = \sum_{B' \in \mathcal{B}: B' \preceq B} \tilde{Q}_{B'} \tag{6}$$

where for each $B \in \mathcal{B}$, $\tilde{Q}_B$ is an orthogonal projection on a subspace $\tilde{V}_B$ where $\mathbb{R}^n = \oplus_{B \in \mathcal{B}} \tilde{V}_B$ and $\tilde{V}_B$ is a sum of V_F subspaces. We want this because then we obtain the decomposition

$$\Sigma = \sum_{B \in \mathcal{B}} \sigma_B^2 n_B P_B = \sum_{B \in \mathcal{B}} \lambda_B \tilde{Q}_B \quad \text{where} \quad \lambda_B = \sum_{B' \in \mathcal{B}: B \preceq B'} n_{B'} \sigma_{B'}^2$$

and hence a decomposition of Y into independent terms $\tilde{Q}_B Y$, $B \in \mathcal{B}$ (here we interchanged order of summation and secondly interchanged B and B').

The following theorem tells us when we may get what we want.

Theorem 3. *A necessary and sufficient condition for* (6) *is that: for all $F \in \mathcal{D}$ there exists a $B \in \mathcal{B}$ such that $F \preceq B$ and $B \preceq B'$ for all other B' with $F \preceq B'$. If this condition holds we can define*

$$\tilde{V}_{B'} = \sum_{F \in \mathcal{D}: B(F) = B'} V_F \quad \textit{and} \quad \tilde{Q}_{B'} = \sum_{F \in \mathcal{D}: B(F) = B'} Q_F.$$

Remark 1. *Note that this condition implies that I has to be in $\mathcal{B}$. This is not a serious restriction since we will always include measurement error in our models. It also combined with assumption 1 above implies that the B mentioned in the condition is unique. In the expressions for $\tilde{V}_{B'}$ and $\tilde{Q}_{B'}$ above and in the proof below we use the notation $B(F)$ for the unique B corresponding to F, $F \in \mathcal{D}$.*

Proof. We first show the sufficiency. Each of the factors $F \preceq B$ on the right hand side of (5) has $B(F) = B'$ for precisely one $B' \in \mathcal{B}$ for which it must hold that $B' \preceq B$ (since $F \preceq B$ - here B and B' interchanged compared to the condition). Hence

$$P_B = \sum_{B' \in \mathcal{B}: B' \preceq B} \sum_{F \in \mathcal{D}: B(F) = B'} Q_F$$

Thus we can define for $B' \in \mathcal{B}$, a subspace and associated orthogonal projection

$$\tilde{V}_{B'} = \sum_{F \in \mathcal{D}: B(F) = B'} V_F \quad \text{and} \quad \tilde{Q}_{B'} = \sum_{F \in \mathcal{D}: B(F) = B'} Q_F.$$

Each $F \in \mathcal{D}$ belongs to precisely one $\tilde{V}_{B'}$ so $\mathbb{R}^n = \oplus_{B' \in \mathcal{B}} \tilde{V}_{B'}$.

To show that the condition is necessary consider any $F \in \mathcal{D}$. Then $V_F \subseteq \tilde{V}_B$ for some unique $B \in \mathcal{B}$. This means by (6) that Q_F is a part of the sum defining P_B. It follows from (5) that $F \preceq B$. Assume $F \preceq B_i$, $i = 1, \ldots, m$ where $B_1 = B$. If $m = 1$ there is nothing more to prove. Suppose $m \geq 2$ and assume to get a contradiction that $B \not\preceq B_2$. Since $F \preceq B_2$, by (5), Q_F is a part of the sum defining P_{B_2}. It follows that $\tilde{Q}_B$ must be a part of the sum defining P_{B_2} according to (6). But this is a contradiction since $B \not\preceq B_2$. □

8 Concluding remarks

This note essentially gives a simplified exposition of results presented in [3]. A general exposition can also be found in [2]. We here considered the use of orthogonal decompositions only in the context of balanced ANOVAs. However, the same principles apply for other types of linear mixed models allowing for relevant orthogonal decompositions of the sample space. One example is a random coefficient model for the well-known orthodontic data set consisting of data for 27 children each with 4 measurements. Here one can e.g. observe that the MLEs of the mean parameters are the same in the models with and without random child effects.

References

[1] Henrik Madsen and Poul Thyregod. 2011. Introduction to general and generalized linear models. Chapman & Hall/CRC Texts in Statistical Science.

[2] Jesper Møller. 1995. *Centrale statistiske modeller og likelihood baserede metoder*, Department of Mathematical Sciences, Aarhus University.

[3] Tue Tjur. 1984. Analysis of variance models in orthogonal designs, *International Statistical Review*, 52, 33-65.

A Orthogonal projections

Suppose L is a subspace of $\mathbb{R}^n$, $n \geq 1$. Let $L^\perp = \{v \in \mathbb{R}^n | v^\mathsf{T} w = 0 \text{ for all } w \in L\}$ denote its orthogonal complement.

Orthogonal decomposition: each $x \in \mathbb{R}^n$ has a unique decomposition

$$x = u + v$$

where $u \in L$ and $v \in L^\perp$.

Orthogonal projection: u and v above are the orthogonal projections $p_L(x)$ and $p_{L^\perp}(x)$ of x on respectively L and $L^\perp$.

Due to orthogonality of u and v, we obtain Pythagoras' theorem:

$$\|x\|^2 = \|u\|^2 + \|v\|^2.$$

A few facts regarding orthogonal projections:

(a) the orthogonal projection $p_L : \mathbb{R}^n \to L$ on L is a linear mapping. It is thus given by a unique matrix-transformation $p_L(x) = Px$ where P is an $n \times n$ matrix.

(b) the projection matrix P is symmetric ($P^\mathsf{T} = P$) and idempotent ($P^2 = P$).

(c) conversely, if a matrix Q is symmetric, idempotent and $L = \mathrm{col}Q$ then Q is the matrix of the orthogonal projection on L.

(d) if $L = \mathrm{col}X$ and X has full rank then $P = X(X^\mathsf{T}X)^{-1}X^\mathsf{T}$.

(e) if P is the matrix of the orthogonal projection on L then $I - P$ is the matrix of the orthogonal projection on $L^\perp$.

(f) if L_1 and L_2 are subspaces with $L_1 \subset L_2$ and associated orthogonal projections P_1 and P_2, then the orthogonal projection on $L_2 \ominus L_1$ is $P_2 - P_1$.

Example: the orthogonal projection on a subspace spanned by a single vector v is $\frac{v^\mathsf{T}x}{\|v\|^2}x$.

Example: the orthogonal projection on a subspace spanned by orthogonal vectors $v_1, \ldots, v_p$ is $\sum_{i=1}^p \frac{v_i^\mathsf{T}x}{\|v_i\|^2}x$.

B Exercises

1. Let $L_1 \subset L_2$ with orthogonal projections P_1 and P_2. Show that $P_2 - P_1$ is the orthogonal projection on $L_2 \ominus L_1$.

2. Show that an orthogonal projection only has eigen values 1 or 0.

3. For a symmetric matrix A show that the determinant $|A|$ is the product of A's eigen values.

4. Show $Q_I P_F = 0$

5. Let $S = aP + bQ$ where P and Q are orthogonal projections with $P + Q = I$ and $a, b \neq 0$. Show that the eigen values of S are the non-zero eigen values a and b of aP and bQ. Show that $S^{-1} = a^{-1}P + b^{-1}Q$.

6. Show that $\|Y\|^2 \sim \sigma^2\chi^2(d)$ if $Y \sim N(0, \sigma^2 P)$ and P is an orthogonal projection on a subspace of dimension d (hint: use spectral decomposition and the result above regarding the eigen values of P).

7. Check that $P_P P_T = P_0$ when $P \times T$ is balanced.

8. Check that $L_T + L_P = V_0 \oplus V_P \oplus V_T$.