then work with standard tools of multivariate statistics such as multivariate normal distributions, linear models, and methods based on correlation or distance matrices. Underlying the Aitchison approach is a notion of ‘compositional data’. Suppose that D positive-valued measurements are made on each of N objects. The D measurements on an object, all in the same units, relate to D separate ‘parts’ of the object. Write y_{ik} for the measurement relating to the kth part of the ith object. The observations can then be represented as $y_{ik} = t_i p_{ik}$ $(i = 1, \ldots, N;\ k = 1, \ldots, D)'$, with $t_i = \sum_k y_{ik}$, and the unit-sum vector $\mathbf{p}_i = (p_{i1}, \ldots, p_{iD})'$ being the observed composition of object i. Aitchison's methodology operates only on the *log-ratios* $\{\log(p_{ik}) - \log(p_{il})\}$, or more generally on some other defined set of linear contrasts among $\{\log(p_{i1}), \ldots, \log(p_{iD})\}$.

In practice, though, the observed $\{p_{ik}\}$ are usually error-affected *measurements* of the composition(s) of interest. Errors can come from any of the usual sources, for example sampling and other study-design effects, inaccurate instruments, numeric rounding, imperfect mixing, etc. In any given application a more flexible approach than data-transformation, then, is to tailor a statistical model to the application — a model that accounts properly for important sources of error, as well as targeting compositions directly through parameters.

1.2 Extensive variables and the need to model arithmetic means

The directly measured quantities y_{ik} are always *extensive*, in that they have the property of physical additivity. In time-use studies, for example, the aggregate time spent on two or more related activities is the sum of the times spent on each, and the time spent on each activity in a whole day is the sum of the times spent hour by hour. For extensive variables it is natural, even essential, to use statistical models that target arithmetic means on the original scale; see for example Cox and Snell (1981, ch. 2) or Cox and Donnelly (2011, ch. 4).

A broadly useful class of models takes each y_{ik} to be a realization of random variable Y_{ik} with mean $E(Y_{ik}) = \tau_i \pi_{ik}$, or in vector notation $E(\mathbf{Y}_i) = \tau_i \boldsymbol{\pi}_i$. The positive-valued vector a whole day is the sum of the times spent hour by hour. For extensive variables it is natural, even essential, to use statistical models that target arithmetic means on the original scale; see for example Cox and Snell (1981, ch. 2) or Cox and Donnelly (2011, ch. 4).

A broadly useful class of models takes each y_{ik} to be a realization of random variable Y_{ik} with mean $E(Y_{ik}) = \tau_i \pi_{ik}$, or in vector notation $E(\mathbf{Y}_i) = \tau_i \boldsymbol{\pi}_i$. The positive-valued vector parameter $\boldsymbol{\pi}_i = (\pi_{i1}, \ldots, \pi_{iD})'$ has unit sum for every i, and represents the composition that is measured by $\mathbf{y}_i = (y_{i1}, \ldots, y_{iD})'$; and the positive scalar τ_i allows each measurement vector $\mathbf{y}_i$ to have its own expected total (or size). This is a very general formulation based only on the first moment, and as such it allows a wide range of potential error distributions.

Included as a special case is the most standard model for categorical *counts*, namely $\mathbf{Y}_i \sim \text{multinomial}(t_i, \boldsymbol{\pi}_i)$. This is important not only because category counts feature in many applications, but also as a prime context where precision of measurement depends on the size parameters τ_i (which here are the known multinomial totals t_i). Recent work of Fiksel et al. (2022) develops methods in a similar spirit to the approach proposed here, to take account of multinomial-type sampling errors in a specific application context.

Even with *continuous* measurements there is often potential for more complex error structures than are supported by the use of log-ratios. For example, small samples (of rock, for instance) might be affected by physical detection limits that are irrelevant with larger samples; in time-use studies the tendency of survey respondents to report times to the nearest half-hour (say) will affect small and large time-periods quite differently; etc.

1.3 Focus here on multiplicative errors

While the general first-moment specification $E(\mathbf{Y}_i) = \tau_i \boldsymbol{\pi}_i$ provides substantial flexibility to account for a wide variety of different error structures, in the remainder of this paper we focus on the particular case of *multiplicative* errors (which might alternatively be called *relative* or *proportionate* errors). Specifically it is assumed that — for each part k of each object i — $Y_{ik} = \tau_i \pi_{ik} U_{ik}$, or in vector notation

$$\mathbf{Y}_i = \tau_i \mathbf{\Pi}_i \mathbf{U}_i, \tag{1}$$

where $\mathbf{\Pi}_i$ denotes the matrix $\mathrm{diag}(\boldsymbol{\pi}_i)$ and the relative errors U_{ik} all have unit mean. We denote the error variance-covariance matrix by $\mathrm{cov}(\mathbf{U}_i) = \mathbf{\Phi}$, the same for all i; this homoskedasticity can easily be relaxed where appropriate, and we comment briefly on it in the Discussion section below. It is natural to think in terms of the stronger assumption that the relative error vectors $\{\mathbf{U}_i\}$ are i.i.d., but that is not a necessary assumption for what follows.

If the random multipliers $\{U_{ik}\}$ are restricted to be positive, the log-ratios $\log(Y_{ik}/Y_{il})$ are available and are free of the size parameter τ_i. It is easily shown that if the relative-error vectors $\{\mathbf{U}_i\}$ are drawn from a multivariate lognormal distribution, suitably scaled to have $E(U_{ik}) = 1$ for all (i,k), then the resulting parametric model is the family of logistic normal distributions (Aitchison, 1986, ch. 6).

The general multiplicative model (1) is more widely applicable than log-ratios, though, because the relative errors U_{ik} are not restricted to be positive: zeros are allowed, or even negative values (as might be appropriate, for example, in situations where the measurement mechanism involves a differencing operation). In addition, importantly, the multiplicative model (1) always has parameters that relate directly to arithmetic means and totals on the original scale of measurement, regardless of any other distributional details of the unit-mean error vectors $\{\mathbf{U}_i\}$.

2 Variance-covariance function

The multiplicative model (1) describes how $\mathbf{Y}_i$ for each object i is an unbiased, error-affected measurement of the corresponding mean vector $\tau_i \boldsymbol{\pi}_i$. The part of the measurement error that

2 Variance-covariance function

The multiplicative model (1) describes how $\mathbf{Y}_i$ for each object i is an unbiased, error-affected measurement of the corresponding mean vector $\tau_i\boldsymbol{\pi}_i$. The part of the measurement error that relates purely to composition is $\mathbf{Y}_i - T_i\boldsymbol{\pi}_i$, where $T_i = \sum_k Y_{ik}$. We can write $\mathbf{Y}_i - T_i\boldsymbol{\pi}_i = (\boldsymbol{I} - \boldsymbol{\Pi}_i\boldsymbol{J})\mathbf{Y}_i$, where the $D \times D$ matrices $\boldsymbol{I}$ and $\boldsymbol{J}$ are respectively the identity matrix and the matrix of ones. Since $\mathrm{cov}(\mathbf{Y}_i/\tau_i) = \boldsymbol{\Pi}_i\boldsymbol{\Phi}\boldsymbol{\Pi}_i$, the variance-covariance matrix of $(\mathbf{Y}_i - T_i\boldsymbol{\pi}_i)/\tau_i$ also is free of τ_i:

$$\begin{aligned}\mathrm{cov}\left(\frac{\mathbf{Y}_i - T_i\boldsymbol{\pi}_i}{\tau_i}\right) &= (\boldsymbol{I} - \boldsymbol{\Pi}_i\boldsymbol{J})\boldsymbol{\Pi}_i\boldsymbol{\Phi}\boldsymbol{\Pi}_i(\boldsymbol{I} - \boldsymbol{\Pi}_i\boldsymbol{J})' \\ &= (\boldsymbol{\Pi}_i - \boldsymbol{\pi}_i\boldsymbol{\pi}_i')\boldsymbol{\Phi}(\boldsymbol{\Pi}_i - \boldsymbol{\pi}_i\boldsymbol{\pi}_i') \\ &= \boldsymbol{V}(\boldsymbol{\pi}_i;\boldsymbol{\Phi}), \text{ say.}\end{aligned} \tag{2}$$

The matrix variance-covariance function $\boldsymbol{V}$ is the natural extension, beyond the case $D = 2$, of a scalar variance function that was suggested previously in Wedderburn (1974) for a generalized linear model with continuous proportions as response variable. The case $D = 2$ has essentially univariate measurements, since $\pi_{i1} = 1 - \pi_{i2}$ for all i; and the suggestion made in Wedderburn (1974) was to use the variance function $V(\boldsymbol{\pi}_i;\phi) = \phi(\pi_{i1}\pi_{i2})^2$ for quasi-likelihood analysis of a logit-linear model. Wedderburn's suggested variance function is proportional to the square of the Bernoulli variance function $\pi_{i1}\pi_{i2}$, with scalar constant of proportionality ϕ. The more general form (2) uses the multinomial variance-covariance function $\boldsymbol{\Pi}_i - \boldsymbol{\pi}_i\boldsymbol{\pi}_i'$ to extend from Bernoulli, and the error dispersion matrix $\boldsymbol{\Phi}$ in place of scalar dispersion ϕ. Because of this connection, we will call $\boldsymbol{V}(\boldsymbol{\pi}_i;\boldsymbol{\Phi})$ the 'generalized Wedderburn' variance-covariance function.

The multinomial variance-covariance matrix $\boldsymbol{\Pi}_i - \boldsymbol{\pi}_i\boldsymbol{\pi}_i'$ is singular, and its Moore-Penrose pseudo-inverse was derived in Tanabe and Sagae (1992). In the $D \times D$ case, and with $\boldsymbol{C} =$