




Covariance Clustering: Modelling Covariance in Designed Experiments When the Number of Variables is Greater than Experimental Units

Clayton R. FORKNALL , Arūnas P. VERBYLA, Yoni NAZARATHY, Adel YOUSIF, Sarah OSAMA, Shirley H. JONES, Edward KERR, Benjamin L. SCHULZ, Glen P. FOX, and Alison M. KELLY

The size and complexity of datasets resulting from comparative research experiments in the agricultural domain is constantly increasing. Often the number of variables measured in an experiment exceeds the number of experimental units composing the experiment. When there is a necessity to model the covariance relationships that exist between variables in these experiments, estimation difficulties can arise due to the resulting covariance structure being of reduced rank. A statistical method, based in a linear mixed model framework, is presented for the analysis of designed experiments where datasets are characterised by a greater number of variables than experimental units, and for which the modelling of complex covariance structures between variables is desired. Aided by a clustering algorithm, the method enables the estimation of covariance through the introduction of covariance clusters as random effects into the modelling framework, providing an extension of the traditional variance components model for building covariance structures. The method was applied to a multi-phase mass spectrometry-based proteomics experiment, with the aim of exploring changes in the proteome of barley grain over time during the malting process. The modelling approach provides a new linear mixed model-based method for the estimation of covariance structures between variables measured from designed experiments, when there are a small number of experimental units, or observations, informing covariance parameter estimates.

C. R. Forknall (✉) · Y. Nazarathy, School of Mathematics and Physics, The University of Queensland, St Lucia, QLD 4067, Australia (E-mail: clayton.forknall@daf.qld.gov.au).

C. R. Forknall, Queensland Department of Agriculture and Fisheries, PO Box 4486, Toowoomba, QLD 4350, Australia.

A. P. Verbyla · A. M. Kelly, Queensland Alliance for Agriculture and Food Innovation, The University of Queensland, St Lucia, QLD 4067, Australia.

A. Yousif, Tasmanian Institute of Agriculture, University of Tasmania, Sandy Bay, TAS 7001, Australia.

S. Osama, Elizabeth Macarthur Agricultural, Department of Regional New South Wales, Menangle, NSW 2568, Australia.

S. H. Jones, Institute for Life Sciences and the Environment, University of Southern Queensland, Toowoomba, QLD 4350, Australia.

E. Kerr · B. L. Schulz, School of Chemistry and Molecular Biosciences, The University of Queensland, St Lucia, QLD 4067, Australia.

G. P. Fox, University of California, Davis, CA 95616, USA.

© 2023 Crown

Journal of Agricultural, Biological, and Environmental Statistics, Volume 29, Number 2, Pages 232–256

<https://doi.org/10.1007/s13253-023-00574-x>

Key Words: Barley; *k*-Means clustering; Linear mixed model; Mass spectrometry; Multi-phase design; Proteomics.

1. INTRODUCTION

Comparative experiments continue to provide the foundation of agricultural research and thus underpin the improvement and optimisation of the productivity of agricultural systems. Over time, large-scale increases in productivity have become harder to achieve (Fischer and Connor 2018), resulting in greater pressure being placed on the humble comparative experiment to yield more comprehensive and detailed information on the biological processes underpinning the system. Often, the answer to this pressure is to ‘measure more’; more variables, more frequently or in more detail. An outcome of measuring more is that comparative experiments give rise to datasets of a greater size and detail than ever before. However, while sizes of datasets are growing, the size of experiments typically remains the same, leading to a situation where, in many experiments, the number of variables measured is greater than the number of experimental units.

With growing size and detail, the complexity of the datasets also often increases. This complexity can arise from relationships between measured traits or characteristics (Dreccer et al. 2020), structure implicit in or imposed on the experimental material by design, sampling or measurement protocols (Brien and Bailey 2006; De Faveri et al. 2017) or other physical or biological factors inherent to the treatments or material in the experiment (Oakey et al. 2006; Osama et al. 2021).

Comparative experiments conducted in the laboratory often facilitate a ‘measure more’ approach, enabling a detailed investigation of biological samples obtained from the field or other sources. When these samples arise from observational studies or designed experiments, a multi-phase experiment, conducted according to a multi-phase design, is possible (Brien et al. 2011). The simplest multi-phase experiment consists of two phases (McIntyre 1955), where units composing the first phase produce outcomes, such as material and/or response variable values, before material from the first phase is randomised to units in the second phase (Brien and Bailey 2006; Brien et al. 2011). A multi-phase design is then the implementation of an experimental design solution for a multi-phase experiment. Such designs have been applied to great effect in the agricultural domain, enabling the partitioning and investigation of extraneous variation across different phases of experimentation (Smith et al. 2006; Panozzo et al. 2007; Kelly and Forknall 2020).

The linear mixed model (LMM) provides a powerful and flexible framework for the analysis of data arising from comparative experiments, and continues to be relevant and applicable even as the size and complexity of the datasets arising from these experiments increase. In part, this is due to the ability to estimate variance structures of complex form, both between treatments and between residual errors (De Faveri et al. 2015; Verbyla et al. 2021).

In practice, the estimation of complex variance structures relies on sufficient independent pieces of information to reliably estimate the covariances and variances required. In

situations where the number of variables exceeds the number of observations, the resulting variance structures can be of reduced rank, causing computational difficulties (Thompson et al. 2003). Structures that accommodate this reduced rank nature exist, for example, the factor analytic variance structure proposed by Smith et al. (2001b) and formulated for reduced rank estimation by Thompson et al. (2003). However, evidence suggests a degradation in the performance of this structure when the number of observations informing a variance parameter estimate is small (Macdonald 2018; Macdonald et al. 2019). In these situations, the modelling of covariance is often restricted to simplistic structures, such as that resulting from a variance components model (Patterson et al. 1977).

The LMM framework has also proven useful for the modelling of smooth trends in data arising from designed experiments. In cases where such trends display nonlinear forms, the LMM representation of the cubic smoothing spline has been formulated (Verbyla et al. 1999). This approach has been shown to be effective in modelling smoothly varying trends arising from designed agricultural experiments (Verbyla et al. 1999) and can be coupled with the estimation of complex covariance structures between residual errors (De Faveri et al. 2015, 2022).

There are multiple software solutions for the implementation of the LMM framework, both using standalone software or via statistical computing environments such as R (R Core Team 2019; Rogers and Taylor 2019). One of the more powerful and flexible options is the commercial `asreml` R package (Butler et al. 2017). This package implements variance component estimation via residual maximum likelihood (REML) (Patterson and Thompson 1971), using the average information algorithm (Gilmour et al. 1995), and supports the implementation of a wide range of complex variance–covariance structures, along with the LMM representation of the cubic smoothing spline.

While LMM software is well developed, currently, there is a distinct lack of software that supports the efficient implementation of highly complex variance–covariance structures. Despite the `asreml` package enabling the use of such structures, the time taken to fit models that include these structures, to even moderately sized datasets, is often prohibitive. Modelling run times are further exacerbated when the aim is to fit the models using a fully efficient single-stage analysis (Welham et al. 2010). An option to reduce computational issues and accelerate model fitting is to fit the LMM using a two-stage analysis (Smith et al. 2001a; Piepho et al. 2012).

A field of research in which laboratory-based comparative experiments are conducted, that produce large volumes of data from a limited number of samples, is mass spectrometry (MS)-based proteomics. Proteomics is the broad-scale investigation of the proteome of biological material, where the proteome is the set of proteins composing a biological sample (Yu et al. 2010). Standard modern proteomics workflows involve several sample processing steps to extract and digest the proteins composing a sample for measurement using MS (Osama et al. 2021). The data produced by these workflows have as their most fundamental building blocks the fragment ions linked to specific protein identities, along with their relative abundance (Gross 2011). Using knowledge of how ions comprise a peptide and how peptides comprise a protein enables the reconstruction of the abundance of each protein identified in a sample (Zhang et al. 2010).

In the context of field crops research, the application of MS-based proteomics is growing rapidly (Agrawal et al. 2013), with comparative experiments to test for differences in the proteome of plant tissues, as a result of treatments or plant developmental stages, becoming commonplace (Osama et al. 2021). Such experiments often result in the quantification of upwards of hundreds of proteins from the processing of an individual sample (Gross 2011), with the datasets generally characterised by complex relationships (correlations) between proteins (Agrawal et al. 2013; Robotti et al. 2015).

Mass spectrometry-based experiments can suffer from a lack of sound experimental design, with multiple authors reporting a need to improve the design of such studies (Hu et al. 2005; Oberg and Vitek 2009). Furthermore, the two-phase nature of MS studies is well suited for the implementation of multi-phase design solutions. Such solutions would facilitate the partitioning of variation associated with the collection of the biological material, from 'technical' variation that could arise during the subsequent processing of the material using MS techniques. However, we have found no reported occurrences in the literature of such a design solution being implemented in the conduct of an agriculturally motivated MS-based proteomics study.

Statistical methods for the exploration of proteomics datasets are dominated by pattern or cluster analysis and classification-based techniques (Robotti et al. 2015; Chen et al. 2020). Analysis methods to test for differences in proteome composition between treatments in MS-based proteomics experiments also vary and range from simple *t*-tests (Chen et al. 2020), to analysis of variance (Oberg et al. 2008), multivariate statistical techniques (Robotti et al. 2015) and machine learning approaches (Chen et al. 2020). Examples also exist of the LMM framework being used for the analysis of MS-based proteomics data (Oberg et al. 2008; Choi et al. 2014). However, these LMM frameworks are often simplistic (Osama et al. 2021), with likely advancements possible through estimation of complex covariance between proteins to characterise the nature of the relationships that exist within the proteome (Robotti et al. 2015).

Given the size and complexity of datasets now arising from the most simple of comparative experiments, exemplified by MS-based proteomics studies, the aim of this paper was to provide a parsimonious LMM framework for the analysis of designed experiments where datasets are characterised by a greater number of variables than experimental units, and for which the modelling of complex covariance structures between variables is desired. The proposed method is demonstrated through application to an MS-based proteomics experiment, the objective of which is to investigate changes in the proteome of barley grain during the malting process. We label our proposed method *covariance clustering* as, through extension of the traditional variance components model (Patterson et al. 1977), it allows for the modelling of covariance between variables by first clustering variables based on estimated effects and then introducing these covariance clusters into an LMM framework through an additional random term.

In order to address our aim, the paper proceeds as follows. To begin, the multi-phase MS-based proteomics experiment that forms the motivating example is introduced. Following this, the four-step procedure to implement the covariance clustering approach is given. The results from each step of the covariance clustering approach, as applied to the motivating

experiment, are then presented. The paper concludes with a discussion of the proposed covariance clustering method.

2. MOTIVATING EXAMPLE

A recent multi-phase MS-based proteomics experiment, with components previously reported by [Yousif and Evans \(2020\)](#) and [Osama et al. \(2021\)](#), provides a motivating example. In this experiment, MS was used to quantify the proteome composition of barley grain and malt samples, where samples were collected at different times during a commercial malting process. The aims of the experiment were to identify proteins that demonstrated a change in abundance over time in the malting process and characterise the relationship between abundance and time for these proteins.

In what we believe to be a first report in the literature, the multi-phase MS-based proteomics experiment was conducted according to a multi-phase design. This multi-phase design enables an investigation of variation arising in both phases of the experiment, being the (i) malt sample collection phase and (ii) MS processing phase. Both phases are explored in more detail, following a brief overview of the particulars of the malting processes relevant to the motivating example.

2.1. PARTICULARS OF THE MOTIVATING EXAMPLE

The malting process is typically conducted over approximately six days and involves the controlled and limited germination, then drying, of grain. The process initiates the expression and activation of enzymes that break down the complex carbohydrates and proteins contained within the grain endosperm for yeast utilisation in fermentation ([Yousif and Evans 2020](#); [Osama et al. 2021](#)). The malting process is achieved through three stages; steeping, germination and kilning ([Schwarz and Li 2010](#)). Barley is the most commonly malted grain, as barley malt is a key ingredient in traditional beer brewing ([Schwarz and Li 2010](#)).

The barley samples considered in this study are a subset of those acquired from an Australian commercial malting plant, labelled as Plant Two in [Yousif and Evans \(2020\)](#).

2.2. PHASE I: MALT SAMPLE COLLECTION

The collection of barley grain and malt samples constitutes the first phase of the motivating experiment. We consider the processing of two replicate batches of grain in the malting plant. For each replicate batch, grain was sampled a total of 12 times ($t = 12$), with the time at which each sample was collected consistent between the replicate batches (Table 1). Note that without the ability to randomise sampling times within replicate batches, this phase is observational in nature and not itself an experiment. Additionally, note that the sampling times presented in Table 1 do not correspond with those presented in Table 1 of [Yousif and Evans \(2020\)](#) due to an error in the measurement of kilning time previously reported. At each sampling time, samples were frozen and freeze-dried to less than 10% moisture if wet ([Yousif and Evans 2020](#)), before being stored for later processing using MS-based proteomics techniques.

Table 1. Times at which barley grain and malt samples were collected during Phase I of the motivating mass spectrometry-based proteomics experiment

Sampling time (j)	1	2	3	4	5	6	7	8	9	10	11	12
Time (h)	0.0	16.5	27.5	40.5	51.5	64.5	75.5	88.5	99.5	112.5	133.5	151.5

2.3. PHASE II: MASS SPECTROMETRY PROCESSING

The individual grain and malt samples collected in the first phase of the experiment were subsequently processed using MS in the second phase of the experiment. Prior to processing, two subsamples were taken from each of the 24 grain samples. Together, this provides a total of 48 subsamples and four independent samples ($z = 4$) of each sampling time.

Each subsample was prepared and processed according to the workflow outlined in [Osama et al. \(2021\)](#). To briefly summarise, the preparation of individual subsamples involved first grinding them into a homogeneous powder, before a series of steps were undertaken to chemically extract the proteins composing each subsample ([Osama et al. 2021](#)). The processing of the extracted proteins from each subsample was achieved using a Sequential Window Acquisition of all THEoretical ions Mass Spectrometry (SWATH-MS) proteomic analysis technique ([Osama et al. 2021](#)), which facilitated the ionisation and measurement of the proteins composing the subsamples.

Due to the nature of the MS process, subsamples needed to be run through the mass spectrometer sequentially, and a design was applied to the order in which the subsamples were processed. Two replicate blocks were defined, with one subsample from each grain sample processed within each replicate block, and each replicate block was composed of 24 subsamples. Within each replicate block, subsamples were assigned to one of four processing groups according to an incomplete block design (processing groups correspond to incomplete blocks), with the order in which subsamples were processed randomised within each processing group. The blocking of subsamples into processing groups was performed to enable adjustment for potential extraneous variability or trend arising in the sequential MS processing of subsamples. The initial incomplete block design randomisation was optimised using the `odw` R package ([Butler 2022](#)), which searches for an optimal design given the specification of an associated LMM. Following this optimised design, all subsamples were processed using MS.

The proteins composing each subsample were identified and quantified using the methods described in [Osama et al. \(2021\)](#). This involved identifying the proteins by matching the measured protein signatures with those existing in a published proteomics database, before the abundance of the proteins was determined using proprietary software ([Kerr et al. 2019](#); [Osama et al. 2021](#)).

Following this process, a consistent set of 617 proteins ($p = 617$) were identified from each subsample and the abundance of each protein i ($i = 1, \dots, 617$) quantified. This resulted in the motivating dataset consisting of a total of $n = tzp = 29,616$ protein abundance observations. The proteins ranged in raw abundance, from 69 units per subsample to 4,687,460 units per subsample, with a median of 17,499, where the unit of measure-

ment is related to MS intensity and the abundance of constituent ions composing a protein. The dataset arising from the motivating mass spectrometry-based proteomics experiment is available through the ProteomeXchange Consortium via the Proteomics Identifications Database (PRIDE) partner repository (Perez-Riverol et al. 2021), using the dataset identifier PXD019384 (<https://www.ebi.ac.uk/pride/>).

3. STATISTICAL METHODS

Methods for the implementation of covariance clustering are presented in the context of their application to the motivating example. The covariance clustering method is implemented using four steps; (1) fitting a baseline model, (2) forming a series of covariance clusters, (3) fitting a series of covariance cluster models, one for each of the plausible covariance clusters identified, and (4) locating the ‘optimal’ number of covariance clusters and fitting an ‘optimal’ covariance cluster model. Figure 1 presents these steps in the two alternate ways in which they are applied in the analysis of the motivating experiment; the first models residual covariance through forming what we label *residual* covariance clusters, and the second identifies similarly shaped response curves through forming what we label *spline* covariance clusters. In this case, a two-stage LMM framework is used to analyse the raw protein abundance data, with a different covariance clustering procedure implemented in each of the respective stages of the analysis (Fig. 1).

3.1. STAGE 1

The first stage of the analysis involves the modelling of the raw protein abundance data for all p proteins. Protein abundance is log transformed ($\log(x/1000 + 1)$) including an offset of 1 as values of $x/1000$ are close to zero, and because diagnostics obtained from a preliminary analysis showed the LMM adhered better to the assumptions of normality and homogeneity of residual variance using this transformation. Note that the selection of this offset needs careful consideration in practice (Welham et al. 2014). Following transformation, the minimum, median and maximum of the log transformed data are, respectively, 0.067, 2.92, and 8.45. Notably, it is assumed the abundance observations are ordered as subsamples, within sampling times, within proteins.

Step 1.1: Fit baseline linear mixed model: The general form of the baseline LMM fit to the protein abundance data is written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\tau} + \mathbf{Z}_d\mathbf{u}_d + \mathbf{e}, \quad (3.1)$$

where \mathbf{y} is an $n \times 1$ vector of (transformed) protein abundance observations and $\boldsymbol{\tau}$ is a $pt \times 1$ vector containing fixed effects for each protein \times sampling time combination, with associated design matrix \mathbf{X} . The vector \mathbf{u}_d contains random effects corresponding to the multi-phase experimental design structure, with associated design matrix \mathbf{Z}_d , and \mathbf{e} is an $n \times 1$ vector of residual error effects.

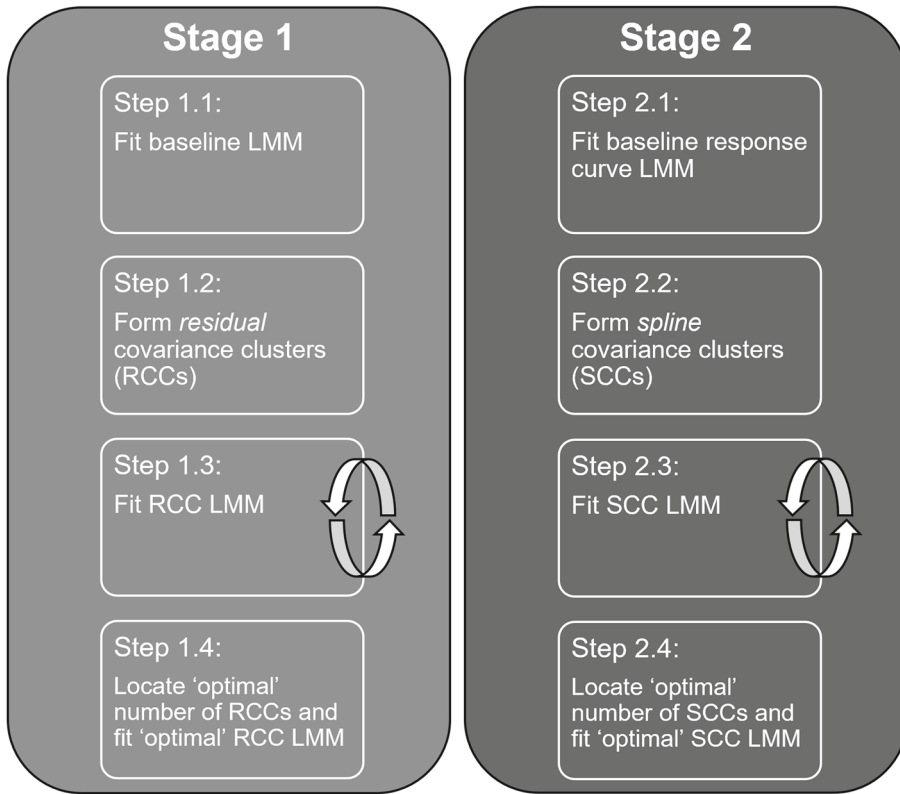


Figure 1. The four steps involved in the covariance clustering procedure ($x.1$ to $x.4$) are implemented in each of two stages ($x = 1, 2$) of a linear mixed model (LMM) framework, to analyse the motivating mass spectrometry-based proteomics experiment. In Stage 1, what we have labelled residual covariance clusters (RCCs) are formed to model residual covariance between proteins. In Stage 2, what we have labelled spline covariance clusters (SCCs) are formed to identify similarly shaped response curves describing the relationship between protein abundance and time in the malting process.

The random effects and residual error effects from (3.1) are assumed to follow a normal distribution with zero mean vector and variance–covariance matrix

$$\text{var} \left(\begin{bmatrix} \mathbf{u}_d \\ \mathbf{e} \end{bmatrix} \right) = \begin{bmatrix} \mathbf{G}_d & \mathbf{0} \\ \mathbf{0} & \mathbf{R} \end{bmatrix}.$$

Of note is the form of \mathbf{R} , an $n \times n$ matrix, which is defined using a three-way separable structure,

$$\mathbf{R} = \mathbf{R}_p \otimes \mathbf{R}_t \otimes \mathbf{I}_z, \tag{3.2}$$

where \otimes is the Kronecker product, $\mathbf{R}_p = \oplus_{i=1}^p \sigma_{e_{p_i}}^2$ is a $p \times p$ diagonal matrix, enabling the estimation of heterogeneous residual variance for each protein i , $\mathbf{R}_t = \oplus_{j=1}^t \sigma_{e_{t_j}}^2$ is a $t \times t$ diagonal matrix, enabling the estimation of heterogeneous residual variance for each sampling time j , and \mathbf{I}_z is a $z \times z$ identity matrix. In practice, to ensure identifiability, $\sigma_{e_{t_j}}^2$

is replaced in the form of \mathbf{R}_t with $\gamma_{e_{t_j}}$, where $\gamma_{e_{t_j}}$ is a scaling parameter and a single $\gamma_{e_{t_j}}$ is constrained to equal 1.

The form of \mathbf{R} presented in (3.2) enables the estimation of heterogeneous residual variance for each protein, while allowing for the differential scaling of these variances across sampling times. However, such a form assumes independence between proteins at the residual level, with any covariance built through the variance of the experimental design terms included in \mathbf{G}_d . This is a potentially restrictive and limiting assumption, given the complex relationships that are known to be present in the proteome (Robotti et al. 2015). As such, (3.1) can be extended to enable the modelling of additional and more complex covariance between proteins at the residual level.

Step 1.2: Form residual covariance clusters: *Residual* covariance clusters are formed by clustering proteins based on their estimated residuals ($\tilde{\mathbf{e}}$) from (3.1). These residuals are obtained from (3.1) as empirical best linear unbiased predictors (e-BLUPs), before being studentised (Gilmour et al. 2015, p. 17).

A k -means clustering algorithm (Hartigan and Wong 1979) is used to generate a range of plausible residual covariance clusters. This involves the grouping of proteins into increasing numbers of residual clusters ($v_r = 2, \dots, q_r$), with the total number of potential cluster groupings labelled as η_r . The k -means algorithm is repeated multiple times, each with a different starting seed for the randomisation process defining the initial cluster allocation. If the total number of seeds considered is labelled ϕ_r , then this results in the formation of $k_r = \phi_r \times \eta_r$ plausible residual covariance cluster groupings ($l_r = 1, \dots, k_r$), based on all seed \times number of cluster combinations.

Step 1.3: Fit residual covariance cluster linear mixed model: To incorporate a plausible residual covariance cluster, (3.1) is extended to include an additional random term containing effects corresponding to an individual seed \times number of cluster combination, h_{l_r} . This model, referred to as the residual covariance cluster LMM, is repeatedly fit whereby the random cluster term is updated to consider each plausible residual covariance cluster. The general form of the model is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\tau} + \mathbf{Z}_d\mathbf{u}_d + \mathbf{Z}_{c_r}\mathbf{u}_{c_r} + \mathbf{e}, \quad (3.3)$$

where \mathbf{u}_{c_r} is an $(v_r t z \times 1)$ vector of random effects corresponding to the $h_{l_r}^{\text{th}}$ residual covariance cluster \times sampling time combination, with design matrix \mathbf{Z}_{c_r} . All other terms are as defined for the model in (3.1).

The random effects and residual error effects from (3.3) are assumed to follow a normal distribution with zero mean vector and variance–covariance matrix

$$\text{var} \left(\begin{bmatrix} \mathbf{u}_d \\ \mathbf{u}_{c_r} \\ \mathbf{e} \end{bmatrix} \right) = \begin{bmatrix} \mathbf{G}_d & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_{c_r} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{R} \end{bmatrix},$$

where \mathbf{R} is as defined for the model in (3.2) and \mathbf{G}_{c_r} is defined using the three-way separable form

$$\mathbf{G}_{c_r} = \mathbf{G}_g \otimes \mathbf{G}_t \otimes \mathbf{I}_z, \quad (3.4)$$

where $\mathbf{G}_g = \bigoplus_{m=1}^{\nu_r} \sigma_{c_m}^2$ is a $\nu_r \times \nu_r$ diagonal matrix, enabling the estimation of heterogeneous variance for each residual covariance cluster m , and $\mathbf{G}_t = \bigoplus_{j=1}^t \gamma_{c_{t_j}}$ is a $t \times t$ diagonal matrix, containing scaling parameters for the residual covariance cluster variances across sampling times.

The inclusion of the residual covariance cluster effects, with the associated variance model in (3.4), acts to induce greater covariance between proteins within a covariance cluster, along with providing greater heterogeneity of variance between clusters.

Step 1.4: Locate ‘optimal’ number of residual covariance clusters and fit ‘optimal’ residual covariance cluster linear mixed model: Upon fitting (3.3) for each seed \times number of cluster combination, the AIC based on the full log-likelihood (Verbyla 2019) is obtained for each model. For brevity, all further mentions of the AIC refers to that based on the full log-likelihood. The ‘optimal’ number of residual covariance clusters, labelled ν_r^* , is that which minimises the AIC, on average across seeds, and thus provides the most parsimonious residual covariance structure.

Upon determining ν_r^* , the AIC of the models fit for each seed is compared within ν_r^* , to locate the seed which results in a clustering of proteins that further minimised the AIC. This ‘optimal’ seed is labelled ω_r^* . The grouping of proteins into ν_r^* clusters, arising from seed ω_r^* , are reintroduced into (3.3) and the model fit to obtain the ‘optimal’ residual covariance cluster LMM, (3.3*).

Upon fitting (3.3*), predictions of abundance for each protein at each sampling time ($\hat{\tau}$) are obtained from the model as empirical best linear unbiased estimators (e-BLUEs). Also obtained from (3.3*) is \mathbf{w} , a $p t \times 1$ vector of weights, corresponding to the diagonal elements of the inverse of the fixed effect variance–covariance matrix (Smith et al. 2001a). The e-BLUEs and associated weights are carried forward to the second stage of analysis.

3.2. STAGE 2

The second stage of the analysis process involves the estimation of response curves to describe the response of proteins to time in the malting process (Fig. 1). In order to allow for nonlinearity in the response of protein abundance, the LMM representation of the cubic smoothing spline can be exploited (Verbyla et al. 1999). Note that the abundance e-BLUEs obtained from (3.3*) are assumed to be ordered according to sampling times, within proteins.

Step 2.1: Fit baseline response curve linear mixed model: The general form of the model is

$$\hat{\tau} = \mathbf{X}_{s_0} \boldsymbol{\beta}_0 + \mathbf{Z}_{s_0} \mathbf{u}_{s_0} + \mathbf{Z}_f \mathbf{u}_f + \mathbf{X}_{s_p} \boldsymbol{\beta}_p + \mathbf{Z}_{s_p} \mathbf{u}_{s_p} + \mathbf{e}_s, \tag{3.5}$$

where $\hat{\tau}$ is a $p t \times 1$ vector containing the abundance e-BLUEs for each protein \times sampling time combination obtained from (3.3*). The terms $\mathbf{X}_{s_0} \boldsymbol{\beta}_0$, $\mathbf{Z}_{s_0} \mathbf{u}_{s_0}$ and $\mathbf{Z}_f \mathbf{u}_f$ correspond to the fitting of an overall (main effect) spline to model the nonlinear response of abundance to sampling time, where $\boldsymbol{\beta}_0 = [\beta_0 \ \beta_1]^T$, a 2×1 vector, contains fixed regression coefficients, \mathbf{u}_{s_0} , a $(t - 2) \times 1$ vector, contains random cubic smoothing spline coefficients and \mathbf{u}_f , a $t \times 1$ vector, enables the estimation of random non-smooth effects that may arise due to replicated sampling at each sampling time (Verbyla et al. 1999). The design matrices accompanying

these vectors are $\mathbf{X}_{s_0} = \mathbf{1}_p \otimes \mathbf{X}_{s_t}$, $\mathbf{Z}_{s_0} = \mathbf{1}_p \otimes \mathbf{Z}_{s_t}$ and \mathbf{Z}_f , respectively, where $\mathbf{X}_{s_t} = [\mathbf{1}_t \ \mathbf{x}]$, \mathbf{Z}_{s_t} is a $t \times (t - 2)$ spline design matrix as defined in Verbyla et al. (2018), $\mathbf{1}_p$ and $\mathbf{1}_t$ are vectors of ones of length p and t , respectively, and \mathbf{x} is a $t \times 1$ vector containing the t sampling times. The terms $\mathbf{X}_{s_p} \boldsymbol{\beta}_p$ and $\mathbf{Z}_{s_p} \mathbf{u}_{s_p}$ allow for the estimation of nonlinear protein specific spline responses. The vectors $\boldsymbol{\beta}_p = \text{vec} \left(\begin{bmatrix} \boldsymbol{\beta}_{p_0}^\top \\ \boldsymbol{\beta}_{p_1}^\top \end{bmatrix} \right)$, a $2p \times 1$ vector, and \mathbf{u}_{s_p} , a $p(t - 2) \times 1$ vector, contain protein specific fixed regression and random spline coefficients, respectively, with associated design matrices $\mathbf{X}_{s_p} = \mathbf{I}_p \otimes \mathbf{X}_{s_t}$ and $\mathbf{Z}_{s_p} = \mathbf{I}_p \otimes \mathbf{Z}_{s_t}$, where \mathbf{I}_p is a $p \times p$ identity matrix. The vector \mathbf{e}_s , of length $pt \times 1$, contains the residual error effects.

The random effects and residual error effects from (3.5) are assumed to follow a normal distribution with zero mean vector and variance–covariance given by

$$\text{var} \left(\begin{bmatrix} \mathbf{u}_{s_0} \\ \mathbf{u}_f \\ \mathbf{u}_{s_p} \\ \mathbf{e}_s \end{bmatrix} \right) = \begin{bmatrix} \mathbf{G}_0 & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_f & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{G}_p & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{R}_s \end{bmatrix},$$

where $\mathbf{G}_0 = \sigma_0^2 \mathbf{I}_{t-2}$, $\mathbf{G}_f = \sigma_f^2 \mathbf{I}_t$, $\mathbf{G}_p = \sigma_p^2 \mathbf{I}_{p(t-2)}$, and σ_0^2 and σ_p^2 are inversely related to the smoothing parameters associated with the overall, and protein specific, nonlinear responses to time in the malting process, respectively. The matrix \mathbf{R}_s has diagonal elements given by the weights vector \mathbf{w} .

Step 2.2: Form spline covariance clusters: *Spline* covariance clusters are formed by clustering proteins based on their estimated protein specific cubic smoothing splines coefficients ($\tilde{\mathbf{u}}_{s_p}$), obtained from (3.5) as e-BLUPs. Using these effects, the procedure documented in Step 1.2 is performed to establish a range of plausible spline covariance clusters, where $v_s = 2, \dots, q_s$ are the number of spline clusters formed, η_s and ϕ_s are the total number of spline cluster groupings and random starting seeds considered, respectively, and $k_s = \phi_s \times \eta_s$ is the total number of plausible spline covariance clusters ($l_s = 1, \dots, k_s$).

Step 2.3: Fit spline covariance cluster linear mixed model: To incorporate a plausible spline covariance cluster, (3.5) is extended to include an additional random term, allowing for the estimation of cluster specific nonlinear responses to time in the malting process, where the clusters correspond to a particular seed \times number of cluster combination (h_{l_s}). This model, labelled as the spline covariance cluster LMM, is repeatedly fit whereby the cluster random term is updated to consider each plausible spline covariance cluster. The general form of the model, assuming the data are ordered as sampling times, then proteins, then covariance clusters, is

$$\hat{\boldsymbol{\tau}} = \mathbf{X}_{s_0} \boldsymbol{\beta}_0 + \mathbf{Z}_{s_0} \mathbf{u}_{s_0} + \mathbf{Z}_f \mathbf{u}_f + \mathbf{X}_{s_p} \boldsymbol{\beta}_p + \mathbf{Z}_{s_p} \mathbf{u}_{s_p} + \mathbf{Z}_{c_s} \mathbf{u}_{c_s} + \mathbf{e}_s, \quad (3.6)$$

where \mathbf{u}_{c_s} is a $v_s(t - 2) \times 1$ vector of random cubic smoothing spline coefficients for the v_s spline covariance clusters fitted, with spline design matrix $\mathbf{Z}_{c_s} = \bigoplus_{m=1}^{v_s} \mathbf{1}_{p_m} \otimes \mathbf{Z}_{s_t}$, where p_m is the number of proteins in spline covariance cluster m . All other terms are as defined for the model in (3.5).

The random effects and residual error effects from (3.6) are assumed to follow a normal distribution with zero mean vector and variance–covariance matrix

$$\text{var} \left(\begin{bmatrix} \mathbf{u}_{s_o} \\ \mathbf{u}_f \\ \mathbf{u}_{s_p} \\ \mathbf{u}_{c_s} \\ \mathbf{e}_s \end{bmatrix} \right) = \begin{bmatrix} \mathbf{G}_o & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_f & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{G}_p & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{G}_{c_s} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{R}_s \end{bmatrix},$$

where the forms of \mathbf{G}_o , \mathbf{G}_f , \mathbf{G}_p and \mathbf{R}_s are as defined for the model in (3.5), $\mathbf{G}_{c_s} = \sigma_{c_s}^2 \mathbf{I}_{v_s(t-2)}$ is a $v_s(t-2) \times v_s(t-2)$ diagonal matrix and $\sigma_{c_s}^2$ is inversely related to the smoothing parameter of the spline covariance cluster specific nonlinear responses to time in the malting process.

Step 2.4: Locate ‘optimal’ number of spline covariance clusters and fit ‘optimal’ spline covariance cluster linear mixed model: The procedure outlined in Step 1.4 is repeated to determine the ‘optimal’ number of spline covariance clusters, where v_s^* is the number of covariance clusters which minimised the AIC, on average across seeds, and ω_s^* is the seed which further minimises the AIC within v_s^* . The protein groupings resulting from ω_s^* with v_s^* clusters are then reintroduced into (3.6) and the model fit to obtain the ‘optimal’ spline covariance cluster LMM, (3.6*), being the model which results in the estimation of the most parsimonious set of ‘typical’ response profiles.

Following the implementation of (3.6*), e-BLUPs of protein abundance using the ‘typical’ protein response profiles are obtained as $\mathbf{X}_{s_o}^* \hat{\boldsymbol{\beta}}_o + \mathbf{Z}_{s_o}^* \tilde{\mathbf{u}}_{s_o} + \mathbf{X}_{s_p}^* \hat{\boldsymbol{\beta}}_p + \mathbf{Z}_{c_s}^* \tilde{\mathbf{u}}_{c_s}$, where $\mathbf{X}_{s_o}^*$, $\mathbf{Z}_{s_o}^*$, $\mathbf{X}_{s_p}^*$ and $\mathbf{Z}_{c_s}^*$ are design matrices formed using the malting times for which protein abundance predictions are sought, with forms described in Verbyla et al. (2021). Additionally, the nonlinear relationships between abundance and time in the malting process at the spline covariance cluster, or ‘typical’ response profile, level can be explored by estimating the e-BLUPs $\mathbf{Z}_{c_s}^* \tilde{\mathbf{u}}_{c_s}$.

3.3. SOFTWARE

All models are fit using the `asreml` package (Butler et al. 2017) in the R statistical computing environment (R Core Team 2019). The AIC based on the full log-likelihood is derived using the `iCREML` function (Verbyla 2019). The k -means clustering approaches are implemented using the `kmeans` package (R Core Team 2019). Functions used to implement the covariance clustering approach are available from <https://github.com/ClaytonForknall/CovarianceClustering>

4. APPLICATION OF METHOD

The motivating MS-based proteomics experiment is used to illustrate the application of the covariance clustering method (Fig. 1). Throughout this section, a consistent subset of 25 proteins are presented. These proteins were selected to span the range of differential response types to time in the malting process. The extent of these differential responses

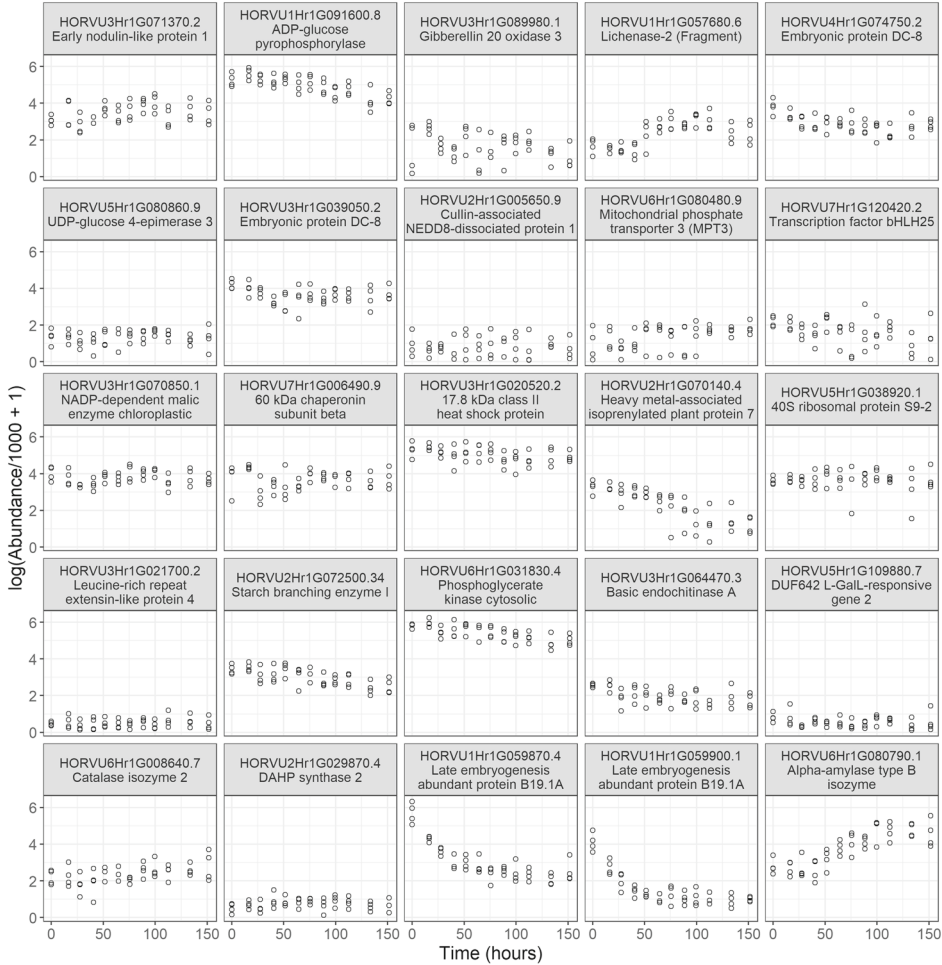


Figure 2. Raw abundance of a subset of 25 proteins quantified using mass spectrometry proteomics techniques from barley grain and malt samples. Samples were taken at different times during the malting process, as part of the motivating mass spectrometry-based proteomics experiment. Proteins are labelled using their unique identifier (HORVU code), followed by their common name.

is shown in Fig. 2, which presents the raw measured abundance for the selected subset of proteins, quantified in subsamples collected at each sampling time during the malting process (Table 1).

Step 1.1: Fit baseline linear mixed model: The fitting of the baseline LMM revealed significant heterogeneity of residual variance, both between proteins and sampling times. Residual variances for proteins ($\sigma_{e_{p_i}}^2$) ranged from 0.005 to 1.45, while the sampling time scaling parameters (γ_{e_j}) ranged from 0.88 to 1.17 relative to sampling time $j = 2$ which was constrained ($\gamma_{e_2} = 1$).

The baseline LMM also revealed non-negligible variation arising in both phases of the motivating experiment. Table 2 presents the variance component estimates of the terms included in the baseline LMM to account for the multi-phase experimental design struc-

Table 2. Variance component estimates corresponding to the terms included in the baseline linear mixed model (Step 1.1) to account for the multi-phase experimental design structure of the motivating mass spectrometry-based proteomics experiment

Term	Variance component
MaltRep	0.001
MaltTime:MaltRep	0.012
LabRep	0.002
LabRep:LabBlock	0.034
MaltTime:MaltRep:Subsample	0.041
Protein:MaltRep	0.001
Protein:MaltTime:MaltRep	1.168×10^{-8}
Protein:LabRep	0.002
Protein:LabRep:LabBlock	0.001

ture ($\tilde{\mathbf{u}}_d$). Terms describing the first phase of the experiment include `MaltRep`, labelling the replicate batches of malt processed in the malting plant, and `MaltTime:MaltRep`, indexing each unique grain sample collected during the malt sample collection phase. Terms describing the second phase of the experiment include `LabRep`, `LabRep:LabBlock` and `MaltTime:MaltRep:Subsample`, which describe the structured approach taken for the processing of the barley grain subsamples. All terms are fit separately and in combination with `Protein`, the latter terms included to capture potential variation between proteins in combination with these structural terms.

The two largest sources of extraneous variation arose during the MS processing phase and corresponded to variation between individual laboratory subsamples (`MaltTime:MaltRep:Subsample`), and variation between the processing groups of subsamples within a `LabRep` (`LabRep:LabBlock`), respectively (Table 2). Variation between proteins and the structural terms were of a smaller magnitude, with one of the four terms involving `Protein` estimated at the boundary of the variance parameter space (~ 0), being `Protein:MaltTime:MaltRep`.

Step 1.2: Form residual covariance clusters: Proteins were grouped into $\eta_r = 28$ different cluster groupings, with the number of clusters considered ranging from 2 to 533 ($2 \leq \nu_r \leq 533$) according to a geometric growth model, with a growth rate of 0.2. The grouping of proteins into clusters was repeated $\phi_r = 10$ times, each time with a different starting seed. This resulted in $k_r = 280$ plausible residual covariance cluster groupings, based on all seed \times number of cluster combinations.

Steps 1.3 & 1.4: Fit residual covariance cluster linear mixed model; Locate ‘optimal’ number of residual covariance clusters and fit ‘optimal’ residual covariance cluster linear mixed model: Fig. 3a presents the AIC for all $k_r = 280$ plausible residual covariance cluster models fit, along with the AIC from the baseline LMM. This shows that the AIC decreases sharply with the inclusion of more residual covariance clusters in the model, and is minimised with the inclusion of $\nu_r^* = 147$ residual covariance clusters (vertical dashed line, Fig. 3a). Within $\nu_r^* = 147$, the seed which further minimised the AIC was $\omega_r^* = 16630$. Increasing the number of residual covariance clusters beyond 147 results in poorer model

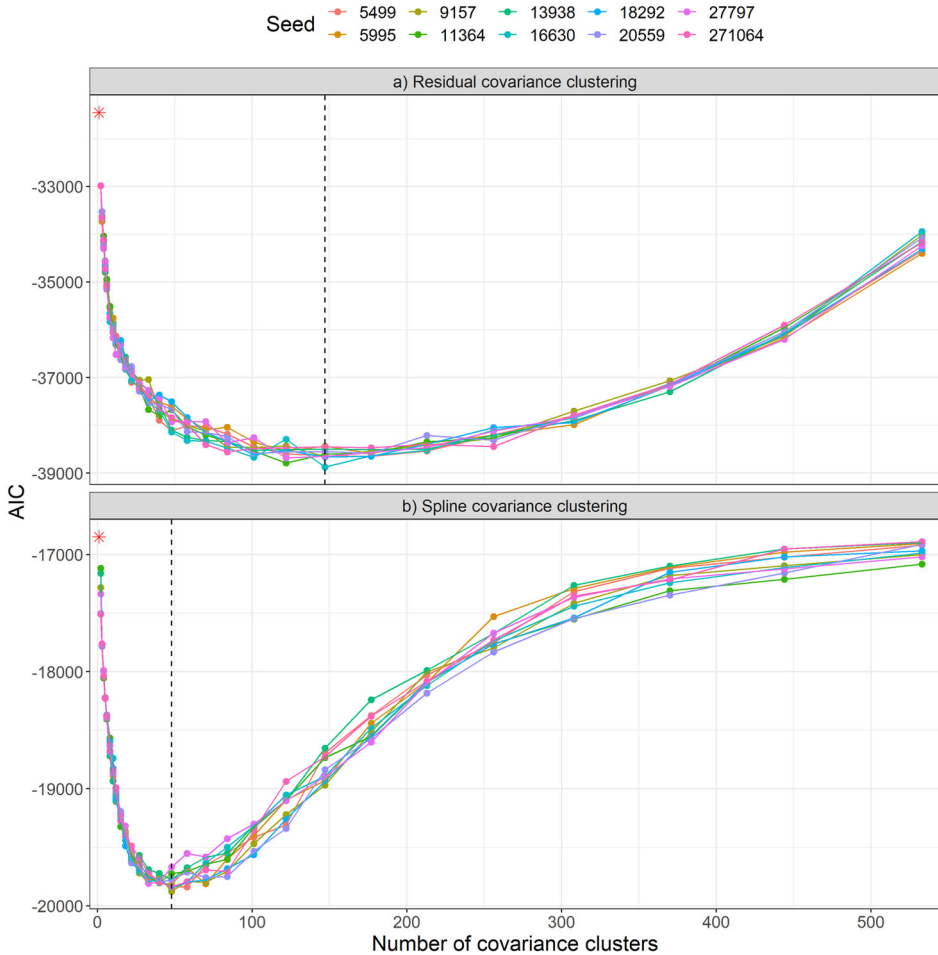


Figure 3. Sample paths of AIC resulting from the fitting of the **a** residual and **b** spline covariance cluster linear mixed models, respectively, for each seed \times number of cluster combination (Steps 1.3 and 2.3, respectively). Coloured points and lines correspond to the result from each of the ten seeds considered, for each covariance cluster model. The AIC corresponds to the Akaike Information Criterion, derived from the full log-likelihood (Verbyla 2019). The dashed lines correspond to the number of covariance clusters that minimised the AIC, on average across the seeds considered, for each of the covariance cluster models (147 and 48, respectively). The red stars correspond to the AIC obtained from the **a** baseline linear mixed model and **b** baseline response curve linear mixed model, respectively (Color figure online).

fit, demonstrated by the AIC increasing. This trend is apparent across all ten seeds, although there is some variability in individual model fits. When compared with the baseline LMM (red star, Fig. 3a), it is seen that the inclusion of any number of residual covariance clusters results in a substantial reduction in AIC.

Using $v_r^* = 147$ and $\omega_r^* = 16630$, the ‘optimal’ residual covariance cluster LMM was fitted. This model revealed heterogeneity of variance between residual covariance clusters, with cluster variances ($\sigma_{c_{rm}}^2$) ranging from 0.005 to 0.21, while the accompanying sampling time scaling parameters ($\gamma_{c_{rj}}$) ranged from 0.82 to 1.39 (relative to sampling time $j = 2$; $\gamma_{c_{r2}} = 1$). The inclusion of residual covariance clusters reduced the magnitude and range

Table 3. Variance component estimates corresponding to the terms included in the baseline response curve linear mixed model (Step 2.1). The use of `spl()` represents a term involving a cubic smoothing spline, as per the notation introduced in Verbyla et al. (1999)

Term	Variance component
<code>spl(malttime)</code>	1.069×10^{-7}
<code>MaltTime</code>	0.011
<code>Protein:spl(malttime)</code>	0.004
<code>Protein:MaltTime</code>	0.006

of residual protein variances, compared to the baseline LMM, with these variances ranging between 4.08×10^{-8} (estimated at the boundary of the parameter space) and 1.31, with the accompanying sampling time scaling parameters ranging from 0.88 to 1.07 (relative to sampling time $j = 2$; $\gamma_{e_2} = 1$).

Step 2.1: Fit baseline response curve linear mixed model: The baseline response curve LMM confirmed that there is non-negligible variation in the nonlinear responses of proteins to time in the malting process (Table 3), with the estimated variance component corresponding to the `Protein:spl(malttime)` term ($\tilde{\sigma}_p^2$) being nonzero.

Step 2.2: Form spline covariance clusters: Proteins were grouped into $\eta_s = 28$ different cluster groupings, with the number of clusters considered spanning $2 \leq \nu_s \leq 533$ according to a geometric growth model (growth rate of 0.2). The same ten starting seeds ($\phi_s = 10$) used in Step 1.2 were used to explore the impact of starting conditions of the clustering procedure. This resulted in $k_s = 280$ plausible spline covariance cluster groupings, based on all seed \times number of cluster combinations.

Steps 2.3 & 2.4: Fit spline covariance cluster linear mixed model; Locate ‘optimal’ number of spline covariance clusters and fit ‘optimal’ spline covariance cluster linear mixed model: The AIC decreases rapidly with the inclusion of spline covariance clusters and is minimised when $\nu_s^* = 48$ clusters are included in the model (vertical dashed line, Fig. 3b). Within $\nu_s^* = 48$, the seed which resulted in the minimum AIC was $\omega_s^* = 9157$. As the number of clusters increases beyond $\nu_s^* = 48$, the AIC increases, indicating a reduction in quality of model fit, and approaches that of the baseline model (red star, Fig. 3b) as the number of clusters increase towards the number of proteins ($p = 617$). This same trend is apparent across all ten seeds.

Using $\nu_s^* = 48$ and $\omega_s^* = 9157$, the ‘optimal’ spline covariance cluster LMM was fitted. Table 4 presents the variance component estimates from the model and shows that, with the inclusion of spline covariance clusters, the `Protein:spl(malttime)` term is estimated at the boundary of the parameter space, while the `Cluster:spl(malttime)` term is the dominate source of variation.

The shape and magnitude of the ‘typical’ nonlinear response profiles ($\mathbf{Z}_{c_s}^* \tilde{\mathbf{u}}_{c_s}$) vary substantially, with some showing only small deviations from zero (spline covariance cluster 16), while others show larger departures (e.g. spline covariance clusters 45 and 11) (Fig. 4).

The impact of the ‘typical’ response profiles is clear when investigating the estimated response curves ($\mathbf{X}_{s_0}^* \hat{\boldsymbol{\beta}}_o + \mathbf{Z}_{s_0}^* \tilde{\mathbf{u}}_{s_0} + \mathbf{X}_{s_p}^* \hat{\boldsymbol{\beta}}_p + \mathbf{Z}_{c_s}^* \tilde{\mathbf{u}}_{c_s}$) from the ‘optimal’ spline covariance

Table 4. Variance component estimates corresponding to the terms included in the ‘optimal’ spline covariance cluster linear mixed model (Step 2.4). The use of `spl()` represents a term involving a cubic smoothing spline, as per the notation introduced in [Verbyla et al. \(1999\)](#)

Term	Variance component
<code>spl(malttime)</code>	1.069×10^{-7}
<code>MaltTime</code>	0.012
<code>Cluster:spl(malttime)</code>	0.036
<code>Protein:spl(malttime)</code>	6.762×10^{-9}
<code>Protein:MaltTime</code>	0.001

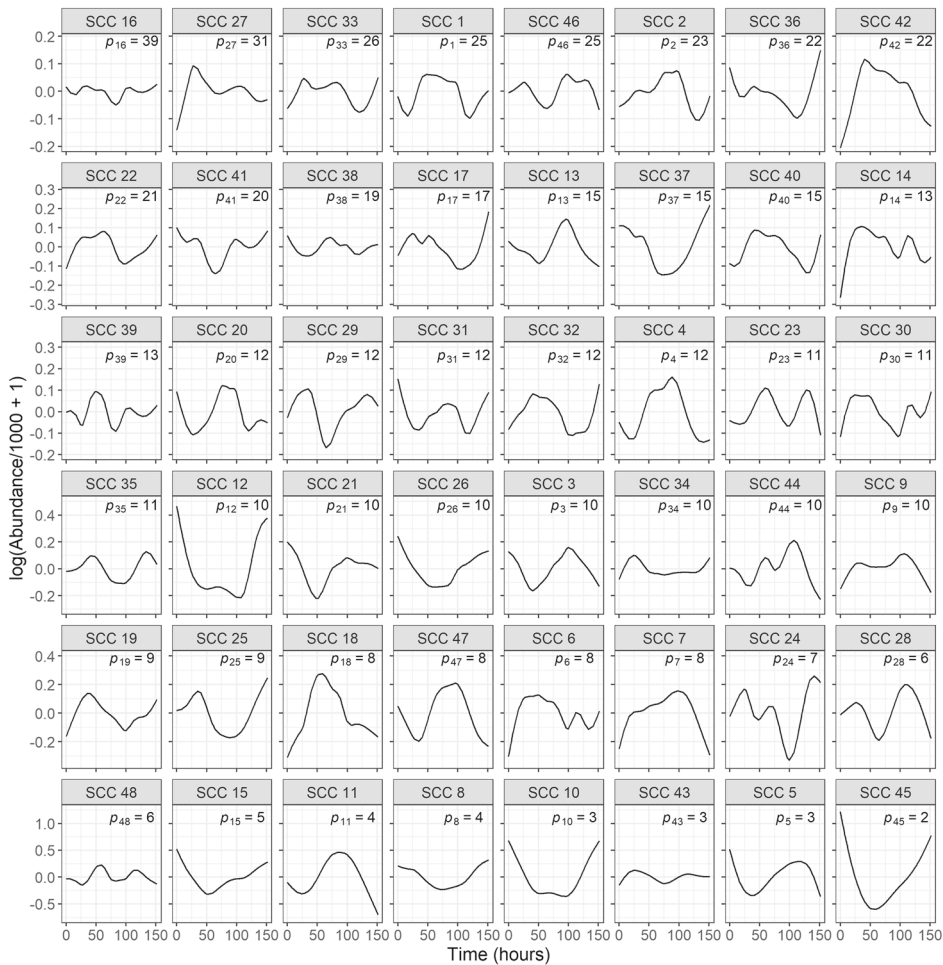


Figure 4. Estimated ‘typical’ response profiles, estimated using the spline covariance cluster (SCC) specific effects from the ‘optimal’ spline covariance cluster linear mixed model (Step 2.4). Labels in the top right-hand corner of each facet present the number of proteins belonging to each spline covariance cluster (p_m). Spline covariance clusters are ordered from largest number of proteins to smallest. Note the magnitude of the abundance axis varies from row to row in the figure.

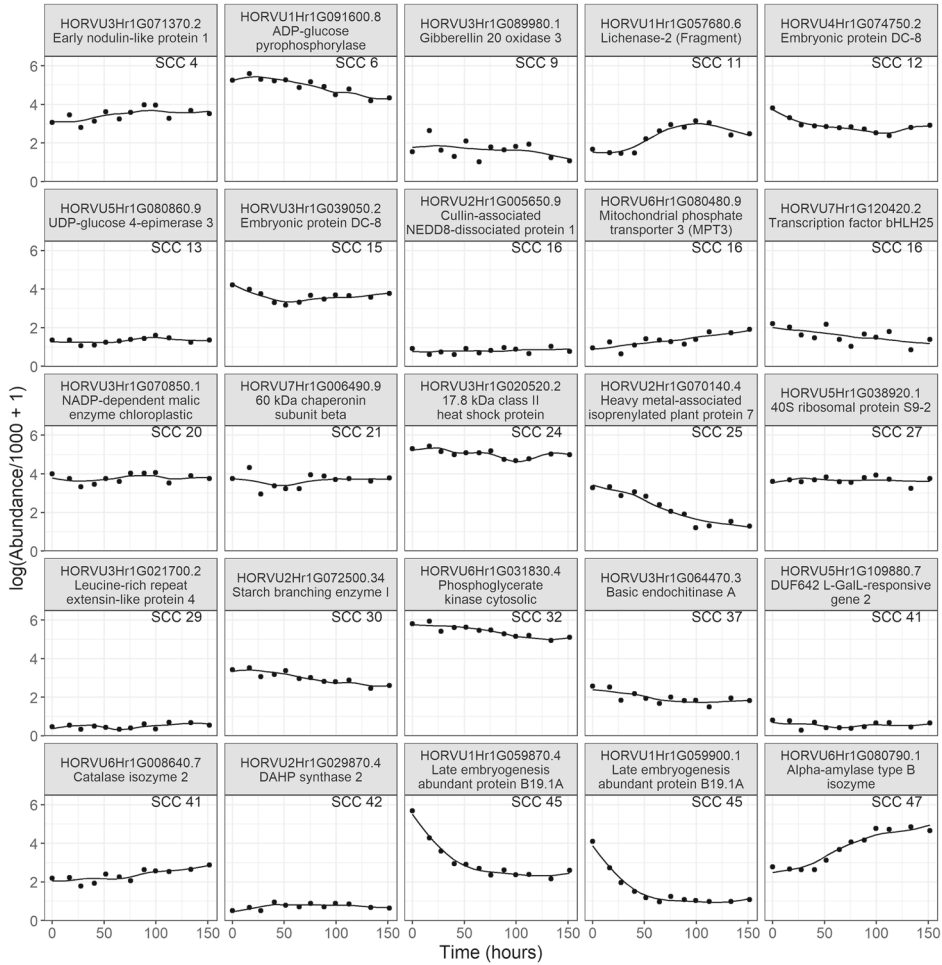


Figure 5. Estimated response curves from the ‘optimal’ spline covariance cluster linear mixed model (Step 2.4), describing the relationship between abundance and time in the malting process, for a subset of 25 proteins. Proteins are labelled using their unique identifier (HORVU code), followed by their common name. Estimates obtained from the response curves are empirical best linear unbiased predictors. Solid black dots correspond to the empirical best linear unbiased estimators of protein abundance obtained from the ‘optimal’ residual covariance cluster linear mixed model (Step 1.4). Cluster labels in the top right-hand corner of each facet present the spline covariance cluster (SCC) to which each protein belongs.

cluster LMM (Fig. 5). For example, although the two proteins composing spline covariance cluster 45 vary in mean abundance, the shape of the respective response curves of the proteins is the same and is determined by the response profile they share by being in the same spline covariance cluster. Another example of this arises for the three proteins presented from spline covariance cluster 16, where, although the proteins demonstrate different responses to time in the malting process (constant, increasing and decreasing responses), all the proteins share a common nonlinear response profile.

5. DISCUSSION AND CONCLUSIONS

We have developed and deployed a method, labelled covariance clustering, that enables the analysis of designed experiments characterised by greater numbers of variables than experimental units, in an LMM framework. The method features a twist on the traditional variance component model (Patterson et al. 1977), building parsimonious covariance between variables by first clustering variables based on estimated effects, and then introducing these covariance clusters into the LMM through an additional random term. In this way, complex covariance between variables can be estimated parsimoniously. Additionally, the act of covariance clustering provides a model-based dimension reduction technique. This occurs when, through the inclusion of a suitable number of covariance clusters in the LMM, the variation originally arising between variables is fully accounted for by the variation arising between covariance clusters. In this way, variation between a large number of variables can potentially be reduced to a smaller set of covariance cluster effects.

Both outcomes of the covariance clustering method, being the estimation of parsimonious covariance between variables and model-based dimension reduction, have been demonstrated through application of the method to a multi-phase MS-based proteomics experiment. Typical of proteomics experiments, it was characterised by the measurement of a greater number of variables than experimental units; 617 proteins (variables) measured on a total of 48 subsamples (experimental units). Covariance clustering was implemented both at the residual level, to capture residual covariance between proteins, and at the treatment level, to cluster proteins displaying similar nonlinear responses to time in the malting process, in doing so demonstrating that the response of 617 proteins can be effectively reduced to 48 differential nonlinear forms.

We also present what we believe to be the first report of a multi-phase design for the conduct of an agriculturally motivated multi-phase MS-based proteomics experiment. Through the use of a multi-phase design, non-negligible variation was found to arise in both the malt sample collection phase and MS processing phase of the motivating experiment (Table 2). The two most dominant sources of variation were found to arise in the MS processing phase, and corresponded to variation between the processing of individual subsamples and batches of subsamples in the laboratory, respectively (Table 2). The potential for extraneous variation associated with processing order or batches in laboratory studies has been reported more generally (Cullis et al. 2003; Smith et al. 2006; Oakey et al. 2013) and also in the particular case of MS-based proteomics studies (Hu et al. 2005). In the case of the motivating experiment, the use of a multi-phase design prevented the introduction of potential bias in the experimental results, that would arise if samples had been processed in a systematic order. The findings of this experiment confirm the importance of sound experimental design in MS-based proteomics studies and demonstrate that multi-phase design is a robust design solution that is well suited for implementation in such studies into the future.

In the methods presented, a variance-component-model-like solution is used to induce covariance between variables. The traditional variance component model (Patterson et al. 1977) builds a simplistic covariance structure, which has been found to be inferior when compared to more complex covariance models in certain applications (Kelly et al. 2007). However, the reliable estimation of the variance–covariance parameters involved in these

more complex models is questionable when the number of variables to be modelled exceeds the number of observations informing the parameter estimates. In these situations, the resulting covariance structures are often of reduced rank and thus computationally challenging to estimate. Although the factor analytic variance structure formulated by [Thompson et al. \(2003\)](#), based on that originally proposed by [Smith et al. \(2001b\)](#), can accommodate the estimation of reduced rank variance structures between variables for moderate to large numbers of observations (≥ 50) ([Kelly et al. 2007](#)), evidence suggests that the performance of such structures, especially in terms of estimating covariance parameters, is reduced when the number of observations informing a parameter estimate is small (10–15) ([Macdonald 2018](#); [Macdonald et al. 2019](#)). In an attempt to avoid the estimation issues associated with these more complex covariance models and have greater flexibility and heterogeneity in the resulting covariance structure than that possible using a traditional variance component model, a ‘middle-ground’ in complexity and parsimony was reached by implementing the covariance clustering approach.

The definition of covariance clusters in the methods was achieved by grouping estimated effects using a k -means clustering algorithm ([Hartigan and Wong 1979](#)). The k -means algorithm was favoured as it resulted in the formation of clusters of more equal size, for smaller numbers of clusters (data not shown). This contrasts with hierarchical clustering methods, which, for small numbers of clusters, often result in one large cluster and multiple clusters consisting of one or two variables ([Nazarathy and Klok 2021](#)). When implemented as a covariance structure, the k -means approach results in structures with greater heterogeneity of variance and covariance between a greater number of variables, whereas the hierarchical approach induces homogeneous variance and covariance between a large number of variables, and heterogeneity between a few (data not shown). In the context of the motivating example, it was found that the covariance structure resulting from the k -means clustering approach was better suited for modelling the complex relationships that exist between proteins in the proteome ([Agrawal et al. 2013](#); [Robotti et al. 2015](#)) (data not shown).

In addition to providing a means of estimating a parsimonious covariance structure, the covariance clustering method can serve as a dimension reduction technique. This is achieved through a partitioning of variance between the covariance cluster and variable terms included in the LMM. Due to the nested nature of the terms and, with the definition and inclusion of sufficient covariance clusters in the model, it would not be unexpected for a situation to occur where the variance arising between variables (estimated to be nonzero in the baseline model) is fully accounted for by variation between covariance clusters (in the covariance cluster LMM). When applied to the motivating experiment, this was evidenced through the inclusion of an optimal number of spline covariance clusters in the corresponding LMM, whereby the variation that was previously attributable to the `Protein:spl (malttime)` term in the baseline response curve LMM (Table 3), was subsequently estimated at the boundary of the parameter space once the covariance clusters were introduced (Table 4). This indicated that the smooth, nonlinear variation between the 617 proteins was effectively accounted for, and could be reduced to, 48 nonlinear response profiles, estimated through the spline covariance cluster effects (Fig. 4). The clustering and identification of ‘typical’ response profiles could have been achieved using a range of techniques, such as a post hoc clustering of spline coefficients, general response curve clustering methods ([Gladish et al. 2021](#)), or

methods established under the [Ramsay and Silverman \(1997\)](#) paradigm of functional data analysis ([James and Sugar 2003](#); [Coffey and Hinde 2011](#); [Coffey et al. 2014](#)). However, the choice of the ‘optimal’ number of clusters to be formed using these approaches can rely on subjectively comparing multiple selection criteria ([Gladish et al. 2021](#)), or subjectively selecting parameters, in an attempt to avoid computational instabilities ([Coffey et al. 2014](#)). Using the LMM representation of the cubic smoothing spline for computationally stable variance parameter estimates, and through the partitioning of variance intrinsic in the covariance clustering method, a model-based and objective choice of the number of clusters, and thus dimension reduction, can be achieved.

The methods outlined in this paper were implemented using a two-stage LMM, in under approximately ten hours. Software limitations involving the commercial R package `asreml` ([Butler et al. 2017](#)) dictated that a single-stage analysis was infeasible due to prohibitive computation times, with the method estimated to take more than 16 days based on initial testing. Although the `asreml` package is unique in its support for fitting many complex variance structures in an LMM framework, the computation times can be prohibitive and impractical when implementing multiple of these structures to even moderately sized datasets, such as that arising from the motivating experiment. Computational limitations can be reduced by using a two-stage approach ([Piepho et al. 2012](#)), and, as such, this was favoured, despite the potential loss of information incurred by using this method as opposed to a single-stage analysis ([Gogel et al. 2018](#)). A weighted two-stage approach was used, as this has been shown to be superior to unweighted two-stage approaches, both in terms of loss of information and accuracy of effects ([Welham et al. 2010](#); [Piepho et al. 2012](#)).

Application of the covariance clustering method to the MS-based proteomics experiment revealed the importance of approximating, through covariance modelling, the complex relationships that can exist between proteins in the proteome ([Robotti et al. 2015](#)). Significant improvement in AIC, and thus model fit, was achieved by including any number of covariance clusters ($\nu_r \geq 2$) in the residual covariance cluster LMM, when compared to the baseline LMM (Fig. 3a). Modelling complex covariance between proteins is not currently performed as part of routine LMM analysis methods for proteomics data ([Oberge et al. 2008](#); [Choi et al. 2014](#)), nor was it modelled in a previous analysis of the data arising from the motivating experiment ([Osama et al. 2021](#)). Theoretically, ignoring such relationships will result in less accurate and potentially biased protein abundance predictions than if appropriate covariance between proteins was modelled ([De Faveri et al. 2017](#)).

Furthermore, through application of the covariance clustering method, new insights into how the barley grain and malt proteome changes over time in the malting process were discovered. The extent of variation between the 48 ‘typical’ response profiles, each with complexly varying forms (Fig. 4), suggests there is novel and exploitable variability within the proteome that could be further explored to better understand and potentially optimise the changes that occur during the malting process. Additionally, preliminary investigations suggest some alignment between the outcomes of the clustering procedure and the biological function of proteins. An example of this is the proteins HORVU1Hr1G059870.4 and HORVU1Hr1G059900.1, which both cluster together (the only two proteins belonging to spline covariance cluster 45; Figs. 4 and 5), demonstrate similar responses to time in the malting process and are identified as being ‘late embryogenesis proteins’ whose abundance

has been previously reported to decrease in the early stages of the malting process (Osama et al. 2021), consistent with their response in this study (Fig. 5). Further investigation of the proteins composing each spline covariance cluster, and the potential alignment of their biological functions in the barley grain and malt proteome, is the topic of ongoing research.

There are multiple possible extensions of the methods proposed in this paper. For instance, the complexity of the covariance structures resulting from covariance clustering could be increased. Currently, as the number of clusters increases, the resulting variance structure is characterised by greater pockets of heterogeneous variance and covariance; however, a greater number of variables are assumed to be related through simple covariance. Alternate implementations of the clustering approach which result in covariance structures with greater heterogeneity, while limiting the extent of simple covariance between variables, are an area open for future investigation. Additionally, further investigation is warranted into the persistence of cluster membership across different initial seeds of the k -means algorithm. Currently, the ‘optimal’ membership of variables to clusters is informed by only one seed \times number of cluster combination. As such, there is scope to explore the consistency of variables occurring together in the same cluster, across all seeds considered, for a given number of clusters. Such an investigation could influence the reporting of the covariance cluster effects (Fig. 4), or lead to alternate specifications of the ‘optimal’ covariance cluster models to respect these persistent variable groupings.

The covariance clustering method we have proposed provides an LMM-based solution for the analysis of designed experiments that yield datasets characterised by a greater number of variables than experimental units, and for which estimation of complex covariance structures between variables is desired. The act of covariance clustering models heterogeneity of variance–covariance between any random effects in an LMM framework, be it to capture relationships between treatments or experimental units. The method is applicable in situations where alternate complex covariance structures may be computationally difficult to estimate due to being of reduced rank. It is envisaged that covariance clustering could provide an alternative to the use of factor analytic variance structures, when the number of observations informing a variance–covariance parameter are small.

ACKNOWLEDGEMENTS

The first author would like to thank the Queensland Department of Agriculture and Fisheries for their financial support of his PhD project, along with the University of Queensland for their provision of a Research Training Program Tuition Fee Offset. The authors would also like to thank the Associate Editor and two reviewers for their suggestions, which considerably improved the quality of the paper. Open Access funding enabled and organized by CAUL and its Member Institutions.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

Funding This work is part of a PhD project being conducted by the first author, for which financial support has been provided by the Queensland Department of Agriculture and Fisheries, along with the University of Queensland via a Research Training Program Tuition Fee Offset.

Declarations

Conflict of interest The authors have no conflicts of interest to declare that are relevant to the content of this article.

[Received June 2023. Revised August 2023. Accepted September 2023. Published Online October 2023.]

REFERENCES

- Agrawal GK, Sarkar A, Righetti PG, Pedreschi R, Carpentier S, Wang T, Barkla BJ, Kohli A, Ndimba BK, Bykova NV, Rampitsch C, Zolla L, Rafudeen MS, Cramer R, Bindschedler LV, Tsakirpaloglou N, Ndimba RJ, Farrant JM, Renaut J, Job D, Kikuchi S, Rakwal R (2013) A decade of plant proteomics and mass spectrometry: translation of technical advancements to food security and safety issues. *Mass Spectrom Rev* 32:335–365
- Brien CJ, Bailey RA (2006) Multiple randomizations. *J R Stat Soc Ser B (Stat Methodol)* 68:571–609
- Brien CJ, Harch BD, Correll RL, Bailey RA (2011) Multiphase experiments with at least one later laboratory phase. I. Orthogonal designs. *J Agric Biol Environ Stat* 16:422–450
- Butler DG (2022) ODW: generate optimal experimental designs. (R Package Version 2.1.4)
- Butler DG, Cullis BR, Gilmour AR, Gogel BJ, Thompson R (2017) ASReml-R reference manual version 4. Report, VSN International Ltd
- Chen C, Hou J, Tanner JJ, Cheng J (2020) Bioinformatics methods for mass spectrometry-based proteomics data analysis. *Int J Mol Sci* 21:2873
- Choi M, Chang C-Y, Clough T, Broudy D, Killeen T, MacLean B, Vitek O (2014) MSstats: an R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments. *Bioinformatics* 30:2524–2526
- Coffey N, Hinde J (2011) Analyzing time-course microarray data using functional data analysis—a review. *Stat Appl Genet Mol Biol*. 10:1–32
- Coffey N, Hinde J, Holian E (2014) Clustering longitudinal profiles using P-splines and mixed effects models applied to time-course gene expression data. *Comput Stat Data Anal* 71:14–29
- Cullis BR, Smith AB, Panozzo JF, Lim P (2003) Barley malting quality: are we selecting the best? *Aust J Agric Res* 54:1261–1275
- De Faveri J, Verbyla AP, Pitchford WS, Venkatanagappa S, Cullis BR (2015) Statistical methods for analysis of multi-harvest data from perennial pasture variety selection trials. *Crop Pasture Sci* 66:947–962
- De Faveri J, Verbyla AP, Cullis BR, Pitchford WS, Thompson R (2017) Residual variance-covariance modelling in analysis of multivariate data from variety selection trials. *J Agric Biol Environ Stat* 22:1–22
- De Faveri J, Verbyla AP, Rebetzke G (2022) Random regression models for multi-environment, multi-time data from crop breeding selection trials. *Crop Pasture Sci* 74:271–283
- Dreccer MF, Condon AG, Macdonald B, Rebetzke GJ, Awasi M-A, Borgognone MG, Peake A, Piñera-Chavez FJ, Hundt A, Jackway P, McIntyre CL (2020) Genotypic variation for lodging tolerance in spring wheat: wider and deeper root plates, a feature of low lodging, high yielding germplasm. *Field Crop Res* 258:107942
- Fischer RA, Connor DJ (2018) Issues for cropping and agricultural science in the next 20 years. *Field Crop Res* 222:121–142
- Gilmour AR, Thompson R, Cullis BR (1995) Average information REML: an efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics* 51:1440–1450
- Gilmour AR, Gogel BJ, Cullis BR, Welham SJ, Thompson R (2015) ASReml User Guide Release 4.1 Functional Specification, Report

- Gladish DW, He D, Wang E (2021) Pattern analysis of Australia soil profiles for plant available water capacity. *Geoderma* 391:114977
- Gogel B, Smith A, Cullis B (2018) Comparison of a one- and two-stage mixed model analysis of Australia's National Variety Trial Southern Region wheat data. *Euphytica* 214:44
- Gross J (2011) *Mass spectrometry: a textbook*, 2nd edn. Springer, Berlin
- Hartigan JA, Wong MA (1979) Algorithm AS 136: a k-means clustering algorithm. *J R Stat Soc Ser C (Appl Stat)* 28:100–108
- Hu J, Coombes KR, Morris JS, Baggerly KA (2005) The importance of experimental design in proteomic mass spectrometry experiments: some cautionary tales. *Brief Funct Genomics* 3:322–331
- James GM, Sugar CA (2003) Clustering for sparsely sampled functional data. *J Am Stat Assoc* 98:397–408
- Kelly A, Forknall C (2020) *Advanced designs for barley breeding experiments*, book section 6. Burleigh Dodds Science Publishing Limited, Milton, pp 159–181
- Kelly AM, Smith AB, Eccleston JA, Cullis BR (2007) The accuracy of varietal selection using factor analytic models for multi-environment plant breeding trials. *Crop Sci* 47:1063–1070
- Kerr ED, Phung TK, Caboche CH, Fox GP, Platz GJ, Schulz BL (2019) The intrinsic and regulated proteomes of barley seeds in response to fungal infection. *Anal Biochem* 580:30–35
- Macdonald B (2018) *How low can you go? Performance of factor analytic models in the analysis of multi-environment trials with small numbers of varieties*, Honours thesis
- Macdonald B, King R, Kelly A (2019) Performance of factor analytic models in the analysis of multi-environment trials with small numbers of varieties. In: *Biometrics by the Botanic Gardens, International Biometric Society Australasian Region Conference*. <https://universityofadelaide.app.box.com/s/ugaby9mg3522m8q7x70y2c2mxchd66jf>
- McIntyre GA (1955) Design and analysis of two phase experiments. *Biometrics* 11:324–334
- Nazarathy Y, Klok H (2021) *Statistics with Julia: Fundamentals for data science, machine learning and artificial intelligence*. Springer, Berlin
- Oakey H, Verbyla A, Pitchford W, Cullis B, Kuchel H (2006) Joint modeling of additive and non-additive genetic line effects in single field trials. *Theor Appl Genet* 113:809–819
- Oakey H, Shafiei R, Comadran J, Uzrek N, Cullis B, Gomez LD, Whitehead C, McQueen-Mason SJ, Waugh R, Halpin C (2013) Identification of crop cultivars with consistently high lignocellulosic sugar release requires the use of appropriate statistical design and modelling. *Biotechnol Biofuels* 6:185
- Oberg AL, Vitek O (2009) Statistical design of quantitative mass spectrometry-based proteomic experiments. *J Proteome Res* 8:2144–2156
- Oberg AL, Mahoney DW, Eckel-Passow JE, Malone CJ, Wolfinger RD, Hill EG, Cooper LT, Onuma OK, Spiro C, Therneau TM, Bergen IIIHR (2008) Statistical analysis of relative labeled mass spectrometry data from complex samples using ANOVA. *J Proteome Res* 7:225–233
- Osama SK, Kerr ED, Yousif AM, Phung TK, Kelly AM, Fox GP, Schulz BL (2021) Proteomics reveals commitment to germination in barley seeds is marked by loss of stress response proteins and mobilisation of nutrient reservoirs. *J Proteomics* 242:104221
- Panozzo JF, Eckermann PJ, Mather DE, Moody DB, Black CK, Collins HM, Barr AR, Lim P, Cullis BR (2007) QTL analysis of malting quality traits in two barley populations. *Aust J Agric Res* 58:858–866
- Patterson HD, Thompson R (1971) Recovery of inter-block information when block sizes are unequal. *Biometrika* 58:545–554
- Patterson HD, Silvey V, Talbot M, Weatherup STC (1977) Variability of yields of cereal varieties in U.K. trials. *J Agric Sci* 89:239–245
- Perez-Riverol Y, Bai J, Bandla C, García-Seisdedos D, Hewapathirana S, Kamatchinathan S, Kundu D, Prakash A, Frericks-Zipper A, Eisenacher M, Walzer M, Wang S, Brazma A, Vizcaíno J (2021) The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. *Nucleic Acids Res* 50:D543–D552

- Piepho H-P, Möhring J, Schulz-Streeck T, Ogutu JO (2012) A stage-wise approach for the analysis of multi-environment trials. *Biom J* 54:844–860
- R Core Team (2019) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.Rproject.org/>
- Ramsay J, Silverman BW (1997) *Functional data analysis*, 1st edn. Springer, New York
- Robotti E, Manfredi M, Marengo E (2015) Biomarkers discovery through multivariate statistical methods: a review of recently developed methods and applications in proteomics. *J Proteom Bioinform* 1–1
- Rogers S, Taylor J (2019), A comparison of linear mixed model packages in R for analysis of plant breeding experiments. In: *Biometrics by the Botanic Gardens, International Biometric Society Australasian Region Conference*. https://ausbiometric2019.org/posters/Sam_Rogers_IBS_poster.pdf
- Schwarz P, Li Y (2010) *Malting and brewing uses of barley*. Blackwell Publishing Ltd, New York, pp 478–521
- Smith A, Cullis B, Gilmour A (2001a) The analysis of crop variety evaluation data in Australia. *Aust N Z J Stat* 43:129–145
- Smith A, Cullis B, Thompson R (2001b) Analyzing variety by environment data using multiplicative mixed models and adjustments for spatial field trend. *Biometrics* 57:1138–1147
- Smith AB, Lim P, Cullis BR (2006) The design and analysis of multi-phase plant breeding experiments. *J Agric Sci* 144:393–409
- Thompson R, Cullis B, Smith A, Gilmour A (2003) A sparse implementation of the Average Information algorithm for factor analytic and reduced rank variance models. *Aust N Z J Stat* 45:445–459
- Verbyla AP (2019) A note on model selection using information criteria for general linear models estimated using REML. *Aust N Z J Stat* 61:39–50
- Verbyla AP, Cullis BR, Kenward MG, Welham SJ (1999) The analysis of designed experiments and longitudinal data by using smoothing splines (with discussion). *J R Stat Soc Ser C (Appl Stat)* 48:269–311
- Verbyla AP, De Faveri J, Deery DM, Rebetzke GJ (2021) Modelling temporal genetic and spatio-temporal residual effects for high-throughput phenotyping data. *Aust N Z J Stat* 63:284–308
- Verbyla AP, De Faveri J, Wilkie JD, Lewis T (2018) Tensor cubic smoothing splines in designed experiments requiring residual modelling. *J Agric Biol Environ Stat* 23:478–508
- Welham SJ, Gogel BJ, Smith AB, Thompson R, Cullis BR (2010) A comparison of analysis methods for late-stage variety evaluation trials. *Aust N Z J Stat* 52:125–149
- Welham SJ, Gezan SA, Clark SJ, Mead A (2014) *Statistical methods in biology: design and analysis of experiments and regression*. CRC Press LLC, Philadelphia
- Yousif AM, Evans DE (2020) Changes in malt quality during production in two commercial malt houses. *J Inst Brew* 126:233–252
- Yu L-R, Stewart NA, Veenstra TD (2010) Chapter 8—Proteomics: the deciphering of the functional genome. Academic Press, San Diego, pp 89–96
- Zhang G, Annan RS, Carr SA, Neubert TA (2010) Overview of peptide and protein analysis by mass spectrometry. *Curr Protocols Protein Sci*. 62:16.1.1–16.1.30

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.