



Privacy-preserving and lossless distributed estimation of high-dimensional generalized additive mixed models

Schalk Daniel^{1,2} · Bischl Bernd^{1,2} · Rügamer David^{1,2}

Received: 7 March 2023 / Accepted: 2 October 2023 / Published online: 7 November 2023
© The Author(s) 2023

Abstract

Various privacy-preserving frameworks that respect the individual's privacy in the analysis of data have been developed in recent years. However, available model classes such as simple statistics or generalized linear models lack the flexibility required for a good approximation of the underlying data-generating process in practice. In this paper, we propose an algorithm for a distributed, privacy-preserving, and lossless estimation of generalized additive mixed models (GAMM) using component-wise gradient boosting (CWB). Making use of CWB allows us to reframe the GAMM estimation as a distributed fitting of base learners using the L_2 -loss. In order to account for the heterogeneity of different data location sites, we propose a distributed version of a row-wise tensor product that allows the computation of site-specific (smooth) effects. Our adaption of CWB preserves all the important properties of the original algorithm, such as an unbiased feature selection and the feasibility to fit models in high-dimensional feature spaces, and yields equivalent model estimates as CWB on pooled data. Next to a derivation of the equivalence of both algorithms, we also showcase the efficacy of our algorithm on a distributed heart disease data set and compare it with state-of-the-art methods.

Keywords Distributed computing · Functional gradient descent · Generalized linear mixed model · Machine learning · Privacy-preserving modelling

1 Introduction

More than ever, data is collected to record the ubiquitous information in our everyday life. However, on many occasions, the physical location of data points is not confined to one place (one global site) but distributed over different locations (sites). This is the case for, e.g., patient records that are gathered at different hospitals but usually not shared between hospitals or other facilities due to the sensitive information they contain. This makes data analysis challenging, particularly if methods require or notably benefit from incorporating all available (but distributed) information. For example, per-

sonal patient information is typically distributed over several hospitals, while sharing or merging different data sets in a central location is prohibited. To overcome this limitation, different approaches have been developed to directly operate at different sites and unite information without having to share sensitive parts of the data to allow privacy-preserving data analysis.

Distributed data Distributed data can be partitioned vertically or horizontally across different sites. Horizontally partitioned data means that observations are spread across different sites with access to all existing features of the available data point, while for vertically partitioned data, different sites have access to all observations but different features (covariates) for each of these observations. In this work, we focus on horizontally partitioned data. Existing approaches for horizontally partitioned data vary from fitting regression models such as generalized linear models (GLMs; Wu et al. 2012; Lu et al. 2015; Jones et al. 2013; Chen et al. 2018), to conducting distributed evaluations (Boyd et al. 2015; Ünal et al. 2021; Schalk et al. 2022), to fitting artificial neural networks (McMahan et al. 2017). Furthermore, various software frameworks are available to run a comprehensive analysis of

✉ Schalk Daniel
daniel.schalk@stat.uni-muenchen.de

Bischl Bernd
bernd.bischl@stat.uni-muenchen.de

Rügamer David
david.ruegamer@stat.uni-muenchen.de

¹ Department of Statistics, LMU Munich, Munich, Germany

² Munich Center for Machine Learning (MCML), Munich, Germany

distributed data. One example is the collection of R (R Core Team 2021) packages `DataSHIELD` (Gaye et al. 2014), which enables data management and descriptive data analysis as well as securely fitting of simple statistical models in a distributed setup without leaking information from one site to the others.

Interpretability and data heterogeneity In many research areas that involve critical decision-making, especially in medicine, methods should not only excel in predictive performance but also be interpretable. Models should provide information about the decision-making process, the feature effects, and the feature importance as well as intrinsically select important features. Generalized additive models (GAMs; see, e.g., Wood 2017) are one of the most flexible approaches in this respect, providing interpretable yet complex models that also allow for non-linearity in the data.

As longitudinal studies are often the most practical way to gather information in many research fields, methods should also be able to account for subject-specific effects and account for the correlation of repeated measurements. Furthermore, when analyzing data originating from different sites, the assumption of having identically distributed observations across all sites often does not hold. In this case, a reasonable assumption for the data-generating process is a site-specific deviation from the general population mean. Adjusting models to this situation is called interoperability (Litwin et al. 1990), while ignoring it may lead to biased or wrong predictions.

1.1 Related literature

Various approaches for distributed and privacy-preserving analysis have been proposed in recent years. In the context of statistical models, Karr et al. (2005) describe how to calculate a linear model (LM) in a distributed and privacy-preserving fashion by sharing data summaries. Jones et al. (2013) propose a similar approach for GLMs by communicating the Fisher information and score vector to conduct a distributed Fisher scoring algorithm. The site information is then globally aggregated to estimate the model parameters. Other privacy-preserving techniques include ridge regression (Chen et al. 2018), logistic regression, and neural networks (Mohassel and Zhang 2017).

In machine learning, methods such as the naive Bayes classifier, trees, support vector machines, and random forests (Li et al. 2020a) exist with specific encryption techniques (e.g., the Paillier cryptosystem; Paillier 1999) to conduct model updates. In these setups, a trusted third party is usually required. However, this is often unrealistic and difficult to implement, especially in a medical or clinical setup. Furthermore, as encryption is an expensive operation, its application is infeasible for complex algorithms that require many encryption calls (Naehrig et al. 2011). Existing privacy-

preserving boosting techniques often focus on the AdaBoost algorithm by using aggregation techniques of the base classifier (Lazarevic and Obradovic 2001; Gambis et al. 2007). A different approach to boosting decision trees in a federated learning setup was introduced by Li et al. (2020b) using a locality-sensitive hashing to obtain similarities between data sets without sharing private information. These algorithms focus on aggregating tree-based base components, making them difficult to interpret, and come with no inferential guarantees.

In order to account for repeated measurements, Luo et al. (2022) propose a privacy-preserving and lossless way to fit linear mixed models (LMMs) to correct for heterogeneous site-specific random effects. Their concept of only sharing aggregated values is similar to our approach, but is limited in the complexity of the model and only allows normally distributed outcomes. Other methods to estimate LMMs in a secure and distributed fashion are Zhu et al. (2020), Anjum et al. (2022), or Yan et al. (2022).

Besides privacy-preserving and distributed approaches, integrative analysis is another technique based on pooling the data sets into one and analyzing this pooled data set while considering challenges such as heterogeneity or the curse of dimensionality (Curran and Hussong 2009; Bazeley 2012; Mirza et al. 2019). While advanced from a technical perspective by, e.g., outsourcing computational demanding tasks such as the analysis of multi-omics data to cloud services (Augustyn et al. 2021), the existing statistical cloud-based methods only deal with basic statistics. The challenges of integrative analysis are similar to the ones tackled in this work, our approach, however, does not allow merging the data sets in order to preserve privacy.

1.2 Our contribution

This work presents a method to fit generalized additive mixed models (GAMMs) in a privacy-preserving and lossless manner¹ to horizontally distributed data. This not only allows the incorporation of site-specific random effects and accounts for repeated measurements in LMMs, but also facilitates the estimation of mixed models with responses following any distribution from the exponential family and provides the possibility to estimate complex non-linear relationships between covariates and the response. To the best of our knowledge, we are the first to provide an algorithm to fit the class of GAMMs in a privacy-preserving and lossless fashion on distributed data.

Our approach is based on component-wise gradient boosting (CWB; Bühlmann and Yu 2003). CWB can be used

¹ In this article, we define a distributed fitting procedure as lossless if the model parameters of the algorithm are the same as the ones computed on the pooled data.

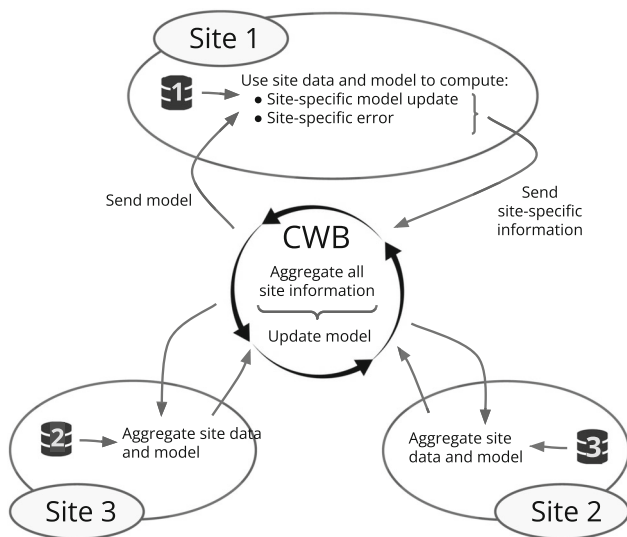


Fig. 1 Method overview of the proposed distributed CWB approach with one main CWB model maintained by a host (center) and distributed computations on different sites (three as an example in this case) that incorporate and provide site-specific information while preserving privacy

to estimate additive models, account for repeated measurements, compute feature importance, and conduct feature selection. Furthermore, CWB is suited for high-dimensional data situations ($n \ll p$). CWB is therefore often used in practice for, e.g., predicting the development of oral cancer (Saintigny et al 2011), classifying individuals with and without patellofemoral pain syndrome (Liew et al. 2020), or detecting synchronization in bioelectrical signals (Rügamer et al. 2018). However, there have so far not been any attempts to allow for a distributed, privacy-preserving, and lossless computation of the CWB algorithm. In this paper, we propose a distributed version of CWB that yields the identical model produced by the original algorithm on pooled data and that accounts for site heterogeneity by including interactions between features and a site variable. This is achieved by adjusting the fitting process using (1) a distributed estimation procedure, (2) a distributed version of row-wise tensor product base learners, and (3) an adaption of the algorithm to conduct feature selection in the distributed setup. Figure 1 sketches the proposed distributed estimation procedure.

We implement our method in R using the DataSHIELD framework and demonstrate its application in an exemplary medical data analysis. Our distributed version of the original CWB algorithm does not have any additional hyperparameters (HPs) and uses optimization strategies from previous research results to define meaningful values for all HPs, effectively yielding a tuning-free method.

The remainder of this paper is structured as follows: First, we introduce the basic notation, terminology, and setup of GAMMs in Sect. 2. We then describe the original CWB algo-

rithm in Sect. 2.3 and its link to GAMMs. In Sect. 3, we present the distributed setup and our novel extension of the CWB algorithm. Finally, Sect. 4 demonstrates both how our distributed CWB algorithm can be used in practice and how to interpret the obtained results.

Implementation We implement our approach as an R package using the DataSHIELD framework and make it available on GitHub.² The code for the analysis can also be found in the repository.³

2 Background

2.1 Notation and terminology

Our proposed approach uses the CWB algorithm as fitting engine. Since this method was initially developed in machine learning, we introduce here both the statistical notation used for GAMMs as well as the respective machine learning terminology and explain how to relate the two concepts.

We assume a p -dimensional covariate or feature space $\mathcal{X} = (\mathcal{X}_1 \times \dots \times \mathcal{X}_p) \subseteq \mathbb{R}^p$ and response or outcome values from a target space \mathcal{Y} . The goal of boosting is to find the unknown relationship f between \mathcal{X} and \mathcal{Y} . In turn, GAMMs (as presented in Sect. 2.2) model the conditional distribution of an outcome variable Y with realizations $y \in \mathcal{Y}$, given features $\mathbf{x} = (x_1, \dots, x_p) \in \mathcal{X}$. Given a data set $\mathcal{D} = \{(\mathbf{x}^{(1)}, y^{(1)}), \dots, (\mathbf{x}^{(n)}, y^{(n)})\}$ with n observations drawn (conditionally) independently from an unknown probability distribution $\mathbb{P}_{\mathbf{x}, y}$ on the joint space $\mathcal{X} \times \mathcal{Y}$, we aim to estimate this functional relationship in CWB with \hat{f} . The goodness-of-fit of a given model \hat{f} is assessed by calculating the empirical risk $\mathcal{R}_{\text{emp}}(\hat{f}) = n^{-1} \sum_{(\mathbf{x}, y) \in \mathcal{D}} L(y, \hat{f}(\mathbf{x}))$ based on a loss function $L : \mathcal{Y} \times \mathbb{R} \rightarrow \mathbb{R}$ and the data set \mathcal{D} . Minimizing \mathcal{R}_{emp} using this loss function is equivalent to estimating f using maximum likelihood by defining $L(y, f(\mathbf{x})) = -\ell(y, h(f(\mathbf{x})))$ with log-likelihood ℓ , response function h and minimizing the sum of log-likelihood contributions.

In the following, we also require the vector $\mathbf{x}_j = (x_j^{(1)}, \dots, x_j^{(n)})^\top \in \mathcal{X}_j$, which refers to the j th feature. Furthermore, let $\mathbf{x} = (x_1, \dots, x_p)$ and y denote arbitrary members of \mathcal{X} and \mathcal{Y} , respectively. A special role is further given to a subset $\mathbf{u} = (u_1, \dots, u_q)^\top$, $q \leq p$, of features \mathbf{x} , which will be used to model the heterogeneity in the data.

² github.com/schalkdaniel/dsCWB.

³ github.com/schalkdaniel/dsCWB/blob/main/usecase/analyse.R.

2.2 Generalized additive mixed models

A very flexible class of regression models to model the relationship between covariates and the response are GAMMs (see, e.g., Wood 2017). In GAMMs, the response $Y^{(i)}$ for observation $i = 1, \dots, n_s$ of measurement unit (or site) s is assumed to follow some exponential family distribution such as the Poisson, binomial, or normal distributions (see, e.g., McCullagh and Nelder 1989), conditional on features $\mathbf{x}^{(i)}$ and the realization of some random effects. The expectation $\mu := \mathbb{E}(Y^{(i)}|\mathbf{x}^{(i)}, \mathbf{u}^{(i)})$ of the response $Y^{(i)}$ for observations $i = 1, \dots, n_s$ of measurement unit (or site) s in GAMMs is given by

$$h^{-1}(\mu^{(i)}) = f^{(i)} = \sum_{j \in \mathcal{J}_1} x_j^{(i)} \beta_j + \sum_{j \in \mathcal{J}_2} u_j^{(i)} \gamma_{j,s} + \sum_{j \in \mathcal{J}_3} \phi_j(x_j^{(i)}). \tag{1}$$

In (1), h is a smooth monotonic response function, f corresponds to the additive predictor, $\gamma_{j,s} \sim \mathcal{N}(0, \psi)$ are random effects accounting for heterogeneity in the data, and ϕ_j are non-linear effects of pre-specified covariates. The different index sets $\mathcal{J}_1, \mathcal{J}_2, \mathcal{J}_3 \subseteq \{1, \dots, p\} \cup \emptyset$ indicate which features are modeled as fixed effects, random effects, or non-linear (smooth) effects, respectively. The modeler usually defines these sets. However, as we will also explain later, the use of CWB as a fitting engine allows for automatic feature selection and therefore does not require explicitly defining these sets. In GAMMs, smooth effects are usually represented by (spline) basis functions, i.e., $\phi_j(x_j) \approx (B_{j,1}(x_j), \dots, B_{j,d_j}(x_j))^T \boldsymbol{\theta}_j$, where $\boldsymbol{\theta}_j \in \mathbb{R}^{d_j}$ are the basis coefficients corresponding to each basis function B_{j,d_j} . The coefficients are typically constrained in their flexibility by adding a quadratic (difference) penalty for (neighboring) coefficients to the objective function to enforce smoothness. GAMMs, as in (1), are not limited to univariate smooth effects ϕ_j , but allow for higher-dimensional non-linear effects $\phi(x_{j_1}, x_{j_2}, \dots, x_{j_k})$. The most common higher-dimensional smooth interaction effects are bivariate effects ($k = 2$) and can be represented using a bivariate or a tensor product spline basis (see Sect. 2.3.1 for more details). Although higher-order splines with $k > 2$ are possible, models are often restricted to bivariate interactions for the sake of interpretability and computational feasibility. In Sect. 3, we will further introduce varying coefficient terms $\phi_{j,s}(x_j)$ in the model (1), i.e., smooth effects f varying with a second variable s . Analogous to random slopes, s can also be the index set defining observation units of random effects \mathcal{J}_2 . Using an appropriate distribution assumption for the basis coefficients $\boldsymbol{\theta}_j$, these varying coefficients can then be considered as random smooth effects.

2.3 Component-wise boosting

Component-wise (gradient) boosting (CWB; Bühlmann and Yu 2003; Bühlmann et al. 2007) is an iterative algorithm that performs block-coordinate descent steps with blocks (or base learners) corresponding to the additive terms in (1). With a suitable choice of base learners and objective function, CWB allows efficient optimization of GAMMs, even in high-dimensional settings with $p \gg n$. We will first introduce the concept of base learners that embed additive terms of the GAMM into boosting and subsequently describe the actual fitting routine of CWB. Lastly, we will describe the properties of the algorithm and explain its connection to model (1).

2.3.1 Base learners

In CWB, the l th base learner $b_l : \mathcal{X} \rightarrow \mathbb{R}$ is used to model the contribution of one or multiple features in the model. In this work, we investigate parametrized base learners $b_l(\mathbf{x}, \boldsymbol{\theta}_l)$ with parameters $\boldsymbol{\theta}_l \in \mathbb{R}^{d_l}$. For simplicity, we will use $\boldsymbol{\theta}$ as a wildcard for the coefficients of either fixed effects, random effects, or spline bases in the following. We assume that each base learner can be represented by a generic basis representation $g_l : \mathcal{X} \rightarrow \mathbb{R}^{d_l}$, $\mathbf{x} \mapsto g_l(\mathbf{x}) = (g_{l,1}(\mathbf{x}), \dots, g_{l,d_l}(\mathbf{x}))^T$ and is linear in the parameters, i.e., $b_l(\mathbf{x}, \boldsymbol{\theta}_l) = g_l(\mathbf{x})^T \boldsymbol{\theta}_l$. Note that the basis transformation g_l of the l th base learner does not necessarily select the j th feature x_j . This is required to, e.g., let two base learners l and k depend on the same feature x_j . For n observations, we define the design matrix of a base learner b_l as $\mathbf{Z}_l := (g_l(\mathbf{x}^{(1)}), \dots, g_l(\mathbf{x}^{(n)}))^T \in \mathbb{R}^{n \times d_l}$. Note that base learners are typically not defined on the whole feature space but on a subset $\mathcal{X}_l \subseteq \mathcal{X}$. For example, a common choice for CWB is to define one base learner for every feature $x_l \in \mathcal{X}_l$ to model the univariate contributions of that feature.

A base learner $b_l(\mathbf{x}, \boldsymbol{\theta}_l)$ can depend on HPs $\boldsymbol{\alpha}_l$ that are set prior to the fitting process. For example, choosing a base learner using a P-spline (Eilers and Marx 1996) representation requires setting the degree of the basis functions, the order of the difference penalty term, and a parameter λ_l determining the smoothness of the spline. Regularized base learners, in addition, will have pre-defined penalty matrices \mathbf{K}_l . For convenience, we further denote the penalty matrix already augmented with the corresponding smoothing parameter with \mathbf{P}_l , e.g., $\mathbf{P}_l = \lambda_l \mathbf{K}_l$. In order to represent GAMMs in CWB, the following four base learner types are used.

(Regularized) linear base learners A linear base learner is used to include linear effects of a features $x_{j_1}, \dots, x_{j_{d_l}}$ into the model. The basis transformation is given by $g_l(\mathbf{x}) =$

$(g_{l,1}(\mathbf{x}), \dots, g_{l,d_l+1}(\mathbf{x}))^\top = (1, x_{j_1}, \dots, x_{j_{d_l}})^\top$. Linear base learners can be regularized by incorporating a ridge penalization (Hoerl and Kennard 1970) with tunable penalty parameter λ_l as an HP α_l . Fitting a ridge penalized linear base learner to a response vector $\mathbf{y} \in \mathbb{R}^n$ results in the penalized least squares estimator $\hat{\theta}_l = (\mathbf{Z}_l^\top \mathbf{Z}_l + \mathbf{P}_l)^{-1} \mathbf{Z}_l^\top \mathbf{y}$ with penalty matrix $\mathbf{P}_l = \lambda_l \mathbf{K}_l$, $\mathbf{K}_l = \mathbf{I}_{d_l+1}$, where \mathbf{I}_d denotes the d -dimensional identity matrix. Often, an unregularized linear base learner is also included to model the contribution of one feature x_j as a linear base learner without penalization. The basis transformation is then given by $g_l(\mathbf{x}) = (1, x_j)^\top$ and $\lambda_l = 0$.

Spline base learners These base learners model smooth effects using univariate splines. A common choice is penalized B-splines (P-Splines; Eilers and Marx 1996), where the feature x_j is transformed using a B-spline basis transformation $g_l(\mathbf{x}) = (B_{l,1}(x_j), \dots, B_{l,d_l}(x_j))^\top$ with d_l basis functions $g_{l,m} = B_{l,m}$, $m = 1, \dots, d_l$. In this case, the choice of the spline order B , the number of basis functions d_l , the penalization term λ_l , and the order v of the difference penalty (represented by a matrix $\mathbf{D}_l \in \mathbb{R}^{d_l-v \times d_l}$) are considered HPs α_l of the base learner. The base learner’s parameter estimator in general is given by the penalized least squares solution $\hat{\theta}_l = (\mathbf{Z}_l^\top \mathbf{Z}_l + \mathbf{P}_l)^{-1} \mathbf{Z}_l^\top \mathbf{y}$, with penalization matrix $\mathbf{P}_l = \lambda_l \mathbf{K}_l$ and $\mathbf{K}_l = \mathbf{D}_l^\top \mathbf{D}_l$ in the case of P-splines.

Categorical and random effect base learners Categorical features $x_j \in \{1, \dots, G\}$ with $G \in \mathbb{N}$, $G \geq 2$ classes are handled by a binary encoding $g_l(\mathbf{x}) = (\mathbb{1}_{\{1\}}(x_j), \dots, \mathbb{1}_{\{G\}}(x_j))^\top$ with the indicator function $\mathbb{1}_A(x) = 1$ if $x \in A$ and $\mathbb{1}_A(x) = 0$ if $x \notin A$. A possible alternative encoding is the dummy encoding with $\check{g}_l(\mathbf{x}) = (1, \mathbb{1}_{\{1\}}(x_j), \dots, \mathbb{1}_{\{G-1\}}(x_j))^\top$ with reference group G . Similar to linear and spline base learners, it is possible to incorporate a ridge penalization with HP $\alpha_l = \lambda_l$. This results in the base learner’s penalized least squared estimator $\hat{\theta}_l = (\mathbf{Z}_l^\top \mathbf{Z}_l + \mathbf{P}_l)^{-1} \mathbf{Z}_l^\top \mathbf{y}$ with penalization matrix $\mathbf{P}_l = \lambda_l \mathbf{K}_l$, $\mathbf{K}_l = \mathbf{I}_G$. Due to the mathematical equivalence of ridge penalized linear effects and random effects with normal prior (see, e.g., Brumback et al. 1999), this base learner can further be used to estimate random effect predictions $\hat{\gamma}_j$ when using categorical features u_j and thereby account for heterogeneity in the data. Hence, this base learner can also be used to model site-specific effects in a distributed system, as outlined later. While such random effects do not directly provide a variance estimate of the different measurement units and are primarily used to account for intra-class correlation, an approximation of the variance components can be retrieved post-model fitting by, e.g., computing the empirical variance of $(\hat{\gamma}_j)_{j \in \mathcal{J}_3}$.

Row-wise tensor product base learners This type of base learner is used to model a pairwise interaction between two features x_j and x_k . Given two base learners b_j and b_k with basis representations $g_j(\mathbf{x}) = (g_{j,1}(x_j), \dots, g_{j,d_j}(x_j))^\top$

and $g_k(\mathbf{x}) = (g_{k,1}(x_k), \dots, g_{k,d_k}(x_k))^\top$, the basis representation of the row-wise tensor product base learner $b_l = b_j \times b_k$ is defined as $g_l(\mathbf{x}) = (g_j(\mathbf{x})^\top \otimes g_k(\mathbf{x})^\top)^\top = (g_{j,1}(x_j)g_{k,1}(x_k), \dots, g_{j,d_j}(x_j)g_{k,d_k}(x_k))^\top \in \mathbb{R}^{d_l}$ with $d_l = d_j d_k$. The HPs $\alpha_l = \{\alpha_j, \alpha_k\}$ of a row-wise tensor product base learner are induced by the HPs α_j and α_k of the respective individual base learners. Analogously to other base learners, the penalized least squared estimator in this case is $\hat{\theta}_l = (\mathbf{Z}_l^\top \mathbf{Z}_l + \mathbf{P}_l)^{-1} \mathbf{Z}_l^\top \mathbf{y}$ with penalization matrix $\mathbf{P}_l = \tau_j \mathbf{K}_j \otimes \mathbf{I}_{d_k} + \mathbf{I}_{d_j} \otimes \tau_k \mathbf{K}_k \in \mathbb{R}^{d_l \times d_l}$. This Kronecker sum penalty, in particular, allows for anisotropic smoothing with penalties τ_j and τ_k when using two spline bases for g_j and g_k , and varying coefficients or random splines when combining a (penalized) categorical base learner and a spline base learner.

2.3.2 Fitting algorithm

CWB first initializes an estimate \hat{f} of the additive predictor with a loss-optimal constant value $\hat{f}^{[0]} = \arg \min_{c \in \mathbb{R}} \mathcal{R}_{\text{emp}}(c)$. It then proceeds and estimates Eq. (1) using an iterative steepest descent minimization in function space by fitting the previously defined base learners to the model’s functional gradient $\nabla_f L(y, f)$ evaluated at the current model estimate \hat{f} . Let $\hat{f}^{[m]}$ denote the model estimation after $m \in \mathbb{N}$ iterations. In each step in CWB, the pseudo residuals $\tilde{r}^{[m](i)} = -\nabla_f L(y^{(i)}, f(\mathbf{x}^{(i)}))|_{f=\hat{f}^{[m-1]}}$ for $i \in \{1, \dots, n\}$ are first computed. CWB then selects the best-fitting base learner from a pre-defined pool of base-learners denoted by $\mathcal{B} = \{b_l\}_{l \in \{1, \dots, |\mathcal{B}|\}}$ and adds the base learner’s contribution to the previous model $\hat{f}^{[m]}$. The selected base learner is chosen based on its sum of squared errors (SSE) when regressing the pseudo residuals $\tilde{\mathbf{r}}^{[m]} = (\tilde{r}^{[m](1)}, \dots, \tilde{r}^{[m](n)})^\top$ onto the base learner’s features using the L_2 -loss. Further details of CWB are given in Algorithm 1 (see, e.g., Schalk et al. 2023).

Controlling HPs of CWB

Good estimation performance can be achieved by selecting a sufficiently small learning rate, e.g., 0.01, as suggested in Bühlmann et al. (2007), and adaptively selecting the number of boosting iterations via early stopping on a validation set. To enforce a fair selection of model terms and thus unbiased effect estimation, regularization parameters are set such that all base learners have the same degrees-of-freedom (Hofner et al. 2011). As noted by Bühlmann et al. (2007), choosing smaller degrees-of-freedom induces more penalization (and thus, e.g., smoother estimated function for spline base learners), which yields a model with lower variance at the cost of a larger bias. This bias induces a shrinkage in the estimated coefficients towards zero but can be reduced by running the optimization process for additional iterations.

Algorithm 1 Vanilla CWB algorithm

Input Train data \mathcal{D} , learning rate ν , number of boosting iterations M , loss function L , set of base learner \mathcal{B}

Output Model $\hat{f}^{[M]}$ defined by fitted parameters $\hat{\theta}^{[1]}, \dots, \hat{\theta}^{[M]}$

```

1: procedure CWB( $\mathcal{D}$ ,  $\nu$ ,  $L$ ,  $\mathcal{B}$ )
2:   Initialize:  $\hat{f}^{[0]}(\mathbf{x}) = \arg \min_{c \in \mathbb{R}} \mathcal{R}_{\text{emp}}(c)$ 
3:   for  $m \in \{1, \dots, M\}$  do
4:      $\tilde{r}^{[m](i)} = -\nabla_f L(y^{(i)}, f(\mathbf{x}^{(i)}))|_{f=\hat{f}^{[m-1]}}$ ,  $\forall i \in \{1, \dots, n\}$ 
5:     for  $l \in \{1, \dots, |\mathcal{B}|\}$  do
6:        $\hat{\theta}_l^{[m]} = (\mathbf{Z}_l^T \mathbf{Z}_l + \mathbf{P}_l)^{-1} \mathbf{Z}_l^T \tilde{r}^{[m]}$ 
7:        $\text{SSE}_l = \sum_{i=1}^n (\tilde{r}^{[m](i)} - b_l(\mathbf{x}^{(i)}, \hat{\theta}_l^{[m]}))^2$ 
8:     end for
9:      $l^{[m]} = \arg \min_{l \in \{1, \dots, |\mathcal{B}|\}} \text{SSE}_l$ 
10:     $\hat{f}^{[m]}(\mathbf{x}) = \hat{f}^{[m-1]}(\mathbf{x}) + \nu b_{l^{[m]}}(\mathbf{x}, \hat{\theta}_{l^{[m]}}^{[m]})$ 
11:  end for
12:  return  $\hat{f} = \hat{f}^{[M]}$ 
13: end procedure

```

2.3.3 Properties and link to generalized additive mixed models

The estimated coefficients $\hat{\theta}$ resulting from running the CWB algorithm are known to converge to the maximum likelihood solution (see, e.g., Schmid and Hothorn 2008) for $M \rightarrow \infty$ under certain conditions. This is due to the fact that CWB performs a coordinate gradient descent update of a model defined by its additive base learners that exactly represent the structure of an additive mixed model (when defining the base learners according to Sect. 2.3.1) and by the objective function that corresponds to the negative (penalized) log-likelihood. Two important properties of this algorithm are (1) its coordinate-wise update routine, and (2) the nature of model updates using the L_2 -loss. Due to the first property, CWB can be used in settings with $p \gg n$, as only a single additive term is fitted onto the pseudo-residuals in every iteration. This not only reduces the computational complexity of the algorithm for an increasing number of additive predictors (linear instead of quadratic) but also allows variable selection when stopping the routine early (e.g., based on a validation data set), as not all the additive components might have been selected into the model. In particular, this allows users to specify the full GAMM model without manual specification of the type of feature effect (fixed or random, linear or non-linear) and then automatically sparsify this model by an objective and data-driven feature selection. The second property, allows fitting models of the class of *generalized* linear/additive (mixed) models using only the L_2 -loss instead of having to work with some iterative weighted least squares routine. In particular, this allows performing the proposed lossless distributed computations described in this paper, as we will discuss in Sect. 3.

2.4 Distributed computing setup and privacy protection

Before presenting our main results, we now introduce the distributed data setup we will work with throughout the remainder of this paper. The data set \mathcal{D} is horizontally partitioned into S data sets $\mathcal{D}_s = \left\{ \left(\mathbf{x}_s^{(1)}, y_s^{(1)} \right), \dots, \left(\mathbf{x}_s^{(n_s)}, y_s^{(n_s)} \right) \right\}$, $s = 1, \dots, S$ with n_s observations. Each data set \mathcal{D}_s is located at a different site s and potentially follows a different data distributions $\mathbb{P}_{x,y,s}$. The union of all data sets yields the whole data set $\mathcal{D} = \cup_{s=1}^S \mathcal{D}_s$ with mutually exclusive data sets $\mathcal{D}_s \cap \mathcal{D}_l = \emptyset \forall l, s \in \{1, \dots, S\}, l \neq s$. The vector of realizations per site is denoted by $\mathbf{y}_s \in \mathcal{Y}^{n_s}$.

In this distributed setup, multiple ways exist to communicate information without revealing individual information. More complex methods such as differential privacy (Dwork 2006), homomorphic encryption (e.g., the Paillier cryptosystem; Paillier 1999), or k-anonymity (Samarati and Sweeney 1998; Sweeney 2002) allow sharing information without violating an individual's privacy. An alternative option is to only communicate aggregated statistics. This is one of the most common approaches and is also used by DataSHIELD (Gaye et al. 2014) for GLMs or by Luo et al (2022) for LMMs. DataSHIELD, for example, uses a privacy level that indicates how many individual values must be aggregated to allow the communication of aggregated values. For example, setting the privacy level to a value of 5 enables sharing of summary statistics such as sums, means, variances, etc. if these are computed on at least 5 elements (observations).

Host and site setup Throughout this article, we assume the $1, \dots, S$ sites or servers to have access to their respective data set \mathcal{D}_s . Each server is allowed to communicate with a host server that is also the analyst's machine. In this setting, the analyst can potentially see intermediate data used when running the algorithms, and hence each message communicated from the servers to the host must not allow any reconstruction of the original data. The host server is responsible for aggregating intermediate results and communicating these results back to the servers.

3 Distributed component-wise boosting

We now present our distributed version of the CWB algorithm to fit privacy-preserving and lossless GAMMs. In the following, we first describe further specifications of our setup in Sect. 3.1, elaborate on the changes made to the set of base learners in Sect. 3.2, and then show how to adapt CWB's fitting routine in Sect. 3.3.

3.1 Setup

In the following, we distinguish between site-specific and shared effects. As effects estimated across sites typically correspond to fixed effects and effects modeled for each site separately are usually represented using random effects, we use the terms as synonyms in the following, i.e., *shared effects* and *fixed effects* are treated interchangeably and the same holds for *site-specific effects* and *random effects*. We note that this is only for ease of presentation and our approach also allows for site-specific fixed effects and random shared effects. As the data is not only located at different sites but also potentially follows different data distributions $\mathbb{P}_{x,y,s}$ at each site s , we extend Eq. (1) to not only include random effects per site, but also site-specific smooth (random) effects $\phi_{j,s}(x_j)$, $s = 1, \dots, S$ for all features x_j with $j \in \mathcal{J}_3$. For every of these smooth effects $\phi_{j,s}$ we assume an existing shared effect $f_{j,\text{shared}}$ that is equal for all sites. These assumptions—particularly the choice of site-specific effects—are made for demonstration purposes. In a real-world application, the model structure can be defined individually to match the given data situation. However, note again that CWB intrinsically performs variable selection, and there is thus no need to manually define the model structure in practice. In order to incorporate the site information into the model, we add a variable $x_0^{(i)} \in \{1, \dots, S\}$ for the site to the data by setting $\tilde{\mathbf{x}}^{(i)} = (x_0^{(i)}, \mathbf{x}^{(i)})$. The site variable is a categorical feature with S classes.

3.2 Base learners

For shared effects, we keep the original structure of CWB with base learners chosen from a set of possible learners \mathcal{B} . Section 3.3.1 explains how these shared effects are estimated in the distributed setup. We further define a random effect base learner b_0 with basis transformation $g_0(x_0) = (\mathbb{1}_{\{1\}}(x_0), \dots, \mathbb{1}_{\{S\}}(x_0))^T$ and design matrix $\mathbf{Z}_0 \in \mathbb{R}^{n \times S}$. We use b_0 to extend \mathcal{B} with a second set of base learners $\mathcal{B}_\times = \{b_0 \times b \mid b \in \mathcal{B}\}$ to model site-specific random effects. All base learners in \mathcal{B}_\times are row-wise tensor product base learners $b_{l_\times} = b_0 \times b_l$ of the regularized categorical base learner b_0 dummy-encoding every site and all other existing base learners $b_l \in \mathcal{B}$. This allows for potential inclusion of random effects for every fixed effect in the model. More specifically, the l th site-specific effect given by the row-wise tensor product base learner b_{l_\times} uses the basis transformation $g_{l_\times} = g_0 \otimes g_l$

$$g_{l_\times}(\tilde{\mathbf{x}}) = g_0(x_0)^T \otimes g_l(\mathbf{x})^T = \underbrace{(\mathbb{1}_{\{1\}}(x_0)g_l(\mathbf{x})^T)}_{=g_{l_\times,1}}, \dots, \underbrace{\mathbb{1}_{\{S\}}(x_0)g_l(\mathbf{x})^T}_{=g_{l_\times,S}} \quad (2)$$

where the basis transformation g_l is equal for all S sites. After distributed computation (see Eq. (4) in the next section), the estimated coefficients are $\hat{\boldsymbol{\theta}}_{l_\times} = (\hat{\boldsymbol{\theta}}_{l_\times,1}^T, \dots, \hat{\boldsymbol{\theta}}_{l_\times,S}^T)^T$ with $\hat{\boldsymbol{\theta}}_{l_\times,s} \in \mathbb{R}^{d_l}$. The regularization of the row-wise Kronecker base learners not only controls their flexibility but also assures identifiable when additionally including a shared (fixed) effect for the same covariate. The penalty matrix $\mathbf{P}_{l_\times} = \lambda_0 \mathbf{K}_0 \otimes \mathbf{I}_{d_l} + \mathbf{I}_S \otimes \lambda_{l_\times} \mathbf{K}_l \in \mathbb{R}^{Sd_l \times Sd_l}$ is given as Kronecker sum of the penalty matrices \mathbf{K}_0 and \mathbf{K}_l with respective regularization strengths $\lambda_0, \lambda_{l_\times}$. As $\mathbf{K}_0 = \mathbf{I}_S$ is a diagonal matrix, \mathbf{P}_{l_\times} is a block matrix with entries $\lambda_0 \mathbf{I}_{d_l} + \lambda_{l_\times} \mathbf{K}_l$ on the diagonal blocks. Moreover, as g_0 is a binary vector, we can also express the design matrix $\mathbf{Z}_{l_\times} \in \mathbb{R}^{n \times Sd_l}$ as a block matrix, yielding

$$\mathbf{Z}_{l_\times} = \text{diag}(\mathbf{Z}_{l,1}, \dots, \mathbf{Z}_{l,S}), \quad \mathbf{P}_{l_\times} = \text{diag}(\lambda_0 \mathbf{I}_{d_l} + \lambda_{l_\times} \mathbf{K}_l, \dots, \lambda_0 \mathbf{I}_{d_l} + \lambda_{l_\times} \mathbf{K}_l), \quad (3)$$

where $\mathbf{Z}_{l,k}$ are the distributed design matrices of b_l on sites $s = 1, \dots, S$. This Kronecker sum penalty induces a centering of the site-specific effects around zero and, hence, allows the interpretation as deviation from the main effect. Note that possible heredity constraints, such as the one described in Wu and Hamada (2011), are not necessarily met when decomposing effects in this way. However, introducing a restriction that forces the inclusion of the shared effect whenever the respective site-specific effect is selected is a straightforward extension without impairing our proposed framework and without increasing computational costs.

3.3 Fitting algorithm

We now describe the adaptations required to allow for distributed computations of the CWB fitting routine. In Sects. 3.3.1 and 3.3.2, we show the equality between our distributed fitting approach and CWB fitted on pooled data. Section 3.3.3 describes the remaining details such as distributed SSE calculations, distributed model updates, and pseudo residual updates in the distributed setup. Section 3.4 summarizes the distributed CWB algorithm and Sect. 3.5 elaborates on the communication costs of our algorithm.

3.3.1 Distributed shared effects computation

Fitting CWB in a distributed fashion requires adapting the fitting process of the base learner b_l in Algorithm 1 to distributed data. To allow for shared effects computations across different sites without jeopardizing privacy, we take advantage of CWB’s update scheme, which boils down to a (penalized) least squares estimation per iteration for every base learner. This allows us to build upon existing work such as Karr et al. (2005) to fit linear models in a distributed fashion.

ion by just communicating aggregated statistics between sites and the host.

In a first step, the aggregated matrices $F_{l,s} = Z_{l,s}^T Z_{l,s}$ and vectors $u_{l,s} = Z_{l,s}^T y_s$ are computed on each site. In our privacy setup (Sect. 2.4), communicating $F_{l,s}$ and $u_{l,s}$ is allowed as long as the privacy-aggregation level per site is met. In a second step, the site information is aggregated to a global information $F_l = \sum_{s=1}^S F_{l,s} + P_l$ and $u_l = \sum_{s=1}^S u_{l,s}$ and then used to estimate the model parameters $\hat{\theta}_l = F_l^{-1} u_l$. This approach, referred to as distFit, is explained again in detail in Algorithm 2 and used for the shared effect computations of the model by substituting $\hat{\theta}_l^{[m]} = (Z_l^T Z_l + P_l)^{-1} Z_l^T \tilde{r}^{[m]}$ (Algorithm 1 line 6) with $\hat{\theta}_l^{[m]} = \text{distFit}(Z_{l,1}, \dots, Z_{l,S}, \tilde{r}_1^{[m]}, \dots, \tilde{r}_S^{[m]}, P_l)$.

Note that the pseudo residuals $\tilde{r}_k^{[m]}$ are also securely located at each site and are updated after each iteration. Details about the distributed pseudo residuals updates are explained in Sect. 3.3.3. We also note that the computational complexity of fitting CWB can be drastically reduced by pre-calculating and storing $(Z_l^T Z_l + P_l)^{-1}$ in a first initialization step, as the matrix is independent of iteration m , and reusing these pre-calculated matrices in all subsequent iterations (cf. Schalk et al. 2023). Using pre-calculated matrices also reduces the amount of required communication between sites and host.

Algorithm 2 Distributed Effect Estimation.

The line prefixes [S] and [H] indicate whether the operation is conducted at the sites ([S]) or at the host ([H]).

Input Sites design matrices $Z_{l,1}, \dots, Z_{l,S}$, response vectors y_1, \dots, y_S and an optional penalty matrix P_l .

Output Estimated parameter vector $\hat{\theta}_l$.

```

1: procedure distFit( $Z_{l,1}, \dots, Z_{l,S}, y_1, \dots, y_S, P_l$ )
2:   for  $s \in \{1, \dots, S\}$  do
3:     [S]  $F_{l,s} = Z_{l,s}^T Z_{l,s}$ 
4:     [S]  $u_{l,s} = Z_{l,s}^T y_s$ 
5:     [S] Communicate  $F_{l,s}$  and  $u_{l,s}$  to the host
6:   end for
7:   [H]  $F_l = \sum_{s=1}^S F_{l,s} + P_l$ 
8:   [H]  $u_l = \sum_{s=1}^S u_{l,s}$ 
9:   [H] return  $\hat{\theta}_l = F_l^{-1} u_l$ 
10: end procedure
    
```

3.3.2 Distributed site-specific effects computation

If we pretend that the fitting of the base learner b_{l_x} is performed on the pooled data, we obtain

$$\hat{\theta}_{l_x} = (Z_{l_x}^T Z_{l_x} + P_{l_x})^{-1} Z_{l_x}^T y$$

$$= \begin{pmatrix} (Z_{l,1}^T Z_{l,1} + \lambda_0 I_{d_l} + P_l)^{-1} Z_{l,1}^T y_1 \\ \vdots \\ (Z_{l,S}^T Z_{l,S} + \lambda_0 I_{d_l} + P_l)^{-1} Z_{l,S}^T y_S \end{pmatrix}, \tag{4}$$

where (4) is due to the block structure, as described in (3) of Sect. 3.2. This shows that the fitting of the site-specific effects $\hat{\theta}_{l_x}$ can be split up into the fitting of individual parameters

$$\hat{\theta}_{l_x,s} = (Z_{l,s}^T Z_{l,s} + \lambda_0 I_{d_l} + P_l)^{-1} Z_{l,s}^T y_s. \tag{5}$$

It is thus possible to compute site-specific effects at the respective site without the need to share any information with the host. The host, in turn, only requires the SSE of the respective base learner (see next Sect. 3.3.3) to perform the next iteration of CWB. Hence, during the fitting process, the parameter estimates remain at their sites and are just updated if the site-specific base learner is selected. This again minimizes the amount of data communication between sites and host and speeds up the fitting process. After the fitting phase, the aggregated site-specific parameters are communicated once in a last communication step to obtain the final model. A possible alternative implementation that circumvents the need to handle site-specific heterogeneity separately is to apply the estimation scheme of main effects (Algorithm 2). While this simplifies computation, this would increase communication costs and, hence, runtime.

3.3.3 Pseudo residual updates, SSE calculation, and base learner selection

The remaining challenges to run the distributed CWB algorithm are (1) the pseudo residual calculation (Algorithm 1 line 4), (2) the SSE calculation (Algorithm 1 line 7), and (3) base learner selection (Algorithm 1 line 9).

Distributed pseudo residual updates The site-specific response vector y_s containing the values $y^{(i)}$, $i \in \{1, \dots, n_s\}$ is the basis of the pseudo residual calculation. We assume that every site s has access to all shared effects as well as the site-specific information of all site-specific base learners b_{l_x} only containing the respective parameters $\hat{\theta}_{l_x,s}$. Based on these base learners, it is thus possible to compute a *site model* $\hat{f}_s^{[m]}$ as a representative of $\hat{f}^{[m]}$ on every site s . The pseudo residual updates $\tilde{r}_s^{[m]}$ per site are then based on $\hat{f}_s^{[m]}$ via $\tilde{r}_s^{[m](i)} = -\nabla_f L(y^{(i)}, f(\mathbf{x}^{(i)}))|_{f=\hat{f}_s^{[m-1]}}$, $i \in \{1, \dots, n_s\}$ using \mathcal{D}_s . Most importantly, all remaining steps of the distributed CWB fitting procedure do not share the pseudo residuals $\tilde{r}_s^{[m]}$ in order to avoid information leakage about y_s .

Distributed SSE calculation and base learner selection After fitting all base learners $b_l \in \mathcal{B}$ and $b_{l_x} \in \mathcal{B}_x$ to $\tilde{r}_s^{[m]}$, we obtain $\hat{\theta}_l^{[m]}$, $l = 1, \dots, |\mathcal{B}|$, and $\hat{\theta}_{l_x}^{[m]}$, $l_x = 1_x, \dots, |\mathcal{B}_x|$.

Calculating the SSE distributively for the l th and l_x th base learner b_l and b_{l_x} , respectively, requires calculating $2S$ site-specific SSE values:

$$\begin{aligned}
 SSE_{l,s} &= \sum_{i=1}^{n_s} \left(\tilde{r}_s^{[m](i)} - b_l \left(\mathbf{x}_s^{(i)}, \hat{\boldsymbol{\theta}}_l^{[m]} \right) \right)^2 \\
 &= \sum_{i=1}^{n_s} \left(\tilde{r}_s^{[m](i)} - g_l(\mathbf{x}^{(i)})^\top \hat{\boldsymbol{\theta}}_l^{[m]} \right)^2, \\
 SSE_{l_x,s} &= \sum_{i=1}^{n_s} \left(\tilde{r}_s^{[m](i)} - b_{l_x} \left(\mathbf{x}_s^{(i)}, \hat{\boldsymbol{\theta}}_{l_x}^{[m]} \right) \right)^2 \\
 &= \sum_{i=1}^{n_s} \left(\tilde{r}_s^{[m](i)} - g_{l_x}(\mathbf{x}^{(i)})^\top \hat{\boldsymbol{\theta}}_{l_x,s}^{[m]} \right)^2.
 \end{aligned}$$

The site-specific SSE values are then sent to the host and aggregated to $SSE_l = \sum_{s=1}^S SSE_{l,s}$. If privacy constraints have been met in all previous calculations, sharing the individual SSE values is not critical and does not violate any privacy constraints as the value is an aggregation of all n_s observations for all sites s .

Having gathered all SSE values at the host location, selecting the best base learner in the current iteration is done in the exact same manner as for the non-distributed CWB algorithm by selecting $l^{[m]} = \arg \min_{l \in \{1, \dots, |\mathcal{B}|, 1_x, \dots, |\mathcal{B}|_x\}} SSE_l$. After the selection, the index $l^{[m]}$ is shared with all sites to enable the update of the site-specific models $\hat{f}_s^{[m]}$. If a shared effect is selected, the parameter vector $\hat{\boldsymbol{\theta}}_{l^{[m]}}^{[m]}$ is shared with all sites. Caution must be taken when the number of parameters of one base learner is equal to the number of observations, as this allows reverse-engineering private data. In the case of a site-specific effect selection, no parameter needs to be communicated, as the respective estimates are already located at each site.

3.4 Distributed CWB algorithm with site-specific effects

Assembling all pieces, our distributed CWB algorithm is summarized in Algorithm 3.

3.5 Communication costs

While the CWB iterations themselves can be performed in parallel on every site and do not slow down the process compared to a pooled calculation, it is worth discussing the communication costs of distrCWB. During the initialization, data is shared just once, while the fitting phase requires the communication of data in each iteration. Let $d = \max_l d_l$ be the maximum number of basis functions (or, alternatively, assume d basis functions for all base learners). The two main drivers of the communication costs are the number of boost-

Algorithm 3 Distributed CWB Algorithm.

The line prefixes [S] and [H] indicate whether the operation is conducted at the sites ([S]) or at the host ([H]).

Input Sites with site data \mathcal{D}_k , learning rate ν , number of boosting iterations M , loss function L , set of shared effects \mathcal{B} and respective site-specific effects \mathcal{B}_x

Output Prediction model \hat{f}

```

1: procedure distrCWB( $\nu, L, \mathcal{B}, \mathcal{B}_x$ )
2:   Initialization:
3:   [H] Initialize shared model  $\hat{f}_{\text{shared}}^{[0]}(\mathbf{x}) = \arg \min_{c \in \mathbb{R}} \mathcal{R}_{\text{emp}}(c)$ 
4:   [S] Calculate  $\mathbf{Z}_{l,s}$  and  $\mathbf{F}_{l,s} = \mathbf{Z}_{l,s}^\top \mathbf{Z}_{l,s}, \forall l \in \{1, \dots, |\mathcal{B}|\}, s \in \{1, \dots, S\}$ 
5:   [S] Set  $\hat{f}_s^{[0]} = \hat{f}_{\text{shared}}^{[0]}$ 
6:   for  $m \in \{1, \dots, M\}$  or while an early stopping criterion is not met do
7:     [S] Update pseudo residuals:
8:     [S]  $\tilde{r}_s^{[m](i)} = -\nabla_{f} L(y^{(i)}, f(\mathbf{x}^{(i)}))|_{f=\hat{f}_s^{[m-1]}}$ ,  $\forall i \in \{1, \dots, n_s\}$ 
9:     for  $l \in \{1, \dots, |\mathcal{B}|\}$  do
10:      [H] Calculate shared effect:  $\hat{\boldsymbol{\theta}}_l^{[m]} = \text{distFit}(\mathbf{Z}_{l,1}, \dots, \mathbf{Z}_{l,S}, \mathbf{y}_1, \dots, \mathbf{y}_S, \mathbf{P}_l)$ 
11:      [H] Communicate  $\hat{\boldsymbol{\theta}}_l^{[m]}$  to the sites
12:      for  $k \in \{1, \dots, S\}$  do
13:        [S] Fit  $l$ th site-specific effect:  $\hat{\boldsymbol{\theta}}_{l_x,s}^{[m]} = (\mathbf{F}_{l_x,s} + \lambda_0 \mathbf{I}_{d_l} + \mathbf{P}_l)^{-1} \mathbf{Z}_{l_x,s}^\top \tilde{\mathbf{r}}_s^{[m]}$ 
14:        [S] Calculate the SSE for the  $l$ th shared and site-specific effect:
15:        [S]  $SSE_{l,s} = \sum_{i=1}^{n_s} (\tilde{r}_s^{[m](i)} - g_l(\mathbf{x}^{(i)})^\top \hat{\boldsymbol{\theta}}_l^{[m]})^2$ 
16:        [S]  $SSE_{l_x,s} = \sum_{i=1}^{n_s} (\tilde{r}_s^{[m](i)} - g_{l_x}(\mathbf{x}^{(i)})^\top \hat{\boldsymbol{\theta}}_{l_x,s}^{[m]})^2$ 
17:        [S] Send  $SSE_{l,s}$  and  $SSE_{l_x,s}$  to the host
18:      end for
19:      [H] Aggregate SSE values:  $SSE_l = \sum_{s=1}^S SSE_{l,s}$  and  $SSE_{l_x} = \sum_{s=1}^S SSE_{l_x,s}$ 
20:      end for
21:      [H] Select best base learner:  $l^{[m]} = \arg \min_{l \in \{1, \dots, |\mathcal{B}|, 1_x, \dots, |\mathcal{B}|_x\}} SSE_l$ 
22:      if  $b_{l^{[m]}}$  is a shared effect then
23:        [H] Update model:  $\hat{f}_{\text{shared}}^{[m]}(\mathbf{x}) = \hat{f}_{\text{shared}}^{[m-1]}(\mathbf{x}) + \nu b_{l^{[m]}}(\mathbf{x}, \hat{\boldsymbol{\theta}}_{l^{[m]}}^{[m]})$ 
24:        [H] Upload model update  $\hat{\boldsymbol{\theta}}_{l^{[m]}}^{[m]}$  to the sites.
25:      end if
26:      [S] Update site model  $\hat{f}_s^{[m]}$  via parameter updates  $\hat{\boldsymbol{\theta}}_{l^{[m]}}^{[m]} = \hat{\boldsymbol{\theta}}_{l^{[m]}}^{[m-1]} + \nu \hat{\boldsymbol{\theta}}_{l^{[m]}}^{[m]}$ 
27:      end for
28:      [S] Communicate site-specific effects  $\hat{\boldsymbol{\theta}}_{1_x}, \dots, \hat{\boldsymbol{\theta}}_{|\mathcal{B}|_x}$  to the host
29:      [H] Add site-specific effects to the model of shared effects  $\hat{f}_{\text{shared}}^{[M]}$  to obtain the full model  $\hat{f}^{[M]}$ 
30:      [H] return  $\hat{f} = \hat{f}^{[M]}$ 
31: end procedure

```

ing iterations M and the number of base learners $|\mathcal{B}|$. Because of the iterative nature of CWB with a single loop over the boosting iterations, the communication costs (both for the host and each site) scale linearly with the number of boosting iterations M , i.e., $\mathcal{O}(M)$. For the analysis of communication

costs in terms of the number of base learners, we distinguish between the initialization phase and the fitting phase.

Initialization As only the sites share $F_{l,s} \in \mathbb{R}^{d \times d}$, $\forall l \in \{1, \dots, |\mathcal{B}|\}$, the transmitted amount of values is $d^2|\mathcal{B}|$ for each site and therefore scales linearly with $|\mathcal{B}|$, i.e., $\mathcal{O}(|\mathcal{B}|)$. The host does not communicate any values during the initialization.

Fitting In each iteration, every site shares its vector $Z_{l,s}^\top \tilde{r}_s^{[m]} \in \mathbb{R}^d$, $\forall l \in \{1, \dots, |\mathcal{B}|\}$. Over the course of M boosting iterations, each site therefore shares $dM|\mathcal{B}|$ values. Every site also communicates the SSE values, i.e., 2 values (index and SSE value) for every base learner and thus $2M|\mathcal{B}|$ values for all iterations and base learners. In total, each site communicates $M|\mathcal{B}|(d+2)$ values. The communication costs for all sites are therefore $\mathcal{O}(|\mathcal{B}|)$. The host, in turn, communicates the estimated parameters $\hat{\theta}^{[m]} \in \mathbb{R}^d$ of the $|\mathcal{B}|$ shared effects. Hence, $dM|\mathcal{B}|$ values as well as the index of the best base learner in each iteration are transmitted. In total, the host therefore communicates $dM|\mathcal{B}| + M$ values to the sites, and costs are therefore also $\mathcal{O}(|\mathcal{B}|)$.

4 Application

We now showcase our algorithm on a heart disease data set that consists of patient data gathered all over the world. The data were collected at four different sites by the (1) Hungarian Institute of Cardiology, Budapest (Andras Janosi, M.D.), (2) University Hospital, Zurich, Switzerland (William Steinbrunn, M.D.), (3) University Hospital, Basel, Switzerland (Matthias Pfisterer, M.D.), and (4) V.A. Medical Center, Long Beach, and Cleveland Clinic Foundation (Robert Detrano, M.D., Ph.D.), and is thus suited for a multi-site distributed analysis. The individual data sets are freely available at <https://archive.ics.uci.edu/ml/datasets/heart+disease> (Dua and Graff 2017). For our analysis, we set the privacy level (cf. Sect. 2.4) to 5 which is a common default.

4.1 Data description

The raw data set contains 14 covariates, such as the chest pain type (cp), resting blood pressure (trestbps), maximum heart rate (thalach), sex, exercise-induced angina (exang), or ST depression (i.e., abnormal difference of the ST segment from the baseline on an electrocardiogram) induced by exercise relative to rest (oldpeak). A full list of covariates and their abbreviations is given on the data set's website. After removing non-informative (constant) covariates and columns with too many missing values at each site, we obtain $n_{\text{cleveland}} = 303$, $n_{\text{hungarian}} = 292$, $n_{\text{switzerland}} = 116$, and $n_{\text{va}} = 140$ observations and 8 covariates. A table containing the description of the abbreviations

of these covariates is given in Table 1 in the Supplementary Material B.1. For our application, we assume that missing values are completely at random and all data sets are exclusively located at each sites. The task is to determine important risk factors for heart diseases. The target variable is therefore a binary outcome indicating the presence of heart disease or not.

4.2 Analysis and results

We follow the practices to setup CWB as mentioned in Sect. 2.3.2 and run the distributed CWB algorithm with a learning rate of 0.1 and a maximum number of 100,000 iterations. To determine an optimal stopping iteration for CWB, we use 20 % of the data as validation data and set the patience to 5 iterations. In other words, the algorithm stops if no risk improvement on the validation data is observed in 5 consecutive iterations. For the numerical covariates, we use a P-spline with 10 cubic basis functions and second-order difference penalties. All base learners are penalized according to a global degree of freedom that we set to 2.2 (to obtain unbiased feature selection) while the random intercept is penalized according to 3 degrees of freedom (see the Supplementary Material B.2 for more details). Since we are modelling a binary response variable, h^{-1} is the inverse logit function $\text{logit}^{-1}(f) = (1 + \exp(-f))^{-1}$. The model for an observation of site s , conditional on its random effects $\boldsymbol{\gamma}$, is given in the Supplementary Material B.3.

Results The algorithm stops after $m_{\text{stop}} = 5578$ iterations as the risk on the validation data set starts to increase (cf. Figure 1 in the Supplementary Material B.4) and selects covariates oldpeak, cp, trestbps, age, sex, restecg, exang, and thalach. Out of these 5578 iterations, the distributed CWB algorithm selects a shared effect in 782 iterations and site-specific effects in 4796 iterations. This indicates that the data is rather heterogeneous and requires site-specific (random) effects. We want to emphasize that the given data is from an observational study and that the sole purpose of our analysis is to better understand the heterogeneity in the data. Hence, the estimated effects have a non-causal relationship. To alleviate problems that come from such data, and allow for the estimation of causal effects, one could use, e.g., propensity score matching (Rosenbaum and Rubin 1983) before applying our algorithm. From our application, we can, e.g., see that the data from Hungary could potentially be problematic in this respect. However, note that applying such measures would also have to be done in a privacy-preserving manner. Figure 2 (Left) shows traces of how and when the different additive terms (base learners) entered the model during the fitting process and illustrates the selection process of CWB.

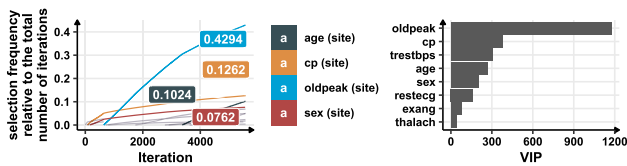


Fig. 2 Left: Model trace showing how and when the four most selected additive terms entered the model. Right: Variable importance (cf. Au et al. 2019) of selected features in decreasing order

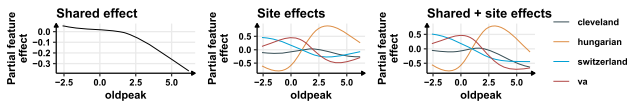


Fig. 3 Decomposition of the effect of `oldpeak` into the shared (left) and the site-specific effects (middle). The plot on the right-hand side shows the sum of shared and site-specific effects

The estimated effect of the most important feature `oldpeak` (cf. Fig. 2, Right) found is further visualized in Fig. 3. Looking at the shared effect, we find a negative influence on the risk of heart disease when increasing ST depression (`oldpeak`). When accounting for site-specific deviations, the effect becomes more diverse, particularly for Hungary.

In the Supplementary Material B.5 and B.6, we provide the partial effects for all features and showcase the conditional predictions of the fitted GAMM model for a given site.

Comparison of estimation approaches The previous example shows partial feature effects that exhibit shrinkage due to the early stopping of CWB’s fitting routine. While this prevents overfitting and induces a sparse model, we can also run CWB for a very large amount of iterations without early stopping to approximate the unregularized and hence unbiased maximum likelihood solution. We illustrate this in the following by training CWB and our distributed version for 100,000 iterations and compare its partial effects to the ones of a classical mixed model-based estimation routine implemented in the R package `mgcv` (Wood 2017). Our R prototype took ≈ 3.5 h for 100,000 iterations and ≈ 700 s with early stopping after 5578 iterations. The corresponding computation on a local machine with `compboost` took ≈ 25 min and ≈ 85 s, respectively. We want to emphasize that the runtime of our algorithm strongly depends on the distributed system that controls the communication as well as the bandwidth.

Results of the estimated partial effects of our distributed CWB algorithm and the original CWB on pooled data show a perfect overlap (cf. Fig. 4). This again underpins the lossless property of the proposed algorithm. The site-specific effects on the pooled data are fitted by defining a row-wise Kronecker base learner for all features and the site as a categorical variable. The same approach is used to estimate a GAMM using `mgcv` fitted on the pooled data with tensor products

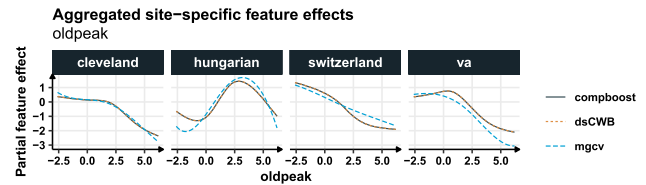


Fig. 4 Comparison of the site-specific effects for `oldpeak` between the distributed (`dsCWB`) and pooled CWB approach (`compboost`) as well as estimates of from `mgcv`

between the main feature and the categorical site variable. A comparison of all partial feature effects is given in the Supplementary Material B.7 showing good alignment between the different methods. For the `oldpeak` effect shown in Fig. 4, we also see that the partial effects of the two CWB methods are very close to the mixed model-based estimation, with only smaller differences caused by a slightly different penalization strength of both approaches. The empirical risk is 0.4245 for our distributed CWB algorithm, 0.4245 for CWB on the pooled data, and 0.4441 for the GAMM on the pooled data.

5 Discussion

We proposed a novel algorithm for distributed, lossless, and privacy-preserving GAMM estimation to analyze horizontally partitioned data. To account for data heterogeneity of different sites we introduced site-specific (smooth) random effects. Using CWB as the fitting engine allows estimation in high-dimensional settings and fosters variable as well as effect selection. This also includes a data-driven selection of shared and site-specific features, providing additional data insights. Owing to the flexibility of boosting and its base learners, our algorithm is easy to extend and can also account for interactions, functional regression settings (Brockhaus et al. 2020), or modeling survival tasks (Bender et al. 2020).

An open challenge for the practical use of our approach is its high communication costs. For larger iterations (in the 10 or 100 thousands), computing a distributed model can take several hours. One option to reduce the total runtime is to incorporate accelerated optimization recently proposed in Schalk et al. (2023). Another driver that influences the runtime is the latency of the technical setup. Future improvements could reduce the number of communications, e.g., via multiple fitting rounds at the different sites before communicating the intermediate results.

A possible future extension of our approach is to account for both horizontally and vertically distributed data. Since the algorithm is performing component-wise (coordinate-wise) updates, the extension to vertically distributed data naturally falls into the scope of its fitting procedure. This would, how-

ever, require a further advanced technical setup and the need to ensure consistency across sites.

As an alternative to CWB, a penalized likelihood approach like `mgcv` could be considered for distributed computing. Unlike CWB, which benefits from parallelized base learners, decomposing the entire design matrix for distributed computing with this approach is more intricate. The parallelization strategy of Wood et al. (2017) could be adapted by viewing cores as sites and the main process as the host. However, ensuring privacy for this approach would require additional attention. A notable obstacle for smoothing parameter estimation is the requirement of the Hessian matrix Wood et al. (2016). Since the Hessian matrix cannot be directly computed from distributed data, methods like subsampling (Umlauf et al. 2023) or more advanced techniques would be necessary to achieve unbiased estimates and convergence of the whole process. In general, unlike CWB which fits pseudo-residuals using the L_2 -loss and estimates smoothness implicitly through iterative gradient updates, penalized likelihood approaches such as the one implemented in `mgcv` are less straightforward to distribute, and a privacy-preserving lossless computation would involve specialized procedures.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s11222-023-10323-2>.

Author Contributions DS developed the idea, its theoretical details and its implementation in R. He further conducted all experiments and practical applications. The manuscript was mainly written by DS. Connection to GAMMs have been worked out by D.R., who also wrote the corresponding text passages. BB and DR helped revising and finalizing the manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

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/>.

References

- Anjum, M.M., Mohammed, N., Li, W., et al.: Privacy preserving collaborative learning of generalized linear mixed model. *J. Biomed. Inform.* **127**(104), 008 (2022)
- Au, Q., Schalk, D., Casalicchio, G., et al.: Component-wise boosting of targets for multi-output prediction. *arXiv preprint arXiv:1904.03943* (2019)
- Augustyn, D.R., Wyciřlik, Ł., Mrozek, D.: Perspectives of using cloud computing in integrative analysis of multi-omics data. *Brief. Funct. Genom.* **20**(4), 198–206 (2021). <https://doi.org/10.1093/bfgp/elab007>
- Bazeley, P.: Integrative analysis strategies for mixed data sources. *Am. Behav. Sci.* **56**(6), 814–828 (2012)
- Bender, A., Rügamer, D., Scheipl, F., et al.: A general machine learning framework for survival analysis. In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, pp. 158–173 (2020)
- Boyd, K., Lantz, E., Page, D.: Differential privacy for classifier evaluation. In: *Proceedings of the 8th ACM Workshop on Artificial Intelligence and Security*, pp. 15–23 (2015)
- Brockhaus, S., Rügamer, D., Greven, S.: Boosting functional regression models with FDboost. *J. Stat. Softw.* **94**(10), 1–50 (2020)
- Brumback, B.A., Ruppert, D., Wand, M.P.: Variable selection and function estimation in additive nonparametric regression using a data-based prior: Comment. *J. Am. Stat. Assoc.* **94**(447), 794–797 (1999)
- Bühlmann, P., Yu, B.: Boosting with the L2 loss: regression and classification. *J. Am. Stat. Assoc.* **98**(462), 324–339 (2003)
- Bühlmann, P., Hothorn, T., et al.: Boosting algorithms: regularization, prediction and model fitting. *Stat. Sci.* **22**(4), 477–505 (2007)
- Chen, Y.R., Rezapour, A., Tzeng, W.G.: Privacy-preserving ridge regression on distributed data. *Inf. Sci.* **451**, 34–49 (2018)
- Curran, P.J., Hussong, A.M.: Integrative data analysis: the simultaneous analysis of multiple data sets. *Psychol. Methods* **14**(2), 81 (2009)
- Dua, D., Graff, C.: UCI machine learning repository. <http://archive.ics.uci.edu/ml> (2017)
- Dwork, C.: Differential privacy. In: *International Colloquium on Automata, Languages, and Programming*. Springer, pp. 1–12 (2006)
- Eilers, P.H., Marx, B.D.: Flexible smoothing with B-splines and penalties. *Stat. Sci.* **11**, 89–102 (1996)
- Gambs, S., Kégl, B., Aïmeur, E.: Privacy-preserving boosting. *Data Min. Knowl. Disc.* **14**(1), 131–170 (2007)
- Gaye, A., Marcon, Y., Isaeva, J., et al.: Datashield: taking the analysis to the data, not the data to the analysis. *Int. J. Epidemiol.* **43**(6), 1929–1944 (2014)
- Hoerl, A.E., Kennard, R.W.: Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* **12**(1), 55–67 (1970)
- Hofner, B., Hothorn, T., Kneib, T., et al.: A framework for unbiased model selection based on boosting. *J. Comput. Graph. Stat.* **20**(4), 956–971 (2011)
- Jones, E.M., Sheehan, N.A., Gaye, A., et al.: Combined analysis of correlated data when data cannot be pooled. *Stat* **2**(1), 72–85 (2013)
- Karr, A.F., Lin, X., Sanil, A.P., et al.: Secure regression on distributed databases. *J. Comput. Graph. Stat.* **14**(2), 263–279 (2005)
- Lazarevic, A., Obradovic, Z.: The distributed boosting algorithm. In: *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 311–316 (2001)
- Li, J., Kuang, X., Lin, S., et al.: Privacy preservation for machine learning training and classification based on homomorphic encryption schemes. *Inf. Sci.* **526**, 166–179 (2020a)

- Li, Q., Wen, Z., He, B.: Practical federated gradient boosting decision trees. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 4642–4649 (2020b)
- Liew, B.X., Rügamer, D., Abichandani, D., et al.: Classifying individuals with and without patellofemoral pain syndrome using ground force profiles—development of a method using functional data boosting. *Gait Posture* **80**, 90–95 (2020)
- Litwin, W., Mark, L., Roussopoulos, N.: Interoperability of multiple autonomous databases. *ACM Comput. Surv. (CSUR)* **22**(3), 267–293 (1990)
- Lu, C.L., Wang, S., Ji, Z., et al.: Webdisco: a web service for distributed cox model learning without patient-level data sharing. *J. Am. Med. Inform. Assoc.* **22**(6), 1212–1219 (2015)
- Luo, C., Islam, M., Sheils, N.E., et al.: DLMM as a lossless one-shot algorithm for collaborative multi-site distributed linear mixed models. *Nat. Commun.* **13**(1), 1–10 (2022)
- McCullagh, P., Nelder, J.A.: *Generalized Linear Models*, 2nd edn. Routledge, Milton Park (1989)
- McMahan, B., Moore, E., Ramage, D., et al.: Communication-efficient learning of deep networks from decentralized data. In: *Artificial Intelligence and Statistics*, PMLR, pp. 1273–1282 (2017)
- Mirza, B., Wang, W., Wang, J., et al.: Machine learning and integrative analysis of biomedical big data. *Genes* **10**(2), 87 (2019)
- Mohassel, P., Zhang, Y.: SecureML: a system for scalable privacy-preserving machine learning. In: *2017 IEEE Symposium on Security and Privacy (SP)*, IEEE, pp. 19–38 (2017)
- Næhrig, M., Lauter, K., Vaikuntanathan, V.: Can homomorphic encryption be practical? In: *Proceedings of the 3rd ACM Workshop on Cloud Computing Security Workshop*, pp. 113–124 (2011)
- Paillier, P.: Public-key cryptosystems based on composite degree residuosity classes. In: *International Conference on the Theory and Applications of Cryptographic Techniques*. Springer, pp. 223–238 (1999)
- R Core Team: *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, <https://www.R-project.org/> (2021)
- Rosenbaum, P.R., Rubin, D.B.: The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**(1), 41–55 (1983)
- Rügamer, D., Brockhaus, S., Gentsch, K., et al.: Boosting factor-specific functional historical models for the detection of synchronization in bioelectrical signals. *J. R. Stat. Soc.: Ser. C (Appl. Stat.)* **67**(3), 621–642 (2018). <https://doi.org/10.1111/rssc.12241>
- Saintigny, P., Zhang, L., Fan, Y.H., et al.: Gene expression profiling predicts the development of oral cancer. *Cancer Prev. Res.* **4**(2), 218–229 (2011)
- Samarati, P., Sweeney, L.: Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Technical Report, <http://www.csl.sri.com/papers/sritr-98-04/> (1998)
- Schalk, D., Hoffmann, V.S., Bischl, B., et al.: Distributed non-disclosive validation of predictive models by a modified ROC-GLM. arXiv preprint [arXiv:2203.10828](https://arxiv.org/abs/2203.10828) (2022)
- Schalk, D., Bischl, B., Rügamer, D.: Accelerated component wise gradient boosting using efficient data representation and momentum-based optimization. *J. Comput. Graph. Stat.* **32**(2), 631–641 (2023). <https://doi.org/10.1080/10618600.2022.2116446>
- Schmid, M., Hothorn, T.: Boosting additive models using component-wise p-splines. *Comput. Stat. Data Anal.* **53**(2), 298–311 (2008)
- Sweeney, L.: k-anonymity: a model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.* **10**(05), 557–570 (2002)
- Umlauf, N., Seiler, J., Wetscher, M., et al.: Scalable estimation for structured additive distributional regression. arXiv preprint [arXiv:2301.05593](https://arxiv.org/abs/2301.05593) (2023)
- Ünal, A.B., Pfeifer, N., Akgün, M.: ppAURORA: privacy preserving area under receiver operating characteristic and precision-recall curves with secure 3-party computation. arXiv 2102 (2021)
- Wood, S.N.: *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC, Boca Raton (2017)
- Wood, S.N., Pya, N., Säfken, B.: Smoothing parameter and model selection for general smooth models. *J. Am. Stat. Assoc.* **111**(516), 1548–1563 (2016). <https://doi.org/10.1080/01621459.2016.1180986>
- Wood, S.N., Li, Z., Shaddick, G., et al.: Generalized additive models for gigadata: modeling the U.K. black smoke network daily data. *J. Am. Stat. Assoc.* **112**(519), 1199–1210 (2017). <https://doi.org/10.1080/01621459.2016.1195744>
- Wu, C.J., Hamada, M.S.: *Experiments: Planning, Analysis, and Optimization*. Wiley, Hoboken (2011)
- Wu, Y., Jiang, X., Kim, J., et al.: Grid binary logistic regression (GLORE): building shared models without sharing data. *J. Am. Med. Inform. Assoc.* **19**(5), 758–764 (2012)
- Yan, Z., Zachrisson, K.S., Schwamm, L.H., et al.: Fed-GLMM: a privacy-preserving and computation-efficient federated algorithm for generalized linear mixed models to analyze correlated electronic health records data. medRxiv (2022)
- Zhu, R., Jiang, C., Wang, X., et al.: Privacy-preserving construction of generalized linear mixed model for biomedical computation. *Bioinformatics* **36**(Supplement-1), i128–i135 (2020). <https://doi.org/10.1093/bioinformatics/btaa478>

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