Electronic Journal of Statistics
Vol. 14 (2020) 272–301
ISSN: 1935-7524
https://doi.org/10.1214/19-EJS1666

Assessing prediction error at interpolation and extrapolation points[*]

Assaf Rabinowicz

e-mail: assafrab@gmail.com
Department of Statistics, Tel-Aviv University, Tel-Aviv, Israel, 69978

and

Saharon Rosset

e-mail: saharon@tauex.tau.ac.il
Department of Statistics, Tel-Aviv University, Tel-Aviv, Israel, 69978

Abstract: Common model selection criteria, such as AIC and its variants, are based on in-sample prediction error estimators. However, in many applications involving predicting at interpolation and extrapolation points, in-sample error does not represent the relevant prediction error. In this paper new prediction error estimators, tAI and $Loss(w_t)$ are introduced. These estimators generalize previous error estimators, however are also applicable for assessing prediction error in cases involving interpolation and extrapolation. Based on these prediction error estimators, two model selection criteria with the same spirit as AIC and Mallow's C_p are suggested. The advantages of our suggested methods are demonstrated in a simulation and a real data analysis of studies involving interpolation and extrapolation in linear mixed model and Gaussian process regression.

Keywords and phrases: Model assessment, model selection, AIC, expected optimism, linear mixed models, Kriging.

Received September 2018.

1. Introduction

Predicting a phenomenon at different points than the points appearing in the training sample plays an important role across many research fields such as in geostatistics (Li and Heap, 2014; Kyriakidis and Journel, 1999), health (Manton, Singer and Suzman, 2012) and econometrics (Baltagi, 2008). In many of these use cases the new predicted points are interpolation or extrapolation points with respect to space or to time. For example, Brown and Comrie (2002) interpolated climate values in Southwestern U.S., where the coverage of climate information is sparse. By predicting at interpolation points, they created a high-resolution map of seasonal temperature and precipitation in this area. An extrapolation example is given by Stewart, Cutler and Rosen (2009) who forecast the effects

[*] The authors gratefully acknowledge *Israel Science Foundation, grant 1804/16.*

of obesity and smoking on U.S. life expectancy in 2020 by using a data set for the years 2003 through 2006.

Modeling approaches involving prediction at pre-specified interpolation and extrapolation points were studied in machine learning, mainly in the context of transductive support-vector machines (Joachims, 1999), however also in regression (Le et al., 2006).

Assessing prediction error at interpolation and extrapolation points, or more generally at transduction points, cannot be done using traditional in-sample prediction error estimators as in AIC (Akaike, 1974) and its variants. Similarly, K-fold cross-validation, which estimates the generalization error, is also unsuitable in these cases, where prediction points are specified.

This paper introduces a prediction error estimator, tAI, which generalizes previous in-sample prediction error estimators like $mAIC$ (Vaida and Blanchard, 2005) and $cAIC$ (Vaida and Blanchard, 2005), however, it does not assume that the predicted points are the same as the points appearing in the training sample and therefore is applicable to a wider range of use cases, such as cases involving prediction at interpolation and extrapolation points. Since prediction error assessment is highly related to model selection, a new model selection criterion, $tAIC$, which is based on tAI, is proposed as well. tAI is suitable when the observations are normally distributed, whether they are correlated or not and therefore is applicable for various parametric models with different variance structure assumptions such as linear mixed model (LMM), Gaussian process regression (GPR), generalized least squares (GLS) and linear regression. Relaxing the normality requirement of tAI, we also propose in Section 4 an approach for inference on interpolation and extrapolation that is based on squared error loss rather than likelihood, and hence generalizes the optimism approach in model selection (Efron, 1986).

In many use cases involving predicting at interpolation and extrapolation points, the dependent variable has a correlation structure (Li and Heap, 2014; Kyriakidis and Journel, 1999). For example, in the use case given by Brown and Comrie (2002), it is natural to assume a spatial correlation structure on the Southwestern U.S. area. Similarly, in repeated measures studies that forecast long-term treatment effects, a correlation structure with respect to time is commonly assumed (Ho et al., 2011). Therefore, use cases involving correlated data and models that are implemented on correlated data, such as LMM, GPR and GLS, are good platforms for analyzing how predicting at interpolation and extrapolation points influences prediction error estimation and model selection. Before introducing tAI, we define a setup which puts LMM, GPR and GLS under a unified framework:

Let $\boldsymbol{y} \in \mathbb{R}^n$ and the fixed matrices $\{X \in \mathbb{R}^{n\times p}, Z \in \mathbb{R}^{n\times q}\}$ be a training sample, $\boldsymbol{y^*} \in \mathbb{R}^{n^*}$ and the fixed matrices $\{X^* \in \mathbb{R}^{n^*\times p}, Z^* \in \mathbb{R}^{n^*\times q}\}$ be a prediction set, where

$$\boldsymbol{y} \sim N(\boldsymbol{\mu}, V),\ \boldsymbol{y^*} \sim N(\boldsymbol{\mu^*}, V^*), \tag{1}$$

$\boldsymbol{\mu} = X\boldsymbol{\beta}$, $\boldsymbol{\mu^*} = X^*\boldsymbol{\beta}$, V is a function of Z, and V^* is a function of Z^*. For example, in LMM it is typically assumed that the columns of Z, Z^* are associated with normally distributed random effects with covariance matrix $G \in \mathbb{R}^{q\times q}$ such that the marginal covariance matrices of $\boldsymbol{y}$ and $\boldsymbol{y^*}$ are:

$$\begin{aligned} V &= ZGZ^t + \sigma^2 I_n,\ \ \sigma \in \mathbb{R}^+ \\ V^* &= Z^*GZ^{*t} + \sigma^2 I_{n^*}, \end{aligned}$$

where I_n, I_{n^*} are the identity matrices with dimensions n and n^* respectively. In GPR it is often assumed that the marginal covariance matrices of $\boldsymbol{y}$ and $\boldsymbol{y^*}$ are:

$$\begin{aligned} V &= K(Z,Z) + \sigma^2 I_n \\ V^* &= K(Z^*,Z^*) + \sigma^2 I_{n^*}, \end{aligned}$$

where K is some kernel function.

In addition, denote the following conditional covariance matrices:

$$\begin{aligned} R^* &= \mathrm{Var}(\boldsymbol{y^*}|\boldsymbol{y}) \qquad (2) \\ R &= \mathrm{Var}(\boldsymbol{y^{new}}|\boldsymbol{y}), \qquad (3) \end{aligned}$$

where $\boldsymbol{y^{new}}$ is $\boldsymbol{y^*}$ for the special case when X^*, Z^*, and V^* are restricted to be equal to X, Z, and V. This also yields that $\boldsymbol{y^{new}}$ is distributed as $\boldsymbol{y}$, i.e., $\boldsymbol{y^{new}} \sim N(\boldsymbol{\mu}, V)$.

By normality of $\boldsymbol{y}$ and $\boldsymbol{y^*}$,

$$\mathbb{E}(\boldsymbol{y^*}|\boldsymbol{y}) = X^*\boldsymbol{\beta} + \mathrm{Cov}(\boldsymbol{y^*},\boldsymbol{y})V^{-1}(\boldsymbol{y} - X\boldsymbol{\beta}).$$

Given V, $\mathrm{Cov}(\boldsymbol{y^*},\boldsymbol{y})$ and the ML estimator of $\boldsymbol{\beta}$,

$$\hat{\boldsymbol{\beta}} = (X^tV^{-1}X)^{-1}X^tV^{-1}\boldsymbol{y},$$

$\mathbb{E}(\boldsymbol{y^*}|\boldsymbol{y})$ can be used for predicting $\boldsymbol{y^*}$ as follows

$$\begin{aligned} \hat{\boldsymbol{f}}^{\boldsymbol{*}} =& \hat{\mathbb{E}}(\boldsymbol{y^*}|\boldsymbol{y}) \qquad (4) \\ =& X^*(X^tV^{-1}X)^{-1}X^tV^{-1}\boldsymbol{y} \\ &+ \mathrm{Cov}(\boldsymbol{y^*},\boldsymbol{y})V^{-1}\left\{I_n - X(X^tV^{-1}X)^{-1}X^tV^{-1}\right\}\boldsymbol{y}. \end{aligned}$$

This procedure generalizes standard prediction practices in LMM, GPR and GLS. In addition, $\hat{\boldsymbol{f}}^{\boldsymbol{*}}$ is the Best Linear Unbiased Predictor (BLUP) (Harville et al., 1976).

tAI is an estimator of the following prediction error,

$$-\frac{1}{n^*}\mathbb{E}_{\boldsymbol{y^*}|\boldsymbol{y}}\ell(\boldsymbol{y^*}) = -\frac{1}{n^*}\mathbb{E}_{\boldsymbol{y^*}|\boldsymbol{y}}\log\left[\frac{\exp\left\{\frac{-1}{2}(\boldsymbol{y^*}-\hat{\boldsymbol{f}}^{\boldsymbol{*}})^t{R^*}^{-1}(\boldsymbol{y^*}-\hat{\boldsymbol{f}}^{\boldsymbol{*}})\right\}}{\sqrt{(2\pi)^{n^*}|R^*|}}\right]. \qquad (5)$$

Correspondingly, given a set of candidate models, $tAIC$ would be defined as a model selection criterion selecting a model with the minimal tAI. In such a way, $tAIC$ selects the model minimizing the unbiased estimate of eq. (5).

This methodology, of estimating the prediction errors for different models and then selecting the model with the minimal prediction error, is the same as is implemented in AIC and its variants.

$\{X^*, Z^*, R^*\} = \{X, Z, R\}$ is not assumed in the setup above and in its associated prediction error measure, eq. (5). Therefore, tAI is applicable in various use cases that require flexibility in defining $\{X^*, Z^*, R^*\}$. For example, in the use case mentioned above of Brown and Comrie (2002), where GPR is used for predicting interpolated climate values (Kriging), it is reasonable to define $\{X^*, Z^*\}$ as the data points at the high-resolution spatial array rather than as the data points at the training sample, $\{X, Z\}$, which cover the area sparsely. Therefore, while prediction error estimators that are based on in-sample error estimation and generalization error estimation are unsuitable to this case, tAI is suitable. For similar considerations, $tAIC$ is required in repeated measures studies in health and biomedicine, when the main interest is to select LMM model minimizing the prediction error at long-term points, $\{X^*, Z^*, R^*\}$, which are different than the points that are used for model building, $\{X, Z, R\}$ (Pope et al., 2002; Li et al., 2008).

Besides downscaling of climate maps and estimating long-term effect in clinical studies, interpolation and extrapolation using LMM and Kriging are important tools for many research topics in mining engineering, agriculture, environmental sciences, especially when sampling is difficult and expensive like in mountainous and deep marine regions (Li and Heap, 2011; Stahl et al., 2006; Vicente-Serrano, Saz-Sánchez and Cuadrat, 2003). tAI and $tAIC$ are relevant for all these research topics as well as for others which do not involve interpolation and extrapolation but still do not satisfy $\{X^*, Z^*, R^*\} = \{X, Z, R\}$. Various use cases will be presented and analyzed in Sections 3 and 5.

2. *tAI* and *tAIC*

tAI is derived by estimating $-\mathbb{E}_{\boldsymbol{y}^*|\boldsymbol{y}}\ell(\boldsymbol{y}^*)/n^*$ by the averaged log-likelihood of the training sample,

$$-\frac{1}{n}\ell(\boldsymbol{y}) = -\frac{1}{n}\log\left[\frac{\exp\left\{\frac{-1}{2}(\boldsymbol{y}-\hat{\boldsymbol{f}})^t R^{-1}(\boldsymbol{y}-\hat{\boldsymbol{f}})\right\}}{\sqrt{(2\pi)^n|R|}}\right], \tag{6}$$

plus a penalty correction

$$C_{tAI} = \mathbb{E}_{\boldsymbol{y}}\left[-\frac{1}{n^*}\mathbb{E}_{\boldsymbol{y}^*|\boldsymbol{y}}\ell(\boldsymbol{y}^*) - \left\{-\frac{1}{n}\ell(\boldsymbol{y})\right\}\right], \tag{7}$$

where

$$\hat{\boldsymbol{f}} = X\hat{\boldsymbol{\beta}} + (V-R)V^{-1}(\boldsymbol{y}-X\hat{\boldsymbol{\beta}})$$

is the estimated conditional expectation, $\hat{\mathbb{E}}(\boldsymbol{y}^{\boldsymbol{new}}|\boldsymbol{y})$.

This approach of estimating prediction error by deriving the bias of the training error is also used in AIC and its variants (Akaike, 1974). Consequently, the estimator

$$tAI = -\frac{1}{n}\ell(\boldsymbol{y}) + C_{tAI}$$

does not contain $\boldsymbol{y^*}$ but still satisfies

$$E_{\boldsymbol{y}} tAI = \mathbb{E}_{\boldsymbol{y}}\left\{-\frac{1}{n^*}\mathbb{E}_{\boldsymbol{y^*}|\boldsymbol{y}}\ell(\boldsymbol{y^*})\right\},$$

and tAI can be seen either as an estimator of $-\mathbb{E}_{\boldsymbol{y^*}|\boldsymbol{y}}\ell(\boldsymbol{y^*})/n^*$ or of its expectation, $-\mathbb{E}_{\boldsymbol{y}}\mathbb{E}_{\boldsymbol{y^*}|\boldsymbol{y}}\ell(\boldsymbol{y^*})/n^* = -\mathbb{E}_{\boldsymbol{y},\boldsymbol{y^*}}\ell(\boldsymbol{y^*})/n^*$. The difference between defining the estimated prediction error with or without expectation over $\boldsymbol{y}$, is also relevant for other AIC-like prediction error measures and a detailed discussion of it is out of the scope of this paper. A relevant discussion can be found in Hastie, Tibshirani and Friedman (2009).

The calculation of C_{tAI} in eq. (7) is addressed by Theorem 1 below. Note that Theorem 1 introduces a transductive correction for a setup that is more general than the setup in eq. (7), i.e., Theorem 1 is relevant for other models besides LMM and GLS.

Theorem 1. *Consider the setup given in eq. (1), (2) and (3). Also, let*

$$\ell_{H^*}(\boldsymbol{y^*}) = \log\left[\frac{\exp\left\{\frac{-1}{2}(\boldsymbol{y^*} - H^*\boldsymbol{y})^t R^{*^{-1}}(\boldsymbol{y^*} - H^*\boldsymbol{y})\right\}}{\sqrt{(2\pi)^{n^*}|R^*|}}\right]$$

$$\ell_H(\boldsymbol{y}) = \log\left[\frac{\exp\left\{\frac{-1}{2}(\boldsymbol{y} - H\boldsymbol{y})^t R^{-1}(\boldsymbol{y} - H\boldsymbol{y})\right\}}{\sqrt{(2\pi)^{n}|R|}}\right],$$

where $H\boldsymbol{y} \in \mathbb{R}^n$ and $H^*\boldsymbol{y} \in \mathbb{R}^{n^*}$ are linear predictors of $\boldsymbol{y}$ and $\boldsymbol{y^*}$ respectively, i.e., H and H^* do not contain $\boldsymbol{y}$, $\boldsymbol{y^*}$. If $H\boldsymbol{\mu} = \boldsymbol{\mu}$, $H^*\boldsymbol{\mu} = \boldsymbol{\mu^*}$ are satisfied, then:

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{y}}\left\{-\frac{1}{n^*}\mathbb{E}_{\boldsymbol{y^*}|\boldsymbol{y}}\ell_{H^*}(\boldsymbol{y^*}) + \frac{1}{n}\ell_H(\boldsymbol{y})\right\}\\
&\qquad = \frac{1}{n}\mathrm{tr}\left(R^{-1}HV\right) - \frac{1}{n^*}\mathrm{tr}\left(R^{*^{-1}}H^*\mathrm{Cov}(\boldsymbol{y},\boldsymbol{y^*})\right)\\
&\qquad + \frac{1}{2}\left\{\log\left(\frac{|R^*|^{\frac{1}{n^*}}}{|R|^{\frac{1}{n}}}\right) + \frac{1}{n^*}\mathrm{tr}\left(R^{*^{-1}}V^*\right) - \frac{1}{n}\mathrm{tr}\left(R^{-1}V\right)\right\}\\
&\qquad + \frac{1}{2}\left\{\frac{1}{n^*}\mathrm{tr}\left(R^{*^{-1}}H^*VH^{*t}\right) - \frac{1}{n}\mathrm{tr}\left(R^{-1}HVH^t\right)\right\},
\end{aligned}$$

when tr *is the trace operator.*

The proof is based on standard algebra arguments and is included in Appendix A.1.

The conditions in Theorem 1 are satisfied by the BLUPs, $\hat{\boldsymbol{f}}$ and $\hat{\boldsymbol{f}}^{\boldsymbol{*}}$, where

$$\begin{aligned}
&\hat{\boldsymbol{f}} = H\boldsymbol{y}\\
&H = X(X^tV^{-1}X)^{-1}X^tV^{-1} + (V - R)V^{-1}\left\{I_n - X(X^tV^{-1}X)^{-1}X^tV^{-1}\right\},
\end{aligned}$$

and

$$\hat{\boldsymbol{f}}^{*} = H^{*}\boldsymbol{y}$$
$$H^{*} = X^{*}(X^{t}V^{-1}X)^{-1}X^{t}V^{-1} + \mathrm{Cov}(\boldsymbol{y^{*}}, \boldsymbol{y})V^{-1}\{I_{n} - X(X^{t}V^{-1}X)^{-1}X^{t}V^{-1}\}.$$

Therefore, Theorem 1 can be utilized for implementing tAI.

Besides prediction error estimation, these results can be used for defining the following model selection criterion.

Definition 2. *Given set of models $\mathcal{H}$, satisfying the conditions in Theorem 1, $tAIC$ is the following criterion:*

$$h_{best} = \operatorname*{argmin}_{h \in \mathcal{H}} tAI_{h},$$

where tAI_{h} is tAI for model h.

$tAIC$ is based on the same idea that AIC and other model selection criteria are based on – selecting the model minimizing the unbiased estimate of the in-sample error.

2.1. Comparison with other prediction error estimators

The prediction error estimators that appear in $cAIC$ and $mAIC$ (Vaida and Blanchard, 2005) are AIC versions for normal linear models for the case when the data has a correlation structure, however assuming $\{X, Z, R\} = \{X^{*}, Z^{*}, R^{*}\}$. They use the random effects framework in order to express the data correlation structure. $cAIC$, conditional AIC, is based on conditional likelihood given the random effects. $cAIC$ is mainly used as a model selection criterion for LMM, i.e., when the prediction goal is to predict new observations that share the same random effects realizations with the training set (e.g., predicting observations that relate to the same clusters as the observations in the training set). $mAIC$, marginal AIC, is based on marginal likelihood with respect to the random effects. $mAIC$ is mainly used as a model selection criterion for GLS, i.e., when the prediction goal is to predict new observations without assuming that the new observations and the training set share the same random effects realizations (e.g., predicting observations that do not necessarily relate to the same clusters as the observations in the training set). For more information about the distinction between marginal and conditional inference see Vaida and Blanchard (2005); Verbeke and Molenberghs (2009).

Given known covariance matrices, V and $R = \sigma^{2}I$, $\sigma^{2} \in \mathbb{R}^{+}$, the prediction error estimate in $cAIC$ is:

$$cAI = -\frac{1}{n}\ell(\boldsymbol{y}) + \frac{\mathrm{tr}(H)}{n}.$$ [1]

[1] Vaida and Blanchard (2005) define this prediction error estimator with a factor of $2n$, i.e., as $2n \times cAI$. In addition, they denote the prediction error estimator as $cAIC$. However, here, in order to distinguish between the prediction error estimator and the model selection procedure, the prediction error estimator is denoted as cAI and the criterion as $cAIC$. Similarly with mAI and $mAIC$.

Given known covariance matrix, V, the prediction error estimate in $mAIC$ is:

$$mAI = -\frac{1}{n}\ell(\boldsymbol{y}) + \frac{p}{n}.$$

$\ell(\boldsymbol{y})$ in cAI is the conditional log-likelihood given the random effects, and $\ell(\boldsymbol{y})$ in mAI is the marginal log-likelihood of $\boldsymbol{y}$ with respect to the random effects. The model selection criteria $cAIC$ and $mAIC$ are defined from cAI and mAI similarly to $tAIC$.

Vaida and Blanchard (2005) extend the cAI result for the case when σ^2 is estimated by MLE, but all the other variance parameters in V are known:

$$cAI_{\sigma^2} = -\frac{1}{n}\ell(\boldsymbol{y}) + \frac{(n-p-1)\,(\mathrm{tr}(H)+1)}{(n-p)\,(n-p-2)} + \frac{p+1}{(n-p)\,(n-p-2)}.$$

A detailed discussion about the case when the variance parameters are unknown can be found in Section 2.3.

It is easy to confirm that when $\{X, Z, R\} = \{X^*, Z^*, R^*\}$, the tAI formula is indeed reduced to the mAI formula when $\mathrm{Cov}(\boldsymbol{y^*}, \boldsymbol{y}) = 0$ and to the cAI formula when $\mathrm{Cov}(\boldsymbol{y^*}, \boldsymbol{y}) \neq 0$.

In addition, when GLS is implemented, an interesting interpretation for the difference between mAI and tAI can be shown. With a little algebra we get:

$$\begin{aligned} C_{tAI}(GLS) =& \frac{p}{n} + \frac{1}{2}\log\left(\frac{|V^*|^{\frac{1}{n^*}}}{|V|^{\frac{1}{n}}}\right) \\ &+ \frac{1}{2}\mathrm{tr}\left\{(X^tV^{-1}X)^{-1}\left(\frac{1}{n^*}X^{*t}V^{*^{-1}}X^* - \frac{1}{n}X^tV^{-1}X\right)\right\} \\ =& \frac{p}{n} + \frac{1}{2}\log\left(\frac{|V^*|^{\frac{1}{n^*}}}{|V|^{\frac{1}{n}}}\right) \\ &+ \frac{1}{2}\mathrm{tr}\left[\mathrm{Var}(\hat{\boldsymbol{\beta}})\left\{\frac{1}{n^*}\mathrm{Var}(\hat{\boldsymbol{\beta}}^{\boldsymbol{*}})^{-1} - \frac{1}{n}\mathrm{Var}(\hat{\boldsymbol{\beta}})^{-1}\right\}\right], \end{aligned}$$

where

$$\hat{\boldsymbol{\beta}}^{\boldsymbol{*}} = (X'^*V^{*^{-1}}X^*)^{-1}X'^*V^{*^{-1}}\boldsymbol{y}^{\boldsymbol{*}}.$$

Since $\mathrm{Var}(\hat{\boldsymbol{\beta}})$ achieves the Cramer-Rao bound:

$$C_{tAI}(GLS) = \frac{p}{n} + \frac{1}{2}\log\left(\frac{|V^*|^{\frac{1}{n^*}}}{|V|^{\frac{1}{n}}}\right) + \frac{1}{2}\mathrm{tr}\left[\mathcal{I}(\hat{\boldsymbol{\beta}})^{-1}\left\{\frac{1}{n^*}\mathcal{I}(\hat{\boldsymbol{\beta}}^{\boldsymbol{*}}) - \frac{1}{n}\mathcal{I}(\hat{\boldsymbol{\beta}})\right\}\right], \tag{8}$$

where $\mathcal{I}$ is Fisher-information. The determinants $|V|$ and $|V^*|$ are often called the generalized variance (Wilks, 1932; Johnson et al., 2014).

2.2. Relaxing Theorem 1 conditions

Although this paper focuses on prediction error estimation and model selection for LMM and GPR, Theorem 1 is more general and does not assume the paradigm applied in LMM and GPR, i.e., predicting using $E(\boldsymbol{y^*}|\boldsymbol{y})$ and estimating the marginal mean parameters with MLE. Theorem 1 assumes:

1. Normality of $\boldsymbol{y^*}$ and $\boldsymbol{y}$.
2. $\mathbb{E}\boldsymbol{y} = X\boldsymbol{\beta}$, $\mathbb{E}\boldsymbol{y^*} = X^*\boldsymbol{\beta}$
3. $H\boldsymbol{\mu} = \boldsymbol{\mu}$, $H^*\boldsymbol{\mu} = \boldsymbol{\mu^*}$

and therefore can be used in other cases satisfying the above conditions.

When the normality assumption cannot be made, another model selection criterion, which is based on similar approach as tAI can be implemented. For more details see Section 4.

Proposition 3 extends Theorem 1 results for the case the normality assumption is valid, however the fitted model does not satisfy condition 3 of unbiasedness:

Proposition 3. *Consider the setting of Theorem 1, without the assumption that $H\boldsymbol{\mu} = \boldsymbol{\mu}$ and $H^*\boldsymbol{\mu} = \boldsymbol{\mu^*}$, then:*

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{y}}\left\{-\frac{1}{n^*}\mathbb{E}_{\boldsymbol{y^*}|\boldsymbol{y}}\ell_{H^*}(\boldsymbol{y^*}) + \frac{1}{n}\ell_H(\boldsymbol{y})\right\} \\
&\qquad = \frac{1}{n}\mathrm{tr}\left(R^{-1}HV\right) - \frac{1}{n^*}\mathrm{tr}\left({R^*}^{-1}H^*\mathrm{Cov}(\boldsymbol{y},\boldsymbol{y^*})\right) \\
&\qquad\quad + \frac{1}{2}\left\{\log\left(\frac{|R^*|^{\frac{1}{n^*}}}{|R|^{\frac{1}{n}}}\right) + \frac{1}{n^*}\mathrm{tr}\left({R^*}^{-1}V^*\right) - \frac{1}{n}\mathrm{tr}\left(R^{-1}V\right)\right\} \\
&\qquad\quad + \frac{1}{2}\left\{\frac{1}{n^*}\mathrm{tr}\left({R^*}^{-1}H^*VH^{*t}\right) - \frac{1}{n}\mathrm{tr}\left(R^{-1}HVH^t\right)\right\} \\
&\qquad\quad + \frac{1}{2n}\mathrm{tr}\left(R^{-1}(2H\boldsymbol{\mu}\boldsymbol{\mu^t} - \boldsymbol{\mu}\boldsymbol{\mu^t} - H\boldsymbol{\mu}\boldsymbol{\mu^t}H^t)\right) \\
&\qquad\quad - \frac{1}{2n^*}\mathrm{tr}\left({R^*}^{-1}(2H^*\boldsymbol{\mu}\boldsymbol{\mu^{*t}} - \boldsymbol{\mu^*}\boldsymbol{\mu^{*t}} - H^*\boldsymbol{\mu}\boldsymbol{\mu^t}H^{*t})\right).
\end{aligned}$$

The proof can be found in Appendix A.1 as part of the proof of Theorem 1.

Note that this expression is less useful, as it depends on $\boldsymbol{\mu}$ and $\boldsymbol{\mu^*}$, the objects that we try to estimate.

2.3. tAI for unknown variance parameters

In most applications the parameters of the covariance matrices, V, V^*, R and R^*, are unknown and therefore are estimated using various methods, such as maximum likelihood and restricted maximum likelihood (Verbeke and Molenberghs, 2009). Using the estimated parameters instead of the true variance parameters can add variance to tAI and even can make tAI biased. This issue, which appears also in cAI and mAI, is complex (Vaida and Blanchard, 2005) and several solutions are proposed in context of cAI and mAI, e.g., by Vaida and Blanchard (2005); Liang, Wu and Zou (2008); Greven and Kneib (2010). In many solutions, some of the variance parameters are assumed to be known and only the remaining subset is considered as an estimate. Some solutions require numerical approximations, asymptotic approximations and consume an intensive computational effort.

As was mentioned in Section 2.1, Vaida and Blanchard (2005) propose cAI_{σ^2} for the case when all the parameters are known besides σ^2. They suggest to use cAI_{σ^2} also in cases when all the variance parameters are unknown. Their suggestion is based on simulations and on an asymptotic analysis of a use case.

Liang, Wu and Zou (2008) propose an unbiased prediction error estimator for the case when σ^2 is known, but other variance parameters are unknown:

$$-\frac{1}{n}\ell(\boldsymbol{y}) + \frac{1}{n}\mathrm{tr}(\frac{\partial\widehat{\boldsymbol{y}^{\boldsymbol{t}}}}{\partial\boldsymbol{y}}), \tag{9}$$

where $\widehat{\boldsymbol{y}}$ is the predictor of $\boldsymbol{y}$. This result is based on Stein's equality (Stein, 1981). Since their method uses derivatives of $\widehat{\boldsymbol{y}}$, which are non-linear expressions due to the variance estimates, implementing this approach requires numerical approximation. Therefore, besides the fact that this approach yields an approximation, the computational effort is substantial (Greven and Kneib, 2010). Greven and Kneib (2010) also derive an estimator of $1/n \times \mathrm{tr}(\partial\widehat{\boldsymbol{y}^{\boldsymbol{t}}}/\partial\boldsymbol{y})$, however their estimator includes derivatives and requires partial prior knowledge on the variance matrix.

In Section 5 we demonstrate numerically the effect of using estimated variance parameters in tAI on its bias and variance, as well as on model selection using $tAIC$, in scenarios when all the variance parameters are unknown. Both simulation and real data analyses support the result by Vaida and Blanchard (2005) – although using the estimated variance parameters adds variance, tAI and $tAIC$ still preform well, especially when the sample size of the training set is not very small.

3. Use cases

In this section, typical use cases of using tAI and $tAIC$ are presented.

3.1. *Predicting interpolation and extrapolation in spatial array and longitudinal temporal data*

As was described in the introduction, predicting interpolated and extrapolated data points using LMM and GPR is common in biomedicine, health, climatology and other research fields, where temporal and spatial datasets are common. The flexible definition of X^*, Z^*, R^* and V^* in tAI makes it applicable when the goal is to estimate prediction error at interpolated and extrapolated data points along time and space dimensions.

In Section 5 we analyze numerically a repeated measures clinical study, containing child growth measurements (Potthoff and Roy, 1964), where interpolation and extrapolation objectives can be defined and application of tAI is demonstrated. The following example, built on the application of Tsanas et al. (2010), demonstrates that appropriate use of $tAIC$ can also simplify and improve on existing methodology.

Example 3.1. *Tsanas et al. (2010) introduced a new method for measuring progression of Parkinson's disease. Their motivation is that the standard methodology for measuring Parkinson's disease progression, which uses UPDRS score (Unified Parkinson's Disease Rating Scale), is costly and requires a physician visit. Their alternative methodology is creating a formula that approximates the UPDRS score with speech signals which are not costly. Six months data was collected for their study, containing large amount of longitudinal speech signal measurements per patient, however, UPDRS scores were collected only at a small number of the time points. In order to select the best covariates with respect to the whole speech signals sample, they suggested to interpolate the UPDRS scores using "straightforward linear interpolation", then to fit several alternative models and to select one of them using AIC and other model selection criteria. An alternative paradigm that does not require imputing UPDRS scores is by using* $tAIC$*. Since* $tAIC$ *does not assume* $\{X^*, Z^*\} = \{X, Z\}$*, there is no need in interpolating the UPDRS scores in order to select a model minimizing the estimated prediction error with respect to the whole speech signals sample.*

We note that in Example 3.1, one may think that $\boldsymbol{y}^*$ is used twice, for model building and for prediction error estimation, and therefore over-fitting can occur. However, since in tAI approach, unlike in cross-validation approach, $\boldsymbol{y}^*$ is used as a conceptual idea in order to derive C_{tAI} and not as real observations, $\boldsymbol{y}^*$ is not used twice.

In the spatial data analysis domain, commonly, studies include geographical data (Li and Heap, 2014) and neuroimaging data (Salimi-Khorshidi et al., 2011). Such studies usually use GPR rather than LMM. Although GPR and LMM reflect different perspectives – while GPR is based on functional data analysis approach, LMM is based on multivariate analysis approach – and use different techniques for expressing the covariance matrices, both models use conditional expectation for prediction, hence tAI is also applicable for GPR. A use case of creating high-resolution climate maps given by Brown and Comrie (2002), which was mentioned in the introduction, is an example for the importance of tAI in assessing the prediction error at interpolation and extrapolation points. A similar use case, containing chemical concentration in soil data, is analyzed numerically in Section 5.

3.2. Other transductive settings

LMM and GPR are also used for modeling data without spatial or temporal correlation structure. One interesting example is modeling the effect of SNPs

(Single Nucleotide Polymorphism) on a phenotype as part of a Genome-Wide Association Study (GWAS). In this case the common practice is to consider the SNPs as random effects and other explanatory variables (e.g., age, height and gender) as fixed effects (Zhang et al., 2010). When using LMM for modeling the effect of SNPs on phenotype, tAI allows estimating the prediction error for an extended population compared to the training sample. It is directly useful in the important case when $\{X^*, Z^*\}$ can be collected from other studies which investigate different phenotype, however contain the SNPs and the explanatory variables that are used in the training sample (Wray et al., 2013).

tAI may also be required when there are missing values. Missing values of the dependent variable is a common phenomenon in statistical analysis and in particular in clinical trials with repeated measures study designs (Wood, White and Thompson, 2004; O'neill and Temple, 2012). There are many methods for handling missing values in repeated measures studies, some of the methods involving missing values imputation (Mallinckrodt et al., 2003). In case of having missing data of the dependent variable at some known points but the goal is to estimate the prediction error with respect to the original study design (Hogan, Roy and Korkontzelou, 2004), tAI can be used without imputing the missing values.

4. Optimism for prediction at interpolation and extrapolation points

The formulation of tAI and the derivation of C_{tAI} are based on the normality assumptions of $\boldsymbol{y}$ and $\boldsymbol{y^*}$, which is commonly assumed when LMM and GPR are implemented. However, the approach that is used for developing tAI can be used for creating other prediction error estimators that are not based on the normality assumption of $\boldsymbol{y^*}$ and $\boldsymbol{y}$. For example, in the standard formulation of the prediction error estimator that is based on expected optimism correction (Efron, 1986),

$$Loss(Opt) = \frac{1}{n}\|\boldsymbol{y} - H\boldsymbol{y}\|_2^2 + w,$$

where

$$w = \mathbb{E}_{\boldsymbol{y}}\left(\frac{1}{n}\mathbb{E}_{\boldsymbol{y^*}|\boldsymbol{y}}\|\boldsymbol{y^*} - H\boldsymbol{y}\|_2^2 - \frac{1}{n}\|\boldsymbol{y} - H\boldsymbol{y}\|_2^2\right),$$

it is assumed that $\boldsymbol{y^*}$ and $\boldsymbol{y}$ are drawn from the same distribution and have the same predictor, $H\boldsymbol{y}$. However, as was already discussed in the previous sections, these conditions are not satisfied in many use cases. The following prediction error generalizes $Loss(Opt)$:

$$Loss(Opt_t) = \frac{1}{n}\|\boldsymbol{y} - H\boldsymbol{y}\|_2^2 + w_t,$$

where

$$w_t = \mathbb{E}_{\boldsymbol{y}}\left(\frac{1}{n^*}\mathbb{E}_{\boldsymbol{y^*}|\boldsymbol{y}}\|\boldsymbol{y^*} - H^*\boldsymbol{y}\|_2^2 - \frac{1}{n}\|\boldsymbol{y} - H\boldsymbol{y}\|_2^2\right).$$

Similarly to $tAIC$ definition, given a set of models $\mathcal{H}$, $Loss(Opt_t)$ can be used for model selection as follows:

$$h_{best} = \underset{h \in \mathcal{H}}{\operatorname{argmin}} \, Loss_h(Opt_t), \tag{10}$$

where $Loss_h(Opt_t)$ is $Loss(Opt_t)$ for model h.

For this section, assume that $\boldsymbol{y} \in \mathbb{R}^n$ is a random variable with mean $\boldsymbol{\mu}$ and variance V. Similarly, let $\boldsymbol{y^*} \in \mathbb{R}^{n^*}$ be a random variable with mean $\boldsymbol{\mu^*}$ and variance V^*. In addition, let $H\boldsymbol{y} \in R^n$ and $H^*\boldsymbol{y} \in R^{n^*}$ be the predictors of $\boldsymbol{y}$ and $\boldsymbol{y^*}$ respectively, when H and H^* do not contain $\boldsymbol{y}$ and $\boldsymbol{y^*}$. Theorem 4 introduces a general expression of w_t for predictors that are linear in $\boldsymbol{y}$.

Theorem 4.

$$\begin{aligned} w_t =& \frac{2}{n}\operatorname{tr}(HV) - \frac{2}{n^*}\operatorname{tr}\left(H^*\operatorname{Cov}(\boldsymbol{y}, \boldsymbol{y^*})\right) \\ &+ \frac{1}{n^*}\operatorname{tr}(V^*) - \frac{1}{n}\operatorname{tr}(V) + \frac{1}{n^*}\operatorname{tr}\left(H^*VH^{*t}\right) - \frac{1}{n}\operatorname{tr}\left(HVH^t\right) \\ &+ \frac{1}{n}\operatorname{tr}\left(2H\boldsymbol{\mu}\boldsymbol{\mu}^{\boldsymbol{t}} - \boldsymbol{\mu}\boldsymbol{\mu}^{\boldsymbol{t}} - H\boldsymbol{\mu}\boldsymbol{\mu}^{\boldsymbol{t}}H^t\right) \\ &- \frac{1}{n^*}\operatorname{tr}\left(2H^*\boldsymbol{\mu}\boldsymbol{\mu}^{\boldsymbol{*t}} - \boldsymbol{\mu}^{\boldsymbol{*}}\boldsymbol{\mu}^{\boldsymbol{*t}} - H^*\boldsymbol{\mu}\boldsymbol{\mu}^{\boldsymbol{t}}H^{*t}\right). \end{aligned}$$

Corollary 5. *Using Theorem 4, when $H\boldsymbol{\mu} = \boldsymbol{\mu}$ and $H^*\boldsymbol{\mu} = \boldsymbol{\mu^*}$*

$$\begin{aligned} w_t =& \frac{2}{n}\operatorname{tr}(HV) - \frac{2}{n^*}\operatorname{tr}\left(H^*\operatorname{Cov}(\boldsymbol{y}, \boldsymbol{y^*})\right) \\ &+ \frac{1}{n^*}\operatorname{tr}(V^*) - \frac{1}{n}\operatorname{tr}(V) + \frac{1}{n^*}\operatorname{tr}\left(H^*VH^{*t}\right) - \frac{1}{n}\operatorname{tr}\left(HVH^t\right). \end{aligned}$$

In case $H^* = H$, $V^* = V$ and $V - \operatorname{Cov}(\boldsymbol{y}, \boldsymbol{y^*}) = \sigma^2 I_n$, then

$$w_t = w = \frac{2\sigma^2}{n}\operatorname{tr}(H),$$

which is the same result as was introduced by Hodges and Sargent (2001) for linear hierarchical models. Numerical results that compare between $Loss(Opt)$ and $Loss(Opt_t)$ are presented in Section 5.

$Loss(Opt_t)$ is based on the squared error loss function which reflects Euclidean distance. Other prediction error estimators which are based on different distances, such as on Mahalanobis distance (Mahalanobis, 1936), might be suggested as well. Corollary 6 presents a penalty correction for a prediction error estimator which is based on Mahalanobis distance.

Corollary 6. *Given the definitions in Theorem 4, when $H\boldsymbol{\mu} = \boldsymbol{\mu}$ and $H^*\boldsymbol{\mu} = \boldsymbol{\mu^*}$*

$$\begin{aligned} &\mathbb{E}_{\boldsymbol{y}}\left(\frac{1}{n^*}\mathbb{E}_{\boldsymbol{y^*}|\boldsymbol{y}}\|\boldsymbol{y^*} - H^*\boldsymbol{y}\|_M^2 - \frac{1}{n}\|\boldsymbol{y} - H\boldsymbol{y}\|_M^2\right) \\ &\qquad = \frac{2}{n}\operatorname{tr}\left(R^{-1}HV\right) - \frac{2}{n^*}\operatorname{tr}\left(R^{*^{-1}}H^*\operatorname{Cov}(\boldsymbol{y}, \boldsymbol{y^*})\right) \end{aligned} \tag{11}$$

$$\begin{aligned}
&+\frac{1}{n^*}\mathrm{tr}\left(R^{*^{-1}}V^*\right) - \frac{1}{n}\mathrm{tr}\left(R^{-1}V\right)\\
&+\frac{1}{n^*}\mathrm{tr}\left(R^{*^{-1}}H^*VH^{*t}\right) - \frac{1}{n}\mathrm{tr}\left(R^{-1}HVH^t\right)\\
=\ &2C_{tAI} - \log\left(\frac{|R^*|^{\frac{1}{n^*}}}{|R|^{\frac{1}{n}}}\right),
\end{aligned}$$

where

$$\begin{aligned}
\|\boldsymbol{y^*} - H^*\boldsymbol{y}\|_M^2 &= (\boldsymbol{y^*} - H^*\boldsymbol{y})^t R^{*-1}(\boldsymbol{y^*} - H^*\boldsymbol{y})\\
\|\boldsymbol{y} - H\boldsymbol{y}\|_M^2 &= (\boldsymbol{y} - H\boldsymbol{y})^t R^{-1}(\boldsymbol{y} - H\boldsymbol{y}).
\end{aligned}$$

The relation between eq. (11) and C_{tAI} arises due to the relation between Mahalanobis distance and the normal likelihood which tAI is based on.

It is natural to use $Loss(Opt_t)$ instead of tAI for linear predictors that do not assume normality, such as the predictors that are used in nearest neighbors, Nadaraya-Watson kernel regression and smoothing spline models. Moreover, due to the form of the normal density function, many predictors that seem to be based on the normality assumption can be alternatively interpreted as a solution of a least squares problem or complex versions of least squares problems like weighed least squares and penalized least squares problems. For example, GLS can be interpreted as the solution of weighted least squares problem with the weight matrix V^{-1}. Similarly, $\hat{\boldsymbol{f}}^{\boldsymbol{*}}$ can be interpreted as an estimator of

$$\tilde{\boldsymbol{a}} + \tilde{B}\boldsymbol{y}$$

where

$$\{\tilde{\boldsymbol{a}}, \tilde{B}\} = \underset{\boldsymbol{a}\in\mathbb{R}^{n^*}, B\in\mathbb{R}^{n^*\times n}}{\operatorname{argmin}} \mathbb{E}_{\boldsymbol{y^*},\boldsymbol{y}}\|\boldsymbol{y^*} - (\boldsymbol{a} + B\boldsymbol{y})\|_2^2.$$

The proof is attached in Appendix A.2.

These alternative interpretations are free from normality assumption and therefore $Loss(Opt_t)$ can be suitable for them. Since many predictors can be interpreted in different ways, then the assignation of predictors to tAI or to $Loss(Opt_t)$ should refer to the possibility to assume normality rather than to the predictor type.

5. Numerical results

This section focuses on comparison between the prediction error estimators that were mentioned in the previous sections, as well as between their corresponding model selection criteria using simulation and real data analyses. Relevant R code can be found at https://github.com/AssafRab/Prediction-Error-Interpolation-Extrapolation.

5.1. Simulation analyses

The goal of the following analyses is to investigate the accuracy of tAI, cAI and mAI in estimating $-\mathbb{E}_{\boldsymbol{y^*}|\boldsymbol{y}}\ell(\boldsymbol{y^*})/n^*$, for different sample sizes and variance setups. In addition, $tAIC$, $cAIC$ and $mAIC$ will also be analyzed and compared with respect to the oracle solution

$$h_{best} = \underset{h\in\mathcal{H}}{\operatorname{argmin}} -\frac{1}{n^*}\mathbb{E}_{\boldsymbol{y}^*|\boldsymbol{y}}\ell_h(\boldsymbol{y}^*).$$

Additional numerical results with respect to a potentially different oracle solution

$$h_{best} = \underset{h\in\mathcal{H}}{\operatorname{argmin}} -\frac{1}{n^*}\mathbb{E}_{\boldsymbol{y}}\mathbb{E}_{\boldsymbol{y}^*|\boldsymbol{y}}\ell_h(\boldsymbol{y}^*),$$

and with respect to

$$h_{best} = \underset{h\in\mathcal{H}}{\operatorname{argmin}} -\frac{1}{n^*}\mathbb{E}_{\boldsymbol{y}^*|\boldsymbol{y}}\|\boldsymbol{y}^* - H^*\boldsymbol{y}\|_2^2,$$

will be also presented right afterwards.

Simulation setup The simulation demonstrates prediction error estimation and model selection for the following LMM setting:

$$\phi_{i,j} = 0.5\times time_{i,j} + \sum_{k=0}^{k=2} x_{i,j,k} + 2\times\sum_{k=3}^{k=6} x_{i,j,k} + b_{i,1} + time_{i,j}\times b_{i,2} + \epsilon_{i,j},$$

where $i \in \{1,...,S\}$ is the subject number and $j \in \{1,...,12\}$ is the measurement number.

- $time_{i,j} = j$, $\forall j \leq 10$, $time_{i,11} = 15$, $time_{i,12} = 20$,
- $x_{i,j,0} = 1$, $\forall i \in \{1,..S\}$, $j \in \{1,...,12\}$,
- $x_{i,j,1}$ were drawn independently from Bernoulli distribution with 0.5 probability of success $\forall i \in \{1,..S\}$, $j \in \{1,...,12\}$,
- $x_{i,j,k}$, $\forall k \in \{2,...,6\}$, were drawn independently form standard normal distribution $\forall i \in \{1,..S\}$, $j \in \{1,...,12\}$,
- $b_{i,1}$ and $b_{i,2}$ were drawn independently from normal distribution with zero mean, $\text{Var}(b_{i,1}) = 15$, $\text{Var}(b_{i,2}) = 1$ and $\text{Cov}(b_{i,1}, b_{i,2}) = 0$ $\forall i \in \{1,...,S\}$. Denote $\boldsymbol{\sigma_b} = [\text{Var}(b_{i,1}), \text{Var}(b_{i,2})]$.
- $\epsilon_{i,j}$ were drawn independently from normal distribution with zero mean and variance σ^2 (for different σ^2 values in different simulation setups) $\forall i \in \{1,..S\}$, $j \in \{1,...,12\}$,
- $x_{i,j,k}$, $b_{i,1}$, $b_{i,2}$ and $\epsilon_{i,j}$ are independent $\forall i \in \{1,..S\}$, $j \in \{1,...,12\}$, $k \in \{1,...,6\}$.

The dependent variable in the training set ,$\boldsymbol{y}$, was defined as $\phi_{i,j}\,\forall j \in \{1,..,10\}$, the dependent variable in the prediction set, $\boldsymbol{y}^*$, was defined as $\phi_{i,j}\,\forall j \in \{11,12\}$. Therefore, this setting demonstrates predicting at extrapolation time points. This setting was generated nine times, for different number of subjects, $S \in \{20,100,200\}$, and different residual variance values, $\sigma^2 \in \{15,20,25\}$. mAI, cAI and tAI were calculated and $-\mathbb{E}_{\boldsymbol{y}^*|\boldsymbol{y}}\ell_h(\boldsymbol{y}^*)/n^*$ was approximated using repeated sampling of $\boldsymbol{y^*}$. The simulation was repeated 200 times for creating an approximation of the density of $-\mathbb{E}_{\boldsymbol{y^*}|\boldsymbol{y}}\ell_h(\boldsymbol{y^*})/n^*$ and for calculating its average – the approximation of $-\mathbb{E}_{\boldsymbol{y}}\mathbb{E}_{\boldsymbol{y^*}|\boldsymbol{y}}\ell_h(\boldsymbol{y^*})/n^*$.

Three linear mixed models were fitted given the true covariance matrices, all the models contain the time covariate, in addition, Model 1 contains $x_{i,j,k}$, $\forall k \leq 2$, Model 2 contains $x_{i,j,k}$, $\forall k \leq 4$ and Model 3 contains $x_{i,j,k}$, $\forall k$ which is also the model that generated the data.

Another simulation analysis which demonstrates the performance of tAI in high-dimensional data (150 variables) is presented in Appendix B.2.

Results Figure 1 presents the densities of tAI, cAI, mAI and $-\mathbb{E}_{\boldsymbol{y^*}|\boldsymbol{y}}\ell(\boldsymbol{y^*})/n^*$ for Model 3, when $S = 100$ and $\sigma^2 = 20$, as generated from the 200 simulation runs. Two versions of tAI, cAI and mAI are presented – when the parameters of the covariance matrices, σ^2 and $\boldsymbol{\sigma_b}$, are known, and when they are unknown.

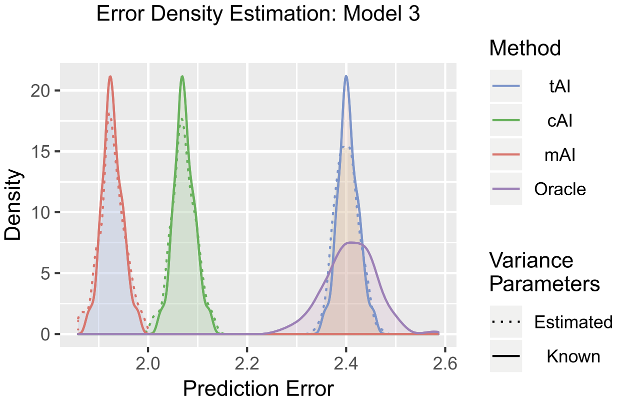


Fig 1. *Densities of* tAI, cAI, mAI *and* $-\mathbb{E}_{\boldsymbol{y^*}|\boldsymbol{y}}\ell(\boldsymbol{y^*})/n^*$ *where number of subjects*$= 100$ *and* $\sigma^2 = 20$. *Two scenarios are presented: when the variance parameters are known and when they are estimated.*

As can be seen from Figure 1, tAI density is concentrated around the mean of $-\mathbb{E}_{\boldsymbol{y^*}|\boldsymbol{y}}\ell(\boldsymbol{y^*})/n^*$. mAI and cAI are stochastically smaller than $-\mathbb{E}_{\boldsymbol{y^*}|\boldsymbol{y}}\ell(\boldsymbol{y^*})/n^*$ since their corrections, $\mathrm{tr}(H)/n$ and p/n, are unsuitable for this case of predicting at extrapolation points. Also, the densities of tAI, cAI and mAI, which are based on estimated variance parameters (dotted lines in the plot) have a larger variance than the densities that are based on known variance parameters, however, they are still very similar.

In addition, for the versions with the known variance parameters, since tAI, cAI and mAI share the same random part, $\ell(\boldsymbol{y})/n$, but different mean, $-\mathbb{E}_{\boldsymbol{y}}\ell(\boldsymbol{y})/n$ plus C_{tAI}, $\mathrm{tr}(H)/n$, p/n respectively, their densities have the same shape, however shifted with respect to the corrections. In contrast, $-\mathbb{E}_{\boldsymbol{y^*}|\boldsymbol{y}}\ell(\boldsymbol{y^*})/n^*$ has the same mean as tAI but different variance, since $\mathrm{Var}\left(-\mathbb{E}_{\boldsymbol{y^*}|\boldsymbol{y}}\ell(\boldsymbol{y^*})/n^*\right)$

depends on H^*, R^* and n^* that do not appear in $\mathrm{Var}\,(tAI) = \mathrm{Var}\,(-\ell(\boldsymbol{y})/n)$. In our case, H^* contains large values compared to H and therefore

$$\mathrm{Var}\left(\mathbb{E}_{\boldsymbol{y}^*|\boldsymbol{y}} - \frac{1}{n^*}\ell(\boldsymbol{y}^*)\right) > \mathrm{Var}\,(tAI)\,.$$

In order to asses the performance of tAI version with the estimated variance parameters, compared to the tAI version with the known variance parameters, a two sample Anderson-Darling test (Anderson and Darling, 1952) was used. The tested statistic is:

$$tAI - \left(-\frac{1}{n^*}\mathbb{E}_{\boldsymbol{y}^*|\boldsymbol{y}}\ell(\boldsymbol{y}^*)\right),$$

where one sample uses tAI version with the estimated variance parameters, and the other sample uses tAI version with the true variance parameters. Implementing the function ad.test of the package kSamples in R software, the p-value of the test is 0.9754. The result indicates that in this setting there is no evidence for significant difference between the distribution of $tAI - \left(-\mathbb{E}_{\boldsymbol{y}^*|\boldsymbol{y}}\ell(\boldsymbol{y}^*)/n^*\right)$ when the variance parameters are known in advance or estimated.

Figure 2 presents the same graph as Figure 1, but for different settings of S and σ^2.

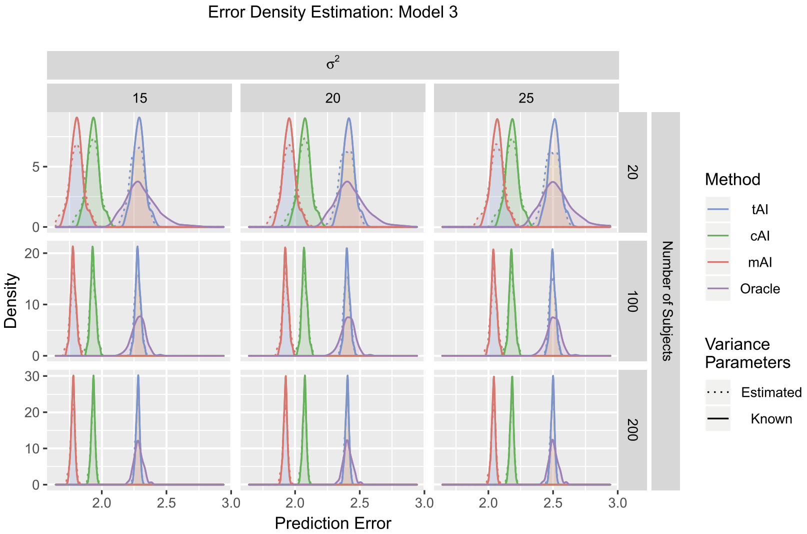


Fig 2. *Densities of tAI, cAI, mAI and $-\mathbb{E}_{\boldsymbol{y}^*|\boldsymbol{y}}\ell(\boldsymbol{y}^*)/n^*$ for different settings. Two scenarios are presented: when the variance parameters are known and when they are estimated.*

As can be seen in Figure 2, even for relatively small sample size and large variance, the versions with the estimated variance parameters are similar to those with the known variance parameters. Also, as expected, as much as the sample size increases the variance of the error decreases.

For demonstrating the model selection performance, Figure 3 presents the agreement rate between each of $tAIC$, $cAIC$ and $mAIC$ and the oracle over the repeated simulation runs. Both versions of $tAIC$, $cAIC$ and $mAIC$, are analyzed – when the variance parameters are known and when they are estimated.

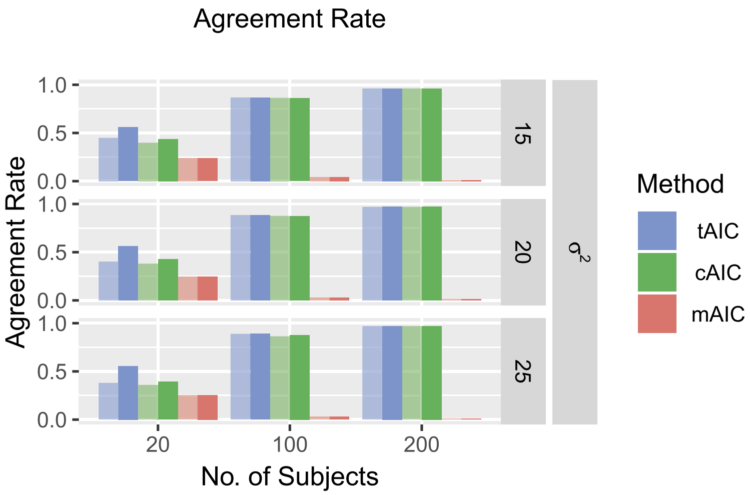


FIG 3. *For the different nine setups, each bar refers to the agreement rates of mAIC, cAIC and tAIC, with the oracle criterion. Also, two scenarios are presented: when the variance parameters are known (in dark color) and when they are estimated (in light color).*

As can be seen from Figure 3, $tAIC$ performs better than $mAIC$ and $cAIC$. For large sample sizes, $tAIC$ and $cAIC$ both perform well. For small sample size (number of subjects = 20), the $tAIC$ version that is based on the known variance parameters outperforms the $tAIC$ version with estimated variance parameters. This demonstrates the expected difficulty in estimating these parameters for small sample size.

Figure 4 presents similar analyses with respect to the prediction errors

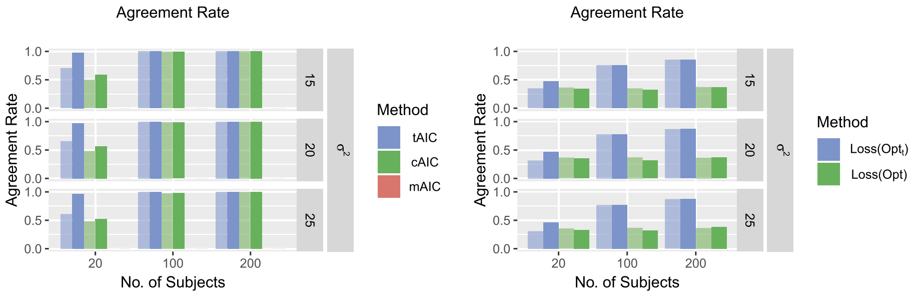
$$\mathbb{E}_{\boldsymbol{y}}\mathbb{E}_{\boldsymbol{y}^*|\boldsymbol{y}} - \frac{1}{n^*}\ell_{h_{best}}(\boldsymbol{y}^*)$$

and

$$\frac{1}{n^*}\mathbb{E}_{\boldsymbol{y}^*|\boldsymbol{y}}\|\boldsymbol{y}^* - H^*\boldsymbol{y}\|_2^2.$$

We see that the general picture – the advantage of $tAIC$ over the other criteria in model selection, and the small effect of variance parameter estimation for large data, compared to the bigger effect for smaller data – is preserved

Another type of model selection performance analysis is presented in Appendix B.1.



(a) (b)

Fig 4. *Figure (4a) presents the agreement rates of mAIC, cAIC and tAIC with* $\underset{h\in\mathcal{H}}{\text{argmin}}\mathbb{E}_{\boldsymbol{y}}\mathbb{E}_{\boldsymbol{y^*}|\boldsymbol{y}} - \frac{1}{n^*}\ell_{h_{best}}(\boldsymbol{y^*})$. *Figure (4b) presents the agreement rates of* $\underset{h\in\mathcal{H}}{\text{argmin}} Loss_h(Opt_t)$ *and* $\underset{h\in\mathcal{H}}{\text{argmin}} Loss_h(Opt)$ *with* $\underset{h\in\mathcal{H}}{\text{argmin}} \frac{1}{n^*}\mathbb{E}_{\boldsymbol{y^*}|\boldsymbol{y}}\|\boldsymbol{y^*} - H^*\boldsymbol{y}\|_2^2$. *In each sub-figure different nine setups are presented. In each setup, each bar refers to the agreement rate of the relevant criterion with the oracle criterion. Also, two scenarios are presented: when the variance parameters are known (in dark color) and when they are estimated (in light color).*

5.2. Real data analyses

The analyses below focus on comparison between tAI, cAI, mAI and

$$-\frac{1}{n^*}\ell(\boldsymbol{y^*}).$$

Here, $-\ell(\boldsymbol{y^*})/n^*$ is used as a ground truth instead of $-\mathbb{E}_{\boldsymbol{y^*}|\boldsymbol{y}}\ell(\boldsymbol{y^*})/n^*$ since the latter is unknown for the real data sets. Also, only tAI, cAI and mAI with the estimated variance parameters version are presented (since the true variance matrices are unknown).

5.2.1. Meuse data

Data description Meuse data set was introduced by Rikken and Van Rijn (1993) and is available in sp package in R software. The data was collected in a floodplain area of the river Meuse, near the village of Stein, Netherlands, and contains 155 measurements of topsoil concentrations of Zinc, Lead, Copper and Cadmium, along with location (latitude and longitude) and other covariates. In addition, another data set, Meuse.grid, is analyzed. Meuse.grid is a higher resolution grid of the same area, containing 3103 observations of location and some of the covariates that are available in the Meuse data set, however it does not contain the metal concentration measurements. The Meuse.grid is available in sp package in R software as well.

Results The Meuse data set was partitioned randomly into training and test samples. Four Gaussian process regression models were fitted to the log of the

Lead concentration.[2] All the models share the same kernel structure, squared-exponential kernel,

$$K(\boldsymbol{Z_i}, \boldsymbol{Z_j}) = \sigma_f^2 \exp\left[-\frac{1}{2}\left\{\frac{1}{l_1^2}\left(Z_{i,1} - Z_{j,1}\right)^2 + \frac{1}{l_2^2}\left(Z_{i,2} - Z_{j,2}\right)^2\right\}\right],$$

where $Z_{i,1}$ refers to the latitude of measurement i, $Z_{i,2}$ refers to longitude of measurement i and l_1, l_2 and σ_f lie in $\mathbb{R}^+$. Each model has a different marginal mean, see Table 1. The descriptions of the covariates can be found in sp package in R software.

Table 1
Meuse data: Covariates

	Covariates		
Model	*Intercept, dist, ffreq, soil*	$dist \times ffreq$	$dist \times Soil$
1	✓		
2	✓	✓	
3	✓		✓
4	✓	✓	✓

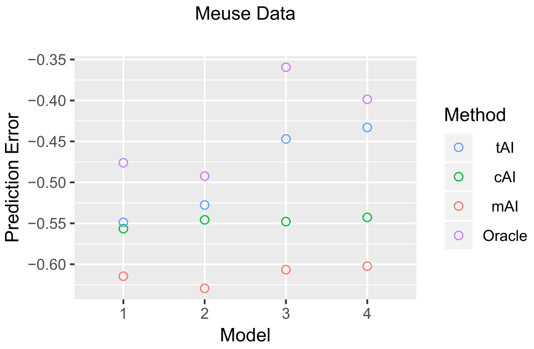


Fig 5. *Estimating prediction error of four predictive spatial models that were fitted to topsoil metal concentration in a floodplain area of the river Meuse (Meuse data). For each model, each color refers to a different prediction error estimate. The oracle is presented as well.*

As can be seen in Figure 5, tAI estimates $-\ell(\boldsymbol{y^*})/n^*$ most accurately. The other prediction error estimators consistently under estimate $-\ell(\boldsymbol{y^*})/n^*$.

Figure 6 is based on Meuse and on Meuse.grid data sets, where the whole Meuse data set is used as training data and the Meuse.grid data set is used as the prediction set, $\{X^*, Z^*\}$. Since the Lead consternation is not given in the Meuse.grid data set, then $-\ell(\boldsymbol{y^*})/n^*$ is unknown. Therefore tAI, cAI and mAI are compared without having a ground truth.

As can be seen from Figure 6, the differences between the mAI, cAI and tAI are sustained and the results are consistent with the previous figures, i.e., mAI, cAI give lower error estimates, which likely underestimate the prediction error.

[2] Only log(*Lead*) can be analyzed under the normality assumption.

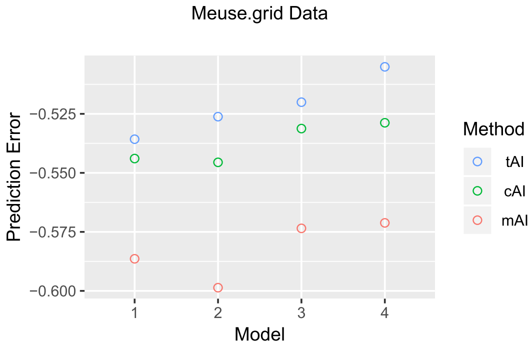


Fig 6. *Estimating prediction error of four predictive spatial models that were fitted to topsoil metal concentration in a floodplain area of the river Meuse (Meuse data). For each model, each color refers to a different prediction error estimate of another data set – Meuse.grid data, which contains the covariates, but not the dependent variable.*

5.2.2. Growth data

Data description The Growth data was introduced by Potthoff and Roy (1964) and contains four skull length measurements for 27 children at ages 8, 10, 12 and 14 (total of 27 × 4 measurements) along with the child's age and gender.

Results Figure 7 presents a scenario where the training sample is defined as the skull length measurements at ages 8, 10, 12 and the prediction set is defined as the skull length measurements at age 14. Three linear mixed models are fitted, all have the same variance structure, containing random intercept per child and random slope for the child's age, however each model has a different set of fixed effects (see Table 2).

Table 2
Growth data: Covariates

	Covariates			
Model	*Intercept*	*Age*	*Gender*	*Age* × *Gender*
1	✓	✓		
2	✓	✓	✓	
3	✓	✓	✓	✓

As can be seen in Figure 7, in general perspective, tAI estimates $-\ell(\boldsymbol{y^*})/n^*$ most accurately. The other prediction error estimators under-estimate $-\ell(\boldsymbol{y^*})/n^*$.

Figure 8 presents three similar analyses as in Figure 7, however where the other time-points measurements are designated as holdout.

When $age = 8$, the results are similar to the results in Figure 7, however, when $age = 10$ and $age = 12$, tAI and cAI have similar performance. This is not surprising since in these cases $\{X^*, Z^*, R^*\}$ is similar to $\{X, Z, R\}$.

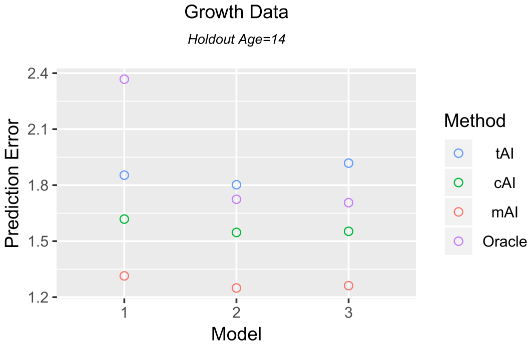


Fig 7. *Estimating prediction error of three predictive longitudinal models that were fitted to repeated skull length measurements of* 27 *children at ages* 8, 10 *and* 12, *but predict the skull length measurements of the same children at age* 14. *For each model, each color refers to a different prediction error estimate. The oracle is presented as well.*

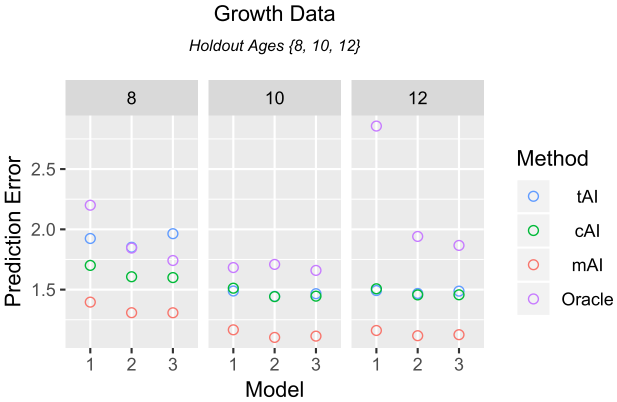


Fig 8. *Estimating prediction error of three predictive longitudinal models that were fitted to repeated skull length measurements of* 27 *children. In each column, measurements that refer to a different age were defined as the prediction set. For each model, each color refers to a different prediction error estimate. The oracle is presented as well.*

6. Discussion and conclusions

tAI is an extension of the prediction error estimators used in $mAIC$ and $cAIC$, extending them to estimate prediction error at interpolation and extrapolation points. As demonstrated in Section 3, these use cases are common in various research fields, and particularly in geostatistics and health, where GPR and LMM are used for predicting at interpolation and extrapolation points. Since GLS, linear regression and smoothing splines can be expressed as LMM (Brumback, Ruppert and Wand, 1999), tAI is applicable for them as well.

The correction in tAI is more complicated than the corrections in $mAIC$ and $cAIC$, which are p/n and $\mathrm{tr}(H)/n$ respectively. The correction in tAI is affected by the relations between $\mathrm{Var}(\boldsymbol{y})$ to $\mathrm{Var}(\boldsymbol{y^*})$, $\mathrm{Var}(\hat{\boldsymbol{f}})$ to $\mathrm{Var}(\hat{\boldsymbol{f}}^{\boldsymbol{*}})$ and between $\mathrm{Cov}(\boldsymbol{y}, \boldsymbol{y^*})$ to $\mathrm{Cov}(\boldsymbol{y}, \boldsymbol{y^{new}})$. When interpreting the correction as a measure of over-fitting, the differences between the corrections offer a new perspective about how the over-fitting is composed as a function of the variance structure of the problem.

In many cases the parameters of the covariance matrices are unknown in advance and therefore are estimated by various procedures prior to model fitting, e.g., REML in LMM (Verbeke and Molenberghs, 2009). Estimating the variance parameters implies an extra variation for tAI, especially when the sample size is small. Our empirical results demonstrate, however, that it remains useful in this setting.

The numerical analyses emphasize the practical importance in using tAI in scenarios where $\{X^*, Z^*, R^*\} \neq \{X, Z, R\}$ are different. It is noticeable especially when predicting at extrapolation points, since in this case the differences between $\mathrm{Var}(\boldsymbol{y})$ to $\mathrm{Var}(\boldsymbol{y^*})$ and between $\mathrm{Var}(\hat{\boldsymbol{f}})$ to $\mathrm{Var}(\hat{\boldsymbol{f}}^{\boldsymbol{*}})$ can be large.

$Loss(Opt_t)$ is another prediction error estimator for cases involving predicting at interpolation and extrapolation points. Unlike tAI, $Loss(Opt_t)$ does not assume that the observations are normally distributed and therefore it is also applicable in various non-parametric applications. Since many predictors that are apparently based on normal linear model can be alternatively interpreted as solutions for the generalized least squares problems, the assignation of predictors to tAI or to $Loss(Opt_t)$ should refer to the possibility to assume normality rather than to the predictor formula.

Appendix A: Proofs

A.1. Proof of Theorem 1

Proof. By the definitions of $\ell_H(\boldsymbol{y})$ and $\ell_{H^*}(\boldsymbol{y^*})$, the bias of $\left\{-\frac{1}{n}\ell_H(\boldsymbol{y})\right\}$ is

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{y}}\left[-\frac{1}{n^*}\mathbb{E}_{\boldsymbol{y^*}|\boldsymbol{y}}\ell_{H^*}(\boldsymbol{y^*}) - \left\{-\frac{1}{n}\ell_H(\boldsymbol{y})\right\}\right] \\
&\qquad = -\frac{1}{n^*}\mathbb{E}_{\boldsymbol{y},\boldsymbol{y^*}}\ell_{H^*}(\boldsymbol{y^*}) + \frac{1}{n}\mathbb{E}_{\boldsymbol{y}}\ell_H(\boldsymbol{y}) \\
&\qquad = \frac{1}{2n^*}\left\{\log|R^*| + n^*\log(2\pi) + \mathbb{E}_{\boldsymbol{y},\boldsymbol{y^*}}(\boldsymbol{y^*} - H^*\boldsymbol{y})^t {R^*}^{-1}(\boldsymbol{y^*} - H^*\boldsymbol{y})\right\} \\
&\qquad\quad - \frac{1}{2n}\left\{\log|R| + n\log(2\pi) + \mathbb{E}_{\boldsymbol{y}}(\boldsymbol{y} - H\boldsymbol{y})^t R^{-1}(\boldsymbol{y} - H\boldsymbol{y})\right\} \\
&\qquad = \frac{1}{2}\log\left(\frac{|R^*|^{\frac{1}{n^*}}}{|R|^{\frac{1}{n}}}\right) \qquad\qquad (12) \\
&\qquad\quad + \frac{1}{2}\left\{\frac{1}{n^*}\mathbb{E}_{\boldsymbol{y},\boldsymbol{y^*}}(\boldsymbol{y^*} - H^*\boldsymbol{y})^t {R^*}^{-1}(\boldsymbol{y^*} - H^*\boldsymbol{y})\right. \\
&\qquad\quad \left. - \frac{1}{n}\mathbb{E}_{\boldsymbol{y}}(\boldsymbol{y} - H\boldsymbol{y})^t R^{-1}(\boldsymbol{y} - H\boldsymbol{y})\right\}.
\end{aligned}
$$

First, let us simplify the last two terms of equation (12):

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{y},\boldsymbol{y^*}}(\boldsymbol{y^*} - H^*\boldsymbol{y})^t R^{*^{-1}}(\boldsymbol{y^*} - H^*\boldsymbol{y}) \\
&\qquad = \operatorname{tr}\left\{R^{*^{-1}}\mathbb{E}_{\boldsymbol{y^*}}(\boldsymbol{y^*}\boldsymbol{y^{*t}})\right\} + \operatorname{tr}\left\{H^{*t}R^{*^{-1}}H^*\mathbb{E}_{\boldsymbol{y}}(\boldsymbol{y}\boldsymbol{y^t})\right\} \\
&\qquad\quad - 2\operatorname{tr}\left\{R^{*^{-1}}H^*\mathbb{E}_{\boldsymbol{y},\boldsymbol{y^*}}(\boldsymbol{y}\boldsymbol{y^{*t}})\right\} \\
&\qquad = \operatorname{tr}\left\{R^{*^{-1}}(V^* + \boldsymbol{\mu^*}\boldsymbol{\mu^{*t}})\right\} + \operatorname{tr}\left\{R^{*^{-1}}H^*(V + \boldsymbol{\mu}\boldsymbol{\mu^t})H^{*t}\right\} \\
&\qquad\quad - 2\operatorname{tr}\left\{R^{*^{-1}}H^*(\operatorname{Cov}(\boldsymbol{y},\boldsymbol{y^*}) + \boldsymbol{\mu}\boldsymbol{\mu^{*t}})\right\},
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{y}}(\boldsymbol{y} - H\boldsymbol{y})^t R^{-1}(\boldsymbol{y} - H\boldsymbol{y}) =& \operatorname{tr}\left\{R^{-1}\mathbb{E}_{\boldsymbol{y}}(\boldsymbol{y}\boldsymbol{y^t})\right\} + \operatorname{tr}\left\{(H^t R^{-1} H\mathbb{E}_{\boldsymbol{y}}(\boldsymbol{y}\boldsymbol{y^t})\right\} \\
& - 2\operatorname{tr}\left\{R^{-1}H\mathbb{E}_{\boldsymbol{y}}(\boldsymbol{y}\boldsymbol{y^t})\right\} \\
=& \operatorname{tr}\left\{R^{-1}(V + \boldsymbol{\mu}\boldsymbol{\mu^t})\right\} + \operatorname{tr}\left\{R^{-1}H(V + \boldsymbol{\mu}\boldsymbol{\mu^t})H^t\right\} \\
& - 2\operatorname{tr}\left\{R^{-1}H(V + \boldsymbol{\mu}\boldsymbol{\mu^t})\right\}.
\end{aligned}$$

Therefore, the difference between the last two terms of equation (12) is

$$\begin{aligned}
&\frac{1}{n^*}\mathbb{E}_{\boldsymbol{y},\boldsymbol{y^*}}(\boldsymbol{y^*} - H^*\boldsymbol{y})^t R^{*^{-1}}(\boldsymbol{y^*} - H^*\boldsymbol{y}) - \frac{1}{n}\mathbb{E}_{\boldsymbol{y}}(\boldsymbol{y} - H\boldsymbol{y})^t R^{-1}(\boldsymbol{y} - H\boldsymbol{y}) \\
&\quad = \frac{2}{n}\operatorname{tr}\left(R^{-1}HV\right) - \frac{2}{n^*}\operatorname{tr}\left(R^{*^{-1}}H^*\operatorname{Cov}(\boldsymbol{y},\boldsymbol{y^*})\right) \\
&\quad\quad + \frac{1}{n^*}\operatorname{tr}\left(R^{*^{-1}}V^*\right) - \frac{1}{n}\operatorname{tr}\left(R^{-1}V\right) \\
&\quad\quad + \frac{1}{n^*}\operatorname{tr}\left(R^{*^{-1}}H^*VH^{*t}\right) - \frac{1}{n}\operatorname{tr}\left(R^{-1}HVH^t\right) \\
&\quad\quad + \frac{1}{n}\operatorname{tr}\left(R^{-1}(2H\boldsymbol{\mu}\boldsymbol{\mu^t} - \boldsymbol{\mu}\boldsymbol{\mu^t} - H\boldsymbol{\mu}\boldsymbol{\mu^t}H^t)\right) \\
&\quad\quad - \frac{1}{n^*}\operatorname{tr}\left(R^{*^{-1}}(2H^*\boldsymbol{\mu}\boldsymbol{\mu^{*t}} - \boldsymbol{\mu^*}\boldsymbol{\mu^{*t}} - H^*\boldsymbol{\mu}\boldsymbol{\mu^t}H^{*t})\right).
\end{aligned}$$

By the assumptions of $H\boldsymbol{\mu} = \boldsymbol{\mu}$ and $H^*\boldsymbol{\mu} = \boldsymbol{\mu^*}$:

$$\begin{aligned}
H\boldsymbol{\mu}\boldsymbol{\mu^t} &= \boldsymbol{\mu}\boldsymbol{\mu^t} \\
H\boldsymbol{\mu}\boldsymbol{\mu^t}H^t &= \boldsymbol{\mu}\boldsymbol{\mu^t} \\
\\
H^*\boldsymbol{\mu}\boldsymbol{\mu^{*t}} &= \boldsymbol{\mu^*}\boldsymbol{\mu^{*t}} \\
H^*\boldsymbol{\mu}\boldsymbol{\mu^t}H^{*t} &= \boldsymbol{\mu^*}\boldsymbol{\mu^{*t}},
\end{aligned}$$

which give

$$\begin{aligned}
\frac{1}{n}\operatorname{tr}\left(R^{-1}(2H\boldsymbol{\mu}\boldsymbol{\mu^t} - \boldsymbol{\mu}\boldsymbol{\mu^t} - H\boldsymbol{\mu}\boldsymbol{\mu^t}H^t)\right) &= 0 \\
\frac{1}{n^*}\operatorname{tr}\left(R^{*^{-1}}(2H^*\boldsymbol{\mu}\boldsymbol{\mu^{*t}} - \boldsymbol{\mu^*}\boldsymbol{\mu^{*t}} - H^*\boldsymbol{\mu}\boldsymbol{\mu^t}H^{*t})\right) &= 0.
\end{aligned}$$

Therefore, the the difference between the last two terms of equation (12) is reduced to:

$$\begin{aligned}\frac{1}{n^*}\mathbb{E}_{\boldsymbol{y},\boldsymbol{y^*}}(\boldsymbol{y^*}-H^*\boldsymbol{y})^t R^{*^{-1}}(\boldsymbol{y^*}-H^*\boldsymbol{y}) - \frac{1}{n}\mathbb{E}_{\boldsymbol{y}}(\boldsymbol{y}-H\boldsymbol{y})^t R^{-1}(\boldsymbol{y}-H\boldsymbol{y})\\ = \frac{2}{n}\mathrm{tr}\left(R^{-1}HV\right) - \frac{2}{n^*}\mathrm{tr}\left(R^{*^{-1}}H^*\mathrm{Cov}(\boldsymbol{y},\boldsymbol{y^*})\right)\\ + \frac{1}{n^*}\mathrm{tr}\left(R^{*^{-1}}V^*\right) - \frac{1}{n}\mathrm{tr}\left(R^{-1}V\right)\\ + \frac{1}{n^*}\mathrm{tr}\left(R^{*^{-1}}H^*VH^{*t}\right) - \frac{1}{n}\mathrm{tr}\left(R^{-1}HVH^t\right),\end{aligned}$$

and the bias is:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{y}}\left[-\frac{1}{n^*}\mathbb{E}_{\boldsymbol{y^*}|\boldsymbol{y}}\ell_{H^*}(\boldsymbol{y^*}) - \left\{-\frac{1}{n}\ell_H(\boldsymbol{y})\right\}\right]\\ = \frac{1}{n}\mathrm{tr}\left(R^{-1}HV\right) - \frac{1}{n^*}\mathrm{tr}\left(R^{*^{-1}}H^*\mathrm{Cov}(\boldsymbol{y},\boldsymbol{y^*})\right)\\ + \frac{1}{2}\left\{\log\left(\frac{|R^*|^{\frac{1}{n^*}}}{|R|^{\frac{1}{n}}}\right) + \frac{1}{n^*}\mathrm{tr}\left(R^{*^{-1}}V^*\right) - \frac{1}{n}\mathrm{tr}\left(R^{-1}V\right)\right\}\\ + \frac{1}{2}\left\{\frac{1}{n^*}\mathrm{tr}\left(R^{*^{-1}}H^*VH^{*t}\right) - \frac{1}{n}\mathrm{tr}\left(R^{-1}HVH^t\right)\right\} \quad \square\end{aligned}$$

A.2. Proof of BLUP optimally for least squares problem with a linear solution

The following theorem will be proven below:

Theorem 7. *Let*

$$\{\boldsymbol{a}^*, B^*\} = \underset{\boldsymbol{a}\in\mathbb{R}^{n^*}, B\in\mathbb{R}^{n^*\times n}}{\mathrm{argmin}} \mathbb{E}_{\boldsymbol{y^*},\boldsymbol{y}}\|\boldsymbol{y^*}-(\boldsymbol{a}+B\boldsymbol{y})\|_2^2.$$

Then

$$\boldsymbol{a^*} + B^*\boldsymbol{y} = \boldsymbol{\mu^*} + \mathrm{Cov}(\boldsymbol{y^*},\boldsymbol{y})V^{-1}(\boldsymbol{y}-\boldsymbol{\mu}).$$

Proof. Under some regularity conditions

$$\begin{aligned}\frac{\partial\mathbb{E}_{\boldsymbol{y},\boldsymbol{y^*}}\|\boldsymbol{y^*}-(\boldsymbol{a}+B\boldsymbol{y})\|_2^2}{\partial\boldsymbol{a}} &= \frac{\mathbb{E}_{\boldsymbol{y},\boldsymbol{y^*}}\partial\|\boldsymbol{y^*}-(\boldsymbol{a}+B\boldsymbol{y})\|_2^2}{\partial\boldsymbol{a}}\\ \frac{\partial\mathbb{E}_{\boldsymbol{y},\boldsymbol{y^*}}\|\boldsymbol{y^*}-(\boldsymbol{a}+B\boldsymbol{y})\|_2^2}{\partial B} &= \frac{\mathbb{E}_{\boldsymbol{y},\boldsymbol{y^*}}\partial\|\boldsymbol{y^*}-(\boldsymbol{a}+B\boldsymbol{y})\|_2^2}{\partial B}.\end{aligned}$$

Since

$$\begin{aligned}\frac{\partial\|\boldsymbol{y^*}-(\boldsymbol{a}+B\boldsymbol{y})\|_2^2}{\partial\boldsymbol{a}} &= \frac{\partial\left(-\boldsymbol{y^{*t}a}-\boldsymbol{a^t y^*}+\boldsymbol{a^t a}+\boldsymbol{a^t}B\boldsymbol{y}+\boldsymbol{y^t}B^t\boldsymbol{a}\right)}{\partial\boldsymbol{a}}\\ &= -2\boldsymbol{y^*}+2\boldsymbol{a}+2B\boldsymbol{y},\end{aligned}$$

then

$$\frac{\partial \mathbb{E}_{\boldsymbol{y},\boldsymbol{y}^*}\|\boldsymbol{y}^* - (\boldsymbol{a} + B\boldsymbol{y})\|_2^2}{\partial \boldsymbol{a}} = -2\boldsymbol{\mu}^* + 2\boldsymbol{a} + 2B\boldsymbol{\mu}.$$

Similarly,

$$\begin{aligned}\frac{\partial\|\boldsymbol{y}^* - (\boldsymbol{a} + B\boldsymbol{y})\|_2^2}{\partial B} &= \frac{\partial\left(-\boldsymbol{y}^{*\boldsymbol{t}}B\boldsymbol{y} + \boldsymbol{a}^{\boldsymbol{t}}B\boldsymbol{y} - \boldsymbol{y}^{\boldsymbol{t}}B^t\boldsymbol{y}^* + \boldsymbol{y}^{\boldsymbol{t}}B^t\boldsymbol{a} + \boldsymbol{y}^{\boldsymbol{t}}B^tB\boldsymbol{y}\right)}{\partial B}\\ &= -2\boldsymbol{y}^*\boldsymbol{y}^{\boldsymbol{t}} + 2\boldsymbol{a}\boldsymbol{y}^{\boldsymbol{t}} + 2B\boldsymbol{y}\boldsymbol{y}^{\boldsymbol{t}}\end{aligned}$$

and therefore

$$\frac{\partial \mathbb{E}_{\boldsymbol{y},\boldsymbol{y}^*}\|\boldsymbol{y}^* - (\boldsymbol{a} + B\boldsymbol{y})\|_2^2}{\partial B} = -2\left(\mathrm{Cov}(\boldsymbol{y}^*, \boldsymbol{y}) + \boldsymbol{\mu}^{*\boldsymbol{t}}\boldsymbol{\mu}\right) + 2\boldsymbol{a}\boldsymbol{\mu}^{\boldsymbol{t}} + 2B\left(V + \boldsymbol{\mu}\boldsymbol{\mu}^{\boldsymbol{t}}\right).$$

Since the optimized function is convex, the solution of the following equations achieves the global minimum where

$$\begin{aligned}0 &= -\boldsymbol{\mu}^* + \boldsymbol{a} + B\boldsymbol{\mu}\\ 0 &= -\mathrm{Cov}(\boldsymbol{y}^*, \boldsymbol{y}) - \boldsymbol{\mu}^*\boldsymbol{\mu}^{\boldsymbol{t}} + \boldsymbol{a}\boldsymbol{\mu}^{\boldsymbol{t}} + B\left(V + \boldsymbol{\mu}\boldsymbol{\mu}^{\boldsymbol{t}}\right).\end{aligned}$$

The solution for B is

$$\begin{aligned}B\left(V + \boldsymbol{\mu}\boldsymbol{\mu}^{\boldsymbol{t}}\right) &= \mathrm{Cov}(\boldsymbol{y}^*, \boldsymbol{y}) + \boldsymbol{\mu}^*\boldsymbol{\mu}^{\boldsymbol{t}} - \boldsymbol{a}\boldsymbol{\mu}^{\boldsymbol{t}}\\ &= \mathrm{Cov}(\boldsymbol{y}^*, \boldsymbol{y}) + \boldsymbol{\mu}^*\boldsymbol{\mu}^{\boldsymbol{t}} - \left(\boldsymbol{\mu}^* - B\boldsymbol{\mu}\right)\boldsymbol{\mu}^{\boldsymbol{t}}\\ &= \mathrm{Cov}(\boldsymbol{y}^*, \boldsymbol{y}) + B\boldsymbol{\mu}\boldsymbol{\mu}^{\boldsymbol{t}},\end{aligned}$$

which gives

$$B = \mathrm{Cov}(\boldsymbol{y}^*, \boldsymbol{y})V^{-1}.$$

The solution for $\boldsymbol{a}$ is

$$\begin{aligned}\boldsymbol{a} &= \boldsymbol{\mu}^* - B\boldsymbol{\mu}\\ &= \boldsymbol{\mu}^* - \mathrm{Cov}(\boldsymbol{y}^*, \boldsymbol{y})V^{-1}\boldsymbol{\mu}\end{aligned}$$

which gives

$$\begin{aligned}\boldsymbol{a} + B\boldsymbol{y} &= \boldsymbol{\mu}^* - \mathrm{Cov}(\boldsymbol{y}^*, \boldsymbol{y})V^{-1}\boldsymbol{\mu} + \mathrm{Cov}(\boldsymbol{y}^*, \boldsymbol{y})V^{-1}\boldsymbol{y}\\ &= \boldsymbol{\mu}^* + \mathrm{Cov}(\boldsymbol{y}^*, \boldsymbol{y})V^{-1}(\boldsymbol{y} - \boldsymbol{\mu}).\end{aligned}$$

Therefore $\hat{\boldsymbol{f}}^*$ can be seen as an estimator of the optimal linear predictor for the squared error loss. □

Appendix B: Additional results

B.1. Model selection performances – an extension

Figure 9 presents the average prediction error

$$\mathbb{E}_{\boldsymbol{y}^*|\boldsymbol{y}} - \frac{1}{n^*}\ell_{h_{best}}(\boldsymbol{y}^*),$$

over the different simulation runs, where h_{best} is the selected model by the relevant criteria, *mAIC*, *cAIC* and *tAIC*. This error reflects the true average prediction error that is obtained when implementing the different model selection criteria. In addition, the average prediction error of the oracle criterion,

$$h_{best} = \underset{h\in\{1,2,3\}}{\operatorname{argmin}} -\frac{1}{n^*}\mathbb{E}_{\boldsymbol{y}^*|\boldsymbol{y}}\ell_h(\boldsymbol{y}^*),$$

is presented.

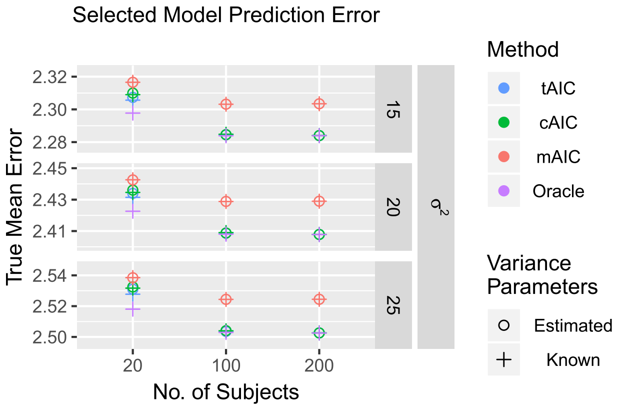


FIG 9. *The figure presents for nine different setups the average true prediction error of the selected model by the criteria: tAIC, cAIC and mAIC with respect to the oracle* $\underset{h\in\mathcal{H}}{\operatorname{argmin}}\mathbb{E}_{\boldsymbol{y}^*|\boldsymbol{y}} - \frac{1}{n^*}\ell_{h_{best}}(\boldsymbol{y}^*)$*, along with the true prediction error of the selected model by the oracle itself. Two versions of the criteria tAIC, cAIC and mAIC are presented: when the variance parameters are known and when they are estimated.*

Figure 10 presented the same analysis with respect to the prediction errors

$$\mathbb{E}_{\boldsymbol{y}}\mathbb{E}_{\boldsymbol{y}^*|\boldsymbol{y}} - \frac{1}{n^*}\ell_{h_{best}}(\boldsymbol{y}^*)$$

and

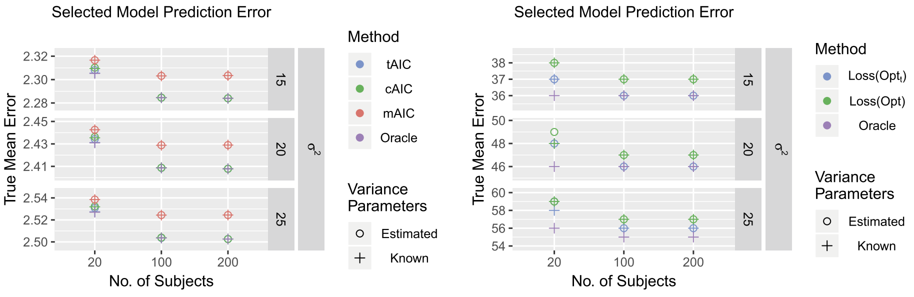
$$\frac{1}{n^*}\mathbb{E}_{\boldsymbol{y}^*|\boldsymbol{y}}\|\boldsymbol{y}^* - H^*\boldsymbol{y}\|_2^2.$$

B.2. *tAI in high-dimensional data setting*

Here, the performance of *tAI* and *tAIC* in a scenario when there are many covariates, is analyzed.

The LMM setting here is as follows:

$$\begin{aligned}\phi_{i,j} =& x_{i,j,0} + 0.1\sum_{k=1}^{k=50} x_{i,j,k} + \sum_{k=51}^{k=150} x_{i,j,k} + 0.5 \times time_{i,j}\\ &+ b_{i,1} + time_{i,j} \times b_{i,2} + \epsilon_{i,j},\end{aligned}$$



(a) (b)

FIG 10. *Figure (10a) presents for nine different setups the average true prediction error of the selected model by the criteria: tAIC, cAIC and mAIC with respect to the oracle* $\underset{h\in\mathcal{H}}{\operatorname{argmin}}\mathbb{E}_{\boldsymbol{y}}\mathbb{E}_{\boldsymbol{y^*}|\boldsymbol{y}} - \frac{1}{n^*}\ell_{h_{best}}(\boldsymbol{y^*})$*, along with the true prediction error of the selected model by the oracle itself. Two versions of the criteria tAIC, cAIC and mAIC are presented: when the variance parameters are known and when they are estimated. Figure (10b) presents the same graph, however with respect to the oracle* $\underset{h\in\mathcal{H}}{\operatorname{argmin}}\frac{1}{n^*}\mathbb{E}_{\boldsymbol{y^*}|\boldsymbol{y}}\|\boldsymbol{y^*} - H^*\boldsymbol{y}\|_2^2$.

where $x_{i,j,0} = 1$ and $x_{i,j,k}$, $\forall k > 0$ was drawn from standard normal distribution. Unlike in Section 5, $0 \le k \le 150$. Other objects, such as $time_{i,j}$, $b_{i,1}$, $b_{i,2}$, $\epsilon_{i,j}$, $\boldsymbol{y}$ and $\boldsymbol{y^*}$ are defined in the same way as they are defined in Section 5. Also, $S = 200$ and $\sigma^2 = 20$.

Figure 11 presents the densities of tAI, mAI, cAI and the oracle, $-\mathbb{E}_{\boldsymbol{y^*}|\boldsymbol{y}}\ell(\boldsymbol{y^*})/n_*$. Table 3 presents the agreement rate of tAI, mAI and cAI with the oracle, as well as the true mean prediction error of tAI, mAI and cAI (for more information, see Section 5).

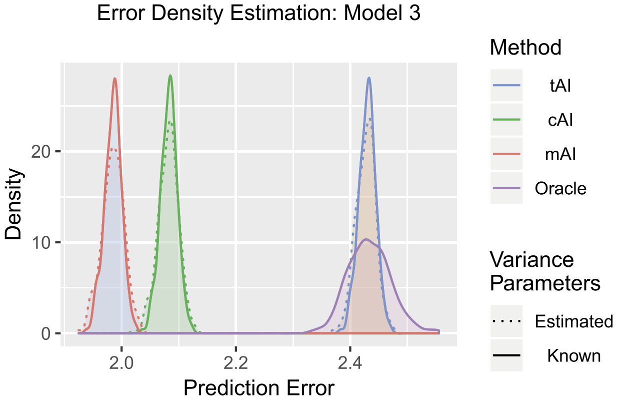


FIG 11. *Densities of tAI, cAI, mAI and* $-\mathbb{E}_{\boldsymbol{y^*}|\boldsymbol{y}}\ell(\boldsymbol{y^*})/n^*$*, where number of subjects*$= 200$ *and* $\sigma^2 = 20$ *in high-dimensional setting.*

TABLE 3
Model Selection Performances
High Dimensional Data

	Agreement Rate with Oracle		Mean Prediction Error	
Estimator/Variance Parameters	True	Estimated	True	Estimated
tAIC	0.96	0.96	2.43	2.43
cAIC	0.88	0.85	2.44	2.44
mAIC	0.02	0.02	2.45	2.45
Oracle	—	—	2.43	2.43

References

AKAIKE, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19** 716–723. MR0423716

ANDERSON, T. W. and DARLING, D. A. (1952). Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. *The Annals of Mathematical Statistics* **23** 193–212. MR0050238

BALTAGI, B. H. (2008). Forecasting with panel data. *Journal of Forecasting* **27** 153–173. MR2420179

BROWN, D. P. and COMRIE, A. C. (2002). Spatial modeling of winter temperature and precipitation in Arizona and New Mexico, USA. *Climate Research* **22** 115–128.

BRUMBACK, B. A., RUPPERT, D. and WAND, M. P. (1999). Variable selection and function estimation in additive nonparametric regression using a data-based prior: Comment. *Journal of the American Statistical Association* **94** 794–797. MR1723272

EFRON, B. (1986). How biased is the apparent error rate of a prediction rule? *Journal of the American Statistical Association* **81** 461–470. MR0845884

GREVEN, S. and KNEIB, T. (2010). On the behaviour of marginal and conditional AIC in linear mixed models. *Biometrika* **97** 773–789. MR2746151

HARVILLE, D. et al. (1976). Extension of the Gauss-Markov theorem to include the estimation of random effects. *The Annals of Statistics* **4** 384–395. MR0398007

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. New York, NY: Springer. MR2722294

HO, B.-C., ANDREASEN, N. C., ZIEBELL, S., PIERSON, R. and MAGNOTTA, V. (2011). Long-term antipsychotic treatment and brain volumes: a longitudinal study of first-episode schizophrenia. *Archives of General Psychiatry* **68** 128–137.

HODGES, J. S. and SARGENT, D. J. (2001). Counting degrees of freedom in hierarchical and other richly-parameterised models. *Biometrika* **88** 367–379. MR1844837

HOGAN, J. W., ROY, J. and KORKONTZELOU, C. (2004). Handling drop-out in longitudinal studies. *Statistics in Medicine* **23** 1455–1497.

Joachims, T. (1999). Transductive inference for text classification using support vector machines. In *Proceedings of ICML-99, 16th International Conference on Machine Learning* **99** 200–209.

Johnson, R. A., Wichern, D. W. et al. (2014). *Applied multivariate statistical analysis* **4**. Prentice-Hall New Jersey. MR1168210

Kyriakidis, P. C. and Journel, A. G. (1999). Geostatistical space–time models: a review. *Mathematical Geology* **31** 651–684. MR1694654

Le, Q. V., Smola, A. J., Gärtner, T. and Altun, Y. (2006). Transductive Gaussian process regression with automatic model selection. In *European Conference on Machine Learning* 306–317. Springer. MR2336654

Li, J. and Heap, A. D. (2011). A review of comparative studies of spatial interpolation methods in environmental sciences: Performance and impact factors. *Ecological Informatics* **6** 228–241.

Li, J. and Heap, A. D. (2014). Spatial interpolation methods applied in the environmental sciences: A review. *Environmental Modelling & Software* **53** 173–189.

Li, G., Zhang, P., Wang, J., Gregg, E. W., Yang, W., Gong, Q., Li, H., Li, H., Jiang, Y., An, Y. et al. (2008). The long-term effect of lifestyle interventions to prevent diabetes in the China Da Qing Diabetes Prevention Study: a 20-year follow-up study. *The Lancet* **371** 1783–1789.

Liang, H., Wu, H. and Zou, G. (2008). A note on conditional AIC for linear mixed-effects models. *Biometrika* **95** 773–778. MR2443190

Mahalanobis, P. C. (1936). On the generalised distance in statistics. *Proceedings of the National Institute of Sciences of India, 1936* 49–55.

Mallinckrodt, C. H., Sanger, T. M., Dubé, S., DeBrota, D. J., Molenberghs, G., Carroll, R. J., Potter, W. Z. and Tollefson, G. D. (2003). Assessing and interpreting treatment effects in longitudinal clinical trials with missing data. *Biological Psychiatry* **53** 754–760.

Manton, K. G., Singer, B. and Suzman, R. M. (2012). *Forecasting the health of elderly populations*. Springer Science & Business Media.

O'neill, R. and Temple, R. (2012). The prevention and treatment of missing data in clinical trials: an FDA perspective on the importance of dealing with it. *Clinical Pharmacology & Therapeutics* **91** 550–554.

Pope, C., T Burnett, R., Thun, M., E Calle, E., Krewski, D., Ito, K. and Thurston, G. (2002). Lung Cancer, Cardiopulmonary Mortality, and Long-Term Exposure to Fine Particulate Air Pollution. *The Journal of the American Medical Association* **287** 1132–1141.

Potthoff, R. F. and Roy, S. (1964). A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika* **51** 313–326. MR0181062

Rikken, M. and Van Rijn, R. (1993). *Soil pollution with heavy metals: in inquiry into spatial variation, cost of mapping and the risk evaluation of Copper, Cadmium, Lead and Zinc in the floodplains of the Meuse West of Stein, The Netherlands: field study report*. University of Utrecht.

Salimi-Khorshidi, G., Nichols, T. E., Smith, S. M. and Woolrich, M. W. (2011). Using Gaussian-process regression for meta-analytic neuroimaging inference based on sparse observations. *IEEE Transactions on Medical Imaging* **30** 1401–1416.

Stahl, K., Moore, R., Floyer, J., Asplin, M. and McKendry, I. (2006). Comparison of approaches for spatial interpolation of daily air temperature in a large region with complex topography and highly variable station density. *Agricultural and Forest Meteorology* **139** 224–236.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics* **9** 1135–1151. MR0630098

Stewart, S. T., Cutler, D. M. and Rosen, A. B. (2009). Forecasting the effects of obesity and smoking on US life expectancy. *New England Journal of Medicine* **361** 2252–2260.

Tsanas, A., Little, M. A., McSharry, P. E. and Ramig, L. O. (2010). Accurate telemonitoring of Parkinson's disease progression by noninvasive speech tests. *IEEE Transactions on Biomedical Engineering* **57** 884–893.

Vaida, F. and Blanchard, S. (2005). Conditional Akaike information for mixed-effects models. *Biometrika* **92** 351–370. MR2201364

Verbeke, G. and Molenberghs, G. (2009). *Linear mixed models for longitudinal data*. Springer Science & Business Media. MR2723365

Vicente-Serrano, S. M., Saz-Sánchez, M. A. and Cuadrat, J. M. (2003). Comparative analysis of interpolation methods in the middle Ebro Valley (Spain): application to annual precipitation and temperature. *Climate Research* **24** 161–180.

Wilks, S. S. (1932). Certain generalizations in the analysis of variance. *Biometrika* **24** 471–494.

Wood, A. M., White, I. R. and Thompson, S. G. (2004). Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. *Clinical Trials* **1** 368–376.

Wray, N. R., Yang, J., Hayes, B. J., Price, A. L., Goddard, M. E. and Visscher, P. M. (2013). Pitfalls of predicting complex traits from SNPs. *Nature Reviews Genetics* **14** 507–515.

Zhang, Z., Ersoz, E., Lai, C.-Q., Todhunter, R. J., Tiwari, H. K., Gore, M. A., Bradbury, P. J., Yu, J., Arnett, D. K., Ordovas, J. M. et al. (2010). Mixed linear model approach adapted for genome-wide association studies. *Nature Genetics* **42** 355–360.