

Burgard, Jan Pablo; Krause, Joscha; Kreber, Dennis; Morales, Domingo

Article — Published Version

The generalized equivalence of regularization and min–max robustification in linear mixed models

Statistical Papers

Provided in Cooperation with:

Springer Nature

Suggested Citation: Burgard, Jan Pablo; Krause, Joscha; Kreber, Dennis; Morales, Domingo (2021) : The generalized equivalence of regularization and min–max robustification in linear mixed models, Statistical Papers, ISSN 1613-9798, Springer, Berlin, Heidelberg, Vol. 62, Iss. 6, pp. 2857-2883,
<https://doi.org/10.1007/s00362-020-01214-z>

This Version is available at:

<https://hdl.handle.net/10419/286749>

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.



<https://creativecommons.org/licenses/by/4.0/>



The generalized equivalence of regularization and min–max robustification in linear mixed models

Jan Pablo Burgard¹ · Joscha Krause¹ · Dennis Kreber² · Domingo Morales³

Received: 21 April 2020 / Revised: 22 October 2020 / Accepted: 29 October 2020 /
Published online: 11 January 2021
© The Author(s) 2020

Abstract

The connection between regularization and min–max robustification in the presence of unobservable covariate measurement errors in linear mixed models is addressed. We prove that regularized model parameter estimation is equivalent to robust loss minimization under a min–max approach. On the example of the LASSO, Ridge regression, and the Elastic Net, we derive uncertainty sets that characterize the feasible noise that can be added to a given estimation problem. These sets allow us to determine measurement error bounds without distribution assumptions. A conservative Jackknife estimator of the mean squared error in this setting is proposed. We further derive conditions under which min-max robust estimation of model parameters is consistent. The theoretical findings are supported by a Monte Carlo simulation study under multiple measurement error scenarios.

Keywords Measurement errors · Regularized regression · Robust best prediction

1 Introduction

Linear regression is a standard tool in many research fields, such as economics and biology. It is widely used for analyzing the statistical relation between responses and covariates. The method relies on the crucial assumption that the data is observed correctly, which implies the absence of measurement errors during data collection (Mitra and Alam 1980). However, for many applications, this assumption is unrealistic. Economic indicators are often subject to uncertainty since their values are estimated

✉ Jan Pablo Burgard
burgardj@uni-trier.de
Domingo Morales
d.morales@umh.es

¹ Department of Economic and Social Statistics, Trier University, Trier, Germany

² Department of Operations Research, Trier University, Trier, Germany

³ Operations Research Center, University Miguel Hernández de Elche, Elche, Spain

from survey data (Alfons et al. 2013). Biomarker measures may be contaminated due to errors in specimen collection and storage (White 2011). Further, since big data sources are more and more used for analysis (Davalos 2017; Yamada et al. 2018), the measurement error is often not controllable. If the correct data values cannot be recovered from their noisy observations, linear regression fails to provide valid results. In that case, methodological adjustments are required to allow for statistically well-founded results. These adjustments are often summarized under the umbrella term *robust estimation* (Li and Zheng 2007).

Robustness is not connoted consistently in statistics. There is a multitude of different methods that account for the effects of measurement interference in the estimation process. Bertsimas et al. (2017) categorize them into two general approaches to robustification, an optimistic and a pessimistic perspective, which they call the min–min and min–max approach. Let $\tilde{\mathbf{X}} = \mathbf{X} + \mathbf{D}$ be a design matrix that is contaminated by measurement errors. Here, $\mathbf{X} \in \mathbb{R}^{n \times p}$ denotes the original design matrix without errors and $\mathbf{D} \in \mathbb{R}^{n \times p}$ is a matrix of error terms. Burgard et al. (2020a, b) assumed that measurement errors are normally distributed. However, no distributional assumptions are made here about \mathbf{D} . Further, \mathbf{X} and \mathbf{D} are not required to be independent. Now, for a loss function $g : \mathbb{R}^n \rightarrow \mathbb{R}_+, \mathbb{R}_+ = [0, \infty)$, a response vector $\mathbf{y} \in \mathbb{R}^n$, a perturbation matrix $\mathbf{\Delta} \in \mathbb{R}^{n \times p}$, and a set $\mathbb{U} \subseteq \mathbb{R}^{n \times p}$, the min–min approach is formulated by

$$\min_{\beta \in \mathbb{R}^p} \min_{\mathbf{\Delta} \in \mathbb{U}} g(\mathbf{y} - (\tilde{\mathbf{X}} + \mathbf{\Delta})\beta), \quad (1)$$

while the min–max approach is characterized by the problem

$$\min_{\beta \in \mathbb{R}^p} \max_{\mathbf{\Delta} \in \mathbb{U}} g(\mathbf{y} - (\tilde{\mathbf{X}} + \mathbf{\Delta})\beta). \quad (2)$$

In both variants, the design matrix is perturbed to account for some possibly contained additive measurement errors. In the min–min approach, the perturbations $\mathbf{\Delta}$ are chosen minimal with respect to \mathbb{U} . This is the most common idea of robustness in statistics. Bertsimas et al. (2017) refer to it as optimistic, because the researcher is allowed to choose which observations to discard for model parameter estimation. The primary concern of the min–min approach is to robustify against distribution outliers. Therefore, oftentimes distribution information about the measurement errors is required. Examples of min–min methods include Least Trimmed Squares (Rousseeuw and Leroy 2003), Trimmed LASSO (Bertsimas et al. 2017), Total Least Squares (Markovsky and Huffel 2007), as well as M-estimation with influence functions (Huber 1973; Schmid and Münnich 2014). The min–max method, on the other hand, introduces perturbations that are chosen maximal with respect to \mathbb{U} . This idea mainly stems from robust optimization theory. The objective is to find solutions that are still good or feasible under a general level of uncertainty. In the process, deterministic assumptions about \mathbb{U} are made, which is then called *uncertainty set*. The researcher chooses it in accordance to how the additive error might be structured. Bertsimas et al. (2017) refer to this approach as pessimistic, since model parameter estimation is performed under a worst-case scenario for the perturbations. Unlike in the min–min approach, the target is not to robustify against errors of a given distribution, but against errors of a given

magnitude. This robustness viewpoint has been studied for example by El Ghaoui and Le Bret (1997), Ben-Tal et al. (2009), and Bertsimas and Copenhaver (2018).

In practice, distribution information on the measurement errors is rarely available. This is particularly the case for big data sources, as the origin of the data is often unknown. In these settings, it makes sense to adopt robust optimization and regard the disturbance of $\tilde{\mathbf{X}}$ pessimistically. Under this premise, we obtain conservative, yet valid results. That is to say, the min–max method can be used to achieve robust estimates in the absence of distribution information on data contamination. Yet, it is not obvious how to efficiently solve a corresponding min–max problem (Bertsimas et al. 2011). But recent results from robust optimization show that it is related to regularized regression problems of the form

$$\min_{\beta \in \mathbb{R}^p} g(\mathbf{y} - \tilde{\mathbf{X}}\beta) + \lambda h(\beta), \quad (3)$$

where $\lambda > 0$ and $h : \mathbb{R}^p \rightarrow \mathbb{R}_+$ is a regularization. From an optimization stand point, problems like (3) can be handled better and solved more efficiently. But given the literature on regularized regression, it is uncommon to regard (3) as robustification. In many cases, regression models are extended by regularization due to at least one of the following aspects: (i) allow for high-dimensional inference (Neykov et al. 2014), (ii) perform variable selection (Zhang and Xiang 2015; Yang et al. 2018), and (iii) deal with multicollinearity (Norouzirad and Arashi 2019). Apart from these common applications, Bertsimas and Copenhaver (2018) provided novel insights by showing that (2) and (3) are equivalent if g is a semi-norm and h is a norm. Unfortunately, many regularized regression methods do not fit naturally in this framework. Popular techniques like the LASSO (Tibshirani 1996), Ridge regression (Hoerl and Kennard 1970), or the Elastic Net (Zou and Hastie 2005) use the squared ℓ_2 -norm as loss function (g), which is not a semi-norm. Further, many regularizations (h) include squares or other functions of norms, which are no norms. Given this limitation, it is desirable to find a more general connection of regularization and robustification that applies to a broader range of methods.

In this paper, we prove that (2) and (3) are equivalent when g and h are functions of semi-norms and norms. On the example of linear mixed models (LMMs; Pinheiro and Bates 2000), we show that regularization obtains efficient estimates in the presence of measurement errors without distribution assumptions. Given the majority of regularized regression applications, this introduces a fairly new perspective on these methods. Past developments mainly focussed on how to robustify regularized regression under contaminated data (Rosenbaum and Tsybakov 2010; Loh and Wainwright 2012; Sørensen et al. 2015). We show that regularization itself is a robustification. Building upon this result, we derive uncertainty sets for the LASSO, Ridge regression and the Elastic Net. They characterize the nature of the respective robustification effect and allow us to find upper bounds for the measurement errors. From the error bounds, we construct a conservative Jackknife estimator of the mean squared error (MSE) for contaminated data. Further, we study conditions under which robust optimization allows for consistency in model parameter estimation.

We proceed as follows. In Sect. 2, the generalized equivalence is established. We use it to derive the uncertainty sets resulting from the three regularizations. Next, we build a robust version of the LMM and show how robust empirical best predictors

from the model are obtained. Section 3 addresses MSE estimation. We first derive error bounds from the uncertainty sets. Then, we present the conservative Jackknife estimator. In Sect. 4, we cover consistency in model parameter estimation. Section 5 contains a Monte Carlo simulation to demonstrate the effectiveness of the methodology. Section 6 closes with some conclusive remarks. The paper is supported by a supplemental material file with eight appendices. Appendix 1 to 7 contain the proofs of the mathematical developments presented in this study. Appendix 8 contains MSE calculations for a general random effect structure. This paper contains insights of a related working paper by Burgard et al. (2019).

2 Min–max robust linear mixed model

2.1 Min–max robustification

In Sect. 1, we introduced min–max robustification (2) as a conservative approach to obtain estimates in the presence of unknown measurement errors. Since it is unclear to efficiently solve the underlying optimization problem, we now present how it is related to regularized regression problems (3). For this purpose, the following result by Bertsimas and Copenhaver (2018) is helpful, as it connects min–max robustification with regularization.

Proposition 1 (Bertsimas and Copenhaver 2018) *If $g : \mathbb{R}^n \rightarrow \mathbb{R}_+$ is a semi-norm which is not identically zero and $h : \mathbb{R}^p \rightarrow \mathbb{R}_+$ is a norm, then for any $\mathbf{z} \in \mathbb{R}^n$, $\boldsymbol{\beta} \in \mathbb{R}^p$ and $\lambda > 0$*

$$\max_{\Delta \in \mathbb{U}} g(\mathbf{z} + \Delta \boldsymbol{\beta}) = g(\mathbf{z}) + \lambda h(\boldsymbol{\beta}),$$

where

$$\mathbb{U} = \left\{ \Delta \in \mathbb{R}^{n \times p} : \max_{\mathbf{y} \in \mathbb{R}^p} \frac{g(\Delta \mathbf{y})}{h(\mathbf{y})} \leq \lambda \right\}.$$

Clearly, the proposition directly implies that

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \max_{\Delta \in \mathbb{U}} g(\mathbf{y} - (\tilde{\mathbf{X}} + \Delta) \boldsymbol{\beta}) = \min_{\boldsymbol{\beta} \in \mathbb{R}^p} g(\mathbf{y} - \tilde{\mathbf{X}} \boldsymbol{\beta}) + \lambda h(\boldsymbol{\beta}),$$

for g , h and \mathbb{U} as in Proposition 1. Thus, the framework provided by Bertsimas and Copenhaver (2018) gives us novel insights into the role of regularization in regression. The choice of a regularization function h with parameter λ directly constrains the uncertainty set \mathbb{U} , which defines a set of perturbations for the design matrix. In other words, the regularization controls the magnitude of noise that can be added to $\tilde{\mathbf{X}}$. Under this interference, $\boldsymbol{\beta}$ is chosen such that the loss is minimal. The effect can be imagined as a two player game where one player tries to minimize the loss by controlling $\boldsymbol{\beta}$, while the other player tries to maximize the deviation by controlling the noise that is added to $\tilde{\mathbf{X}}$. However, many regression methods are formulated using the squared norm or a mix of squared and non-squared norms. For instance, the LASSO is posed

as the optimization problem

$$\min_{\beta \in \mathbb{R}^p} \|y - \tilde{X}\beta\|_2^2 + \lambda \|\beta\|_1. \tag{4}$$

Here, the deviation is squared while the regularization term is not. On the other hand, Ridge regression is posed as the optimization problem

$$\min_{\beta \in \mathbb{R}^p} \|y - \tilde{X}\beta\|_2^2 + \lambda \|\beta\|_2^2, \tag{5}$$

with both, the deviation and regularization, being squared. Both optimization problems do not fit naturally into the framework of Proposition 1 since a squared (semi-)norm $\|\cdot\|^2$ is not a (semi-)norm. We provide a generalization of Proposition 1 that (i) displays a more fundamental connection between regularization and robustification, and (ii) enables us to regard more sophisticated regularizations in light of robustification.

Theorem 1 *Let $\lambda_1, \dots, \lambda_d$ be positive real numbers, $g : \mathbb{R}^n \rightarrow \mathbb{R}_+$ be a semi-norm which is not identically zero, let $h_1, h_2, \dots, h_d : \mathbb{R}^p \rightarrow \mathbb{R}_+$ be norms and $f, f_1, f_2, \dots, f_d : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ be increasing, convex functions, then there exist $\varphi_1 > 0, \dots, \varphi_d > 0$ such that*

$$\operatorname{argmin}_{\beta \in \mathbb{R}^p} \max_{\Delta \in \mathbb{U}} g(y - (\tilde{X} + \Delta)\beta) = \operatorname{argmin}_{\beta \in \mathbb{R}^p} f(g(y - \tilde{X}\beta)) + \sum_{l=1}^d \lambda_l f_l(h_l(\beta)),$$

where

$$\mathbb{U} = \left\{ \Delta \in \mathbb{R}^{n \times p} : g(\Delta\gamma) \leq \sum_{l=1}^d \varphi_l h_l(\gamma) \text{ for all } \gamma \in \mathbb{R}^p \right\}.$$

The proof can be found in Appendix 1 of the supplemental material. Observe the difference between Theorem 1 and Proposition 1. In the original statement, regularization and robustification are equivalent when the loss function is a semi-norm and the regularization is a norm. In the generalization, the equivalence also holds when loss function and regularization are increasing convex functions of semi-norms and norms. This covers a broader range of settings, which we demonstrate hereafter. For Ridge regression, we have $g(z) = \|z\|_2, h_1(z) = \|z\|_2, f(z) = z^2, f_1(z) = z^2$ and $d = 1$ with $\lambda > 0$. Applying Theorem 1 yields

$$\hat{\beta}^{\ell_2} = \operatorname{argmin}_{\beta \in \mathbb{R}^p} \max_{\Delta \in \mathbb{U}_{\ell_2}} \|y - (\tilde{X} + \Delta)\beta\|_2 = \operatorname{argmin}_{\beta \in \mathbb{R}^p} \|y - \tilde{X}\beta\|_2^2 + \lambda \|\beta\|_2^2, \tag{6}$$

with

$$\mathbb{U}_{\ell_2} = \{\Delta \in \mathbb{R}^{n \times p} : \|\Delta\gamma\|_2 \leq \varphi \|\gamma\|_2 \ \forall \gamma \in \mathbb{R}^p\}, \tag{7}$$

for some $\varphi > 0$. For the LASSO, we have $g(z) = \|z\|_2, h_1(z) = \|z\|_1, f(z) = z^2, f_1(z) = z$ and $d = 1$ with $\lambda > 0$. Applying Theorem 1 obtains

$$\hat{\beta}^{\ell_1} = \operatorname{argmin}_{\beta \in \mathbb{R}^p} \max_{\Delta \in \mathbb{U}_{\ell_1}} \left\| \mathbf{y} - (\tilde{\mathbf{X}} + \Delta) \beta \right\|_2 = \operatorname{argmin}_{\beta \in \mathbb{R}^p} \left\| \mathbf{y} - \tilde{\mathbf{X}} \beta \right\|_2^2 + \lambda \|\beta\|_1, \tag{8}$$

with

$$\mathbb{U}_{\ell_1} = \{ \Delta \in \mathbb{R}^{n \times p} : \|\Delta \mathbf{y}\|_2 \leq \varphi \|\mathbf{y}\|_1 \ \forall \mathbf{y} \in \mathbb{R}^p \}. \tag{9}$$

And finally, for the Elastic Net, we have $g(z) = \|z\|_2, h_1(z) = \|z\|_1, h_2(z) = \|z\|_2, f(z) = z^2, f_1(z) = z, f_2(z) = z^2$ and $d = 2$ with $\lambda_1, \lambda_2 > 0$. Applying Theorem 1 yields

$$\hat{\beta}^{EN} = \operatorname{argmin}_{\beta \in \mathbb{R}^p} \max_{\Delta \in \mathbb{U}_{EN}} \left\| \mathbf{y} - (\tilde{\mathbf{X}} + \Delta) \beta \right\|_2 = \operatorname{argmin}_{\beta \in \mathbb{R}^p} \left\| \mathbf{y} - \tilde{\mathbf{X}} \beta \right\|_2^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2, \tag{10}$$

with

$$\mathbb{U}_{EN} = \{ \Delta \in \mathbb{R}^{n \times p} : \|\Delta \mathbf{y}\|_2 \leq \varphi_1 \|\mathbf{y}\|_1 + \varphi_2 \|\mathbf{y}\|_2 \ \forall \mathbf{y} \in \mathbb{R}^p \}. \tag{11}$$

We see that the theorem can be broadly applied and establishes the robustification effect for a variety of regularized regression methods. However, the manner in which robustification is achieved depends on the regularization. By looking at the definition of \mathbb{U} in Theorem 1, we see that the effects of measurement errors with respect to the loss function are bounded by a generic term $\sum_{l=1}^d \varphi_l h_l(\cdot)$ for some $\varphi_1 > 0, \dots, \varphi_d > 0$. The exact form of this term depends on the regularization the researcher wishes to apply. Accordingly, in the light of the three regularized regression approaches considered before, the robustification effect manifests itself differently given the penalty. On that note, Bertsimas and Copenhaver (2018) provided another result that allows for an interpretation of the robustification effects. It is summarized within the subsequent proposition.

Proposition 2 (Bertsimas and Copenhaver 2018) *Let $t \in [1, \infty], \|\beta\|_0$ be the number of non-zero entries of β and Δ_i be the i -th column of Δ for $i = 1, \dots, p$. If*

$$\mathbb{U}' = \{ \Delta \in \mathbb{R}^{n \times p} : \|\Delta \beta\|_2 \leq \varphi \|\beta\|_0 \text{ for all } \|\beta\|_t \leq 1 \},$$

and

$$\mathbb{U}'' = \{ \Delta \in \mathbb{R}^{n \times p} : \|\Delta_i\|_2 \leq \varphi, \ i = 1, \dots, p \},$$

then $\mathbb{U}_{\ell_1} = \mathbb{U}' = \mathbb{U}''$.

Applying this result to our generalization, we see that for \mathbb{U}_{ℓ_2} (Ridge regression), the maximum singular value of Δ is bounded by φ . With respect to \mathbb{U}_{ℓ_1} (LASSO), the columnwise ℓ_2 -norm of Δ is bounded by φ . Thus, while Ridge regression induces an upper bound on the entire noise matrix, the LASSO provides a componentwise bound. With respect to the Elastic Net, the error bound is a linear combination of the componentwise LASSO bound and the general Ridge bound. Unfortunately, it is less apparent how to interpret the corresponding robustification effect.

2.2 Model and robust empirical best prediction

Based on these insights, we now use min-max robustification to construct a robust version of the basic LMM in a finite population setting. Let \mathcal{U} denote a population of $|\mathcal{U}| = N$ individuals indexed by $i = 1, \dots, N$. Assume that \mathcal{U} is segmented into m domains \mathcal{U}_j of size $|\mathcal{U}_j| = N_j$ indexed by $j = 1, \dots, m$ with \mathcal{U}_j and \mathcal{U}_k pairwise disjoint for all $j \neq k$. Let $\mathcal{S} \subset \mathcal{U}$ be a random sample of size $|\mathcal{S}| = n < N$ that is drawn from \mathcal{U} . Assume that the sample design is such that there are domain-specific subsamples $\mathcal{S}_j \subset \mathcal{U}_j$ of size $|\mathcal{S}_j| = n_j > 1$ for all $j = 1, \dots, m$. Let Y be a real-valued response variable of interest. Denote $y_{ij} \in \mathbb{R}$ as the realization of Y for a given individual $i \in \mathcal{U}_j$. For convenience, assume that the objective is to estimate the mean of Y for all population domains, that is

$$\bar{Y}_j = \frac{1}{N_j} \sum_{i \in \mathcal{U}_j} y_{ij}, \quad j = 1, \dots, m. \tag{12}$$

Let X be a p -dimensional real-valued vector of covariates statistically related to Y . Let $\mathbf{x}_{ij} \in \mathbb{R}^{1 \times p}$ denote the realization of X for $i \in \mathcal{U}_j$ and let $\mathbf{z}_{ij} \in \mathbb{R}^{1 \times q}$ be the known incidence vector. In the light of Sect. 1, we assume that the observations of X are impaired by measurement errors. That is, we only observe $\tilde{\mathbf{x}}_{ij} = \mathbf{x}_{ij} + \mathbf{d}_{ij}$, where $\mathbf{d}_{ij} \in \mathbb{R}^{1 \times p}$. The robust LMM is formulated as

$$\mathbf{y}_j = (\mathbf{X}_j + \mathbf{D}_j) \boldsymbol{\beta} + \mathbf{Z}_j \mathbf{b}_j + \mathbf{e}_j, \quad j = 1, \dots, m, \tag{13}$$

where $\mathbf{y}_j = (y_{1j}, \dots, y_{n_j j})'$ is the vector of sample responses in \mathcal{S}_j . Further, we have $\mathbf{X}_j = (\mathbf{x}'_{1j}, \dots, \mathbf{x}'_{n_j j})'$, $\mathbf{D}_j = (\mathbf{d}'_{1j}, \dots, \mathbf{d}'_{n_j j})'$, $\mathbf{Z}_j = (\mathbf{z}'_{1j}, \dots, \mathbf{z}'_{n_j j})'$, and $\boldsymbol{\beta} \in \mathbb{R}^{p \times 1}$ as vector of fixed effect coefficients. The term $\mathbf{Z}_j \in \mathbb{R}^{n_j \times q}$ denotes the random effect design matrix and $\mathbf{b}_j \in \mathbb{R}^{q \times 1}$ with $\mathbf{b}_j \sim N(\mathbf{0}_q, \boldsymbol{\Psi})$ is the vector of random effects. The latter follows a normal distribution with a positive-definite covariance matrix $\boldsymbol{\Psi} \in \mathbb{R}^{q \times q}$ that is parametrized by some vector $\boldsymbol{\psi} \in \mathbb{R}^{q^* \times 1}$. The vector $\mathbf{e}_j \in \mathbb{R}^{n_j \times 1}$ contains random model errors with $\mathbf{e}_j \sim N(\mathbf{0}_{n_j}, \sigma^2 \mathbf{I}_{n_j})$ and a variance parameter σ^2 . We assume that $\mathbf{b}_1, \dots, \mathbf{b}_m, \mathbf{e}_1, \dots, \mathbf{e}_m$ are independent. Under model (13), the response vectors follow independent normals

$$\mathbf{y}_j \sim N\left((\mathbf{X}_j + \mathbf{D}_j)\boldsymbol{\beta}, \mathbf{V}_j(\sigma^2, \boldsymbol{\psi})\right), \quad j = 1, \dots, m, \tag{14}$$

where $\mathbf{V}_j(\sigma^2, \boldsymbol{\psi}) = \mathbf{Z}_j \boldsymbol{\Psi} \mathbf{Z}'_j + \sigma^2 \mathbf{I}_{n_j}$. The conditional distribution $\mathbf{y}_j | \mathbf{b}_j$ is

$$\mathbf{y}_j | \mathbf{b}_j \sim N\left((\mathbf{X}_j + \mathbf{D}_j)\boldsymbol{\beta} + \mathbf{Z}_j \mathbf{b}_j, \sigma^2 \mathbf{I}_{n_j}\right), \quad j = 1, \dots, m. \tag{15}$$

Formulating (13) over all domains yields

$$\mathbf{y} = (\mathbf{X} + \mathbf{D}) \boldsymbol{\beta} + \mathbf{Z} \mathbf{b} + \mathbf{e}, \tag{16}$$

where $\mathbf{y} = (\mathbf{y}'_1, \dots, \mathbf{y}'_m)'$, $\mathbf{b} = (\mathbf{b}'_1, \dots, \mathbf{b}'_m)'$, $\mathbf{e} = (\mathbf{e}'_1, \dots, \mathbf{e}'_m)'$ are stacked vectors and $\mathbf{X} = (\mathbf{X}'_1, \dots, \mathbf{X}'_m)'$ $\in \mathbb{R}^{n \times p}$, $\mathbf{D} = (\mathbf{D}'_1, \dots, \mathbf{D}'_m)'$ $\in \mathbb{R}^{n \times p}$, $\mathbf{Z} = \text{diag}(\mathbf{Z}_1, \dots, \mathbf{Z}_m) \in \mathbb{R}^{n \times mq}$ are stacked matrices. Finally, the model parameters are $\boldsymbol{\theta} := (\boldsymbol{\beta}', \boldsymbol{\eta}')'$ with $\boldsymbol{\eta} = (\sigma^2, \boldsymbol{\psi}')'$. Please note that in accordance with a likelihood-based estimation setting, the random effects \mathbf{b} are not model parameters, but random variables. Thus, in order to obtain robust estimates of (12), random effect and response realizations have to be predicted. For this, we first state the best predictors of \mathbf{b}_j and \bar{Y}_j under the robust LMM with the preliminary assumption that $\boldsymbol{\theta}$ is known. They are obtained from the respective conditional expectations given the response observations under (16). We refer to them as robust best predictors (RBPs). Afterwards, the model parameters are substituted by empirical estimates to obtain the robust empirical best predictors (REBPs). The RBPs are stated in the subsequent Proposition.

Proposition 3 *Under model (16), the RBPs of \mathbf{b}_j and \bar{Y}_j are given by*

$$\begin{aligned} \hat{\mathbf{b}}_j^{RBP}(\boldsymbol{\theta}) &= E(\mathbf{b}_j | \mathbf{y}) = \left(\mathbf{Z}'_j \mathbf{Z}_j + \sigma^2 \boldsymbol{\Psi}^{-1} \right)^{-1} \mathbf{Z}'_j \left[y_j - (\mathbf{X}_j + \mathbf{D}_j) \boldsymbol{\beta} \right], \\ \hat{\bar{Y}}_j^{RBP}(\boldsymbol{\theta}) &= E(\bar{Y}_j | \mathbf{y}) = \frac{1}{N_j} \left[\sum_{i \in \mathcal{S}_j} y_{ij} + \sum_{i \in \mathcal{U}_j \setminus \mathcal{S}_j} \left((\mathbf{x}_{ij} + \mathbf{d}_{ij}) \boldsymbol{\beta} + \mathbf{z}_{ij} \hat{\mathbf{b}}_j^{RBP}(\boldsymbol{\theta}) \right) \right]. \end{aligned}$$

The proof can be found in Appendix 2 of the supplemental material. Note that the RBP of \bar{Y}_j requires covariate observations for all $i \in \mathcal{U}_j$. Such knowledge may be unrealistic in practice, depending on the application. Therefore, we use an alternative expression that is less demanding in terms of data. Battese et al. (1988) suggested the approximation $\bar{Y}_j \approx \mu_j = \bar{\mathbf{X}}_j \boldsymbol{\beta} + \bar{\mathbf{Z}}_j \mathbf{b}_j$ for cases when $n_j / N_j \approx 0$. Here, $\bar{\mathbf{X}}_j$ and $\bar{\mathbf{Z}}_j$ are the domain means of \mathbf{X}_j and \mathbf{Z}_j in domain \mathcal{U}_j . Observe that the unknown μ_j is generated without measurement errors. The RBP of μ_j is

$$\hat{\mu}_j^{RBP}(\boldsymbol{\theta}) = (\bar{\mathbf{X}}_j + \bar{\mathbf{D}}_j)' \boldsymbol{\beta} + \bar{\mathbf{Z}}_j \hat{\mathbf{b}}_j^{RBP}(\boldsymbol{\theta}),$$

where $\bar{\mathbf{D}}_j$ is the hypothetical domain mean of the measurement errors. Based on this approximation and Proposition 3, we can state the REBP of μ_j by substituting the unknown model parameter $\boldsymbol{\theta}$ by an estimator $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}', \hat{\boldsymbol{\eta}}')'$ under the min-max setting. This is done by solving the two optimization problems iteratively. For fixed effect estimation, we solve the regularized weighted least squares problem

$$\begin{aligned} \hat{\boldsymbol{\beta}}(\check{\boldsymbol{\eta}}) &= \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\text{argmin}} \max_{\boldsymbol{\Delta} \in \mathbb{U}} \left\| \mathbf{V}^{-1/2}(\check{\boldsymbol{\eta}}) \left[\mathbf{y} - (\tilde{\mathbf{X}} + \boldsymbol{\Delta}) \boldsymbol{\beta} \right] \right\|_2 \\ &= \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\text{argmin}} \left\| \mathbf{V}^{-1/2}(\check{\boldsymbol{\eta}}) (\mathbf{y} - \tilde{\mathbf{X}} \boldsymbol{\beta}) \right\|_2^2 + \sum_{l=1}^d \lambda_l f_l(h_l(\boldsymbol{\beta})), \end{aligned} \tag{17}$$

given variance parameter candidates $\check{\eta}$ and a predefined regularization $\sum_{l=1}^d \lambda_l f_l(h_l(\boldsymbol{\beta}))$. For variance parameter estimation, we solve the maximum likelihood (ML) problem

$$\hat{\eta}(\hat{\boldsymbol{\beta}}) = \underset{\eta \in \mathbb{R}^{1+q^*}}{\operatorname{argmin}} \frac{1}{2} \left[n \log(2\pi) + \log(|\mathbf{V}(\eta)|) + (\mathbf{y} - \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}})' \mathbf{V}^{-1}(\eta) (\mathbf{y} - \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}) \right] \tag{18}$$

given the robust solution $\hat{\boldsymbol{\beta}}$, where $|\mathbf{V}(\eta)|$ is the determinant of $\mathbf{V}(\eta)$. Both estimation steps are performed conditionally on each other until convergence. For an iteration index $r = 1, 2, \dots$, the complete procedure is summarized in the subsequent algorithm.

Algorithm 1 Model parameter estimation

- 1: Center the response observations \mathbf{y} and standardize the covariate observations $\tilde{\mathbf{X}}$.
 - 2: Find initial model parameter estimates $\hat{\boldsymbol{\theta}}^{[0]}$.
 - 3: **while** not converged **do**
 - 4: Solve (17) to obtain $\hat{\boldsymbol{\beta}}^{[r]}(\hat{\eta}^{[r-1]})$.
 - 5: Solve (18) to find $\hat{\eta}^{[r]}(\hat{\boldsymbol{\beta}}^{[r]})$.
 - 6: **end**
-

In the upper descriptions, centering means transforming \mathbf{y} such that it has zero mean. Further, standardization implies transforming $\tilde{\mathbf{X}}$ such that each of its columns has zero mean and unit length. At this point, we omit a detailed presentation of suitable methods to solve the individual problems. This already been addressed exhaustively in the literature. For the solution of (17), coordinate descent methods are often applied. See for instance Tseng and Yun (2009), Friedman et al. (2010), as well as Bottou et al. (2018). Regarding the solution of (18), a Newton–Raphson algorithm can be used (Lindstrom and Bates 1988; Searle et al. 1992). For an estimate $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}', \hat{\eta}')'$, the REBP of μ_j is

$$\hat{\mu}_j^{REBP} = \hat{\mu}_j^{RBP}(\hat{\boldsymbol{\theta}}) = (\bar{\mathbf{X}}_j + \bar{\mathbf{D}}_j)' \hat{\boldsymbol{\beta}} + \bar{\mathbf{Z}}_j \hat{\mathbf{b}}_j^{REBP}, \tag{19}$$

where

$$\hat{\mathbf{b}}_j^{REBP} = \hat{\mathbf{b}}_j^{RBP}(\hat{\boldsymbol{\theta}}) = (\mathbf{Z}'_j \mathbf{Z}_j + \hat{\sigma}^2 \boldsymbol{\Psi}^{-1}(\hat{\boldsymbol{\psi}}))^{-1} \mathbf{Z}'_j [\mathbf{y}_j - (\mathbf{X}_j + \mathbf{D}_j) \hat{\boldsymbol{\beta}}].$$

3 Conservative MSE estimation under measurement errors

Hereafter, we demonstrate how to obtain conservative estimates for the MSE of the REBP, which is given by $\text{MSE}(\hat{\mu}_j^{REBP}) = \text{E}[(\hat{\mu}_j^{REBP} - \mu_j)^2]$. This is done in two steps. We first derive upper bounds for the MSE of the RBP under known model parameters in the presence of measurement errors, that is, $\text{MSE}(\hat{\mu}_j^{RBP}) = \text{E}[(\hat{\mu}_j^{RBP} -$

$\mu_j)^2]$. Then, we state a Jackknife procedure that accounts for the additional uncertainty resulting from model parameter estimation. Combining both steps ultimately allows for conservative MSE estimates.

3.1 MSE bound for the RBP

For the sake of a compact presentation, we assume that the random effect structure is limited to a random intercept, thus, $\mathbf{b}_j = b_j$ with $b_j \sim N(0, \psi^2)$. Note that there is a related MSE derivation for a general random effect structure in Appendix 8 of the supplemental material. Under the random intercept setting, the MSE of the RBP is given by

$$\begin{aligned} \text{MSE}(\hat{\mu}_j^{RBP}) &= E\left[\left(\hat{\mu}_j^{RBP}(\boldsymbol{\theta}) - \mu_j\right)^2\right] \\ &= E\left[\left((\bar{\mathbf{X}}_j + \bar{\mathbf{D}}_j)' \boldsymbol{\beta} + \gamma_j (\bar{y}_j - (\bar{\mathbf{x}}_j + \bar{\mathbf{d}}_j)' \boldsymbol{\beta}) - \bar{\mathbf{X}}_j' \boldsymbol{\beta} - b_j\right)^2\right] \\ &= E\left[\left(\bar{\mathbf{D}}_j' \boldsymbol{\beta} + \gamma_j (\bar{y}_j - (\bar{\mathbf{x}}_j + \bar{\mathbf{d}}_j)' \boldsymbol{\beta}) - (1 - \gamma_j)b_j\right)^2\right] \\ &= E\left[\left(\bar{\mathbf{D}}_j' \boldsymbol{\beta} + \gamma_j \bar{e}_j - (1 - \gamma_j)b_j\right)^2\right], \end{aligned} \tag{20}$$

where

$$\bar{e}_j = \frac{1}{n_j} \sum_{i \in \mathcal{S}_j} e_{ij}, \quad \bar{\mathbf{x}}_j = \frac{1}{n_j} \sum_{i \in \mathcal{S}_j} \mathbf{x}_{ij}, \quad \bar{\mathbf{d}}_j = \frac{1}{n_j} \sum_{i \in \mathcal{S}_j} \mathbf{d}_{ij}, \quad \gamma_j = \frac{\psi^2}{\psi^2 + \sigma^2/n_j}, \tag{21}$$

with $e_{ij} \stackrel{iid}{\sim} N(0, \sigma^2)$. Recall that b_j and \mathbf{e}_j are independent. Since $\bar{\mathbf{D}}_j$ is a fixed unknown quantity under the min-max setting, it follows that

$$\begin{aligned} \text{MSE}(\hat{\mu}_j^{RBP}) &= E\left[\left(\bar{\mathbf{D}}_j' \boldsymbol{\beta}\right)^2\right] + \gamma_j^2 E\left(\bar{e}_j^2\right) + (1 - \gamma_j)^2 E\left(b_j^2\right) \\ &= \left(\bar{\mathbf{D}}_j' \boldsymbol{\beta}\right)^2 + \gamma_j^2 \frac{\sigma^2}{n_j} + (1 - \gamma_j)^2 \psi^2 = \left(\bar{\mathbf{D}}_j' \boldsymbol{\beta}\right)^2 + \gamma_j \frac{\sigma^2}{n_j}. \end{aligned} \tag{22}$$

We see that MSE contains the term $(\bar{\mathbf{D}}_j' \boldsymbol{\beta})^2$, which is unknown due to the measurement errors being unobservable. Thus, even with all the model parameters known, we cannot calculate the exact value of (22) because $\bar{\mathbf{D}}_j$ is an unknown quantity under the considered setting. Yet, recall that for min-max robustification, we introduce perturbations $\boldsymbol{\Delta}$ to account for the uncertainty resulting from covariate contamination. Therefore, we can replace the term $(\bar{\mathbf{D}}_j' \boldsymbol{\beta})^2$ by a corresponding expression $(\boldsymbol{\Delta}'_j \boldsymbol{\beta})^2$, provided that $\lambda_1, \dots, \lambda_d$ are chosen sufficiently high. Here, $\boldsymbol{\Delta}'_j$ is the j -th row of $\boldsymbol{\Delta}$. As described in Sect. 2.1, the perturbations are element of an underlying uncertainty set \mathbb{U} , which depends on the regularization $\sum_{l=1}^d \lambda_l f_l(h_l(\boldsymbol{\beta}))$. If $\boldsymbol{\Delta} \in \mathbb{U}$, the

uncertainty set induces an upper bound $\sum_{l=1}^d \varphi_l h_l(\boldsymbol{\beta})$ on the total impact of the perturbations on model parameter estimation in accordance with Theorem 1. Provided that the loss function g is the squared ℓ_2 -norm, the total impact is measured by $\|\Delta\boldsymbol{\beta}\|_2$ in accordance with Theorem 1. From this argumentation, we can state the error bound

$$\left(\Delta'_j \boldsymbol{\beta}\right)^2 \leq \|\Delta\boldsymbol{\beta}\|_2^2 \leq \left(\sum_{l=1}^d \varphi_l h_l(\boldsymbol{\beta})\right)^2. \tag{23}$$

Substituting $(\bar{\mathbf{D}}'_j \boldsymbol{\beta})^2$ by $(\sum_{l=1}^d \varphi_l h_l(\boldsymbol{\beta}))^2$ in (22) subsequently yields an upper limit for the MSE. However, by Theorem 1, the error bound (23) depends on the chosen regularization. Naturally, it has to be determined for each uncertainty set individually. Here, we encounter the problem that the uncertainty set parameters $\varphi_1, \dots, \varphi_d$ are unknown. The one-to-one relation between the regularization parameters and the uncertainty set parameters in Proposition 1 is lost in Theorem 1. Hence, the values of the uncertainty set parameters have to be recovered first before corresponding error bounds can be used. In order to recover the uncertainty set parameters, we apply the following basic procedure. Assume that we have computed an optimal solution $\hat{\boldsymbol{\beta}}$ of a regularized regression problem

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} f\left(g\left(\mathbf{y} - \tilde{\mathbf{X}}\boldsymbol{\beta}\right)\right) + \sum_{l=1}^d \lambda_l f_l(h_l(\boldsymbol{\beta})), \tag{24}$$

in accordance with Theorem 1. By Theorem 1, $\hat{\boldsymbol{\beta}}$ is also an optimal solution of

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \max_{\Delta \in \mathbb{U}} g\left(\mathbf{y} - (\tilde{\mathbf{X}} + \Delta)\boldsymbol{\beta}\right) \tag{25}$$

for appropriate $\varphi_1 > 0, \dots, \varphi_d > 0$ forming the uncertainty set \mathbb{U} . Thus, by Proposition 1, we know that $\hat{\boldsymbol{\beta}}$ is an optimal solution of

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} g(\mathbf{y} - \tilde{\mathbf{X}}\boldsymbol{\beta}) + \sum_{l=1}^d \varphi_l h_l(\boldsymbol{\beta}). \tag{26}$$

Following this argumentation, we recover the relation by choosing $\varphi_1, \dots, \varphi_d$ in such a way that $\hat{\boldsymbol{\beta}}$ is an optimal solution of (26). For this, we need to derive the optimality conditions of (26) for a given specification of g and h . We then solve the arising system of equations for the uncertainty set parameters in dependence of the regularization parameters. This yields us the relation between $\varphi_1, \dots, \varphi_d$ and $\lambda_1, \dots, \lambda_d$. In what follows, we demonstrate the procedure for Ridge regression, the LASSO, and the Elastic Net. The obtained results are subsequently used to find upper bounds for the MSE stated in (22). Let $\mathbf{W} := \mathbf{V}^{-1/2} \tilde{\mathbf{X}}$ and denote the columns of \mathbf{W} by $\mathbf{W}^1, \dots, \mathbf{W}^p$. Further, define $\mathbf{v} := \mathbf{V}^{-1/2} \mathbf{y}$.

Error bound for Ridge regression

We have a single uncertainty set parameter φ and the optimization problem

$$\hat{\beta}^{\ell_2} = \operatorname{argmin}_{\beta \in \mathbb{R}^p} \left\| \mathbf{V}^{-1/2} (\mathbf{y} - \tilde{\mathbf{X}}\beta) \right\|_2^2 + \lambda \|\beta\|_2^2. \tag{27}$$

In this setting, φ can be recovered according to the subsequent proposition.

Proposition 4 *Let $\hat{\beta}^{\ell_2} \neq \mathbf{0}_p$ be an optimal solution to the optimization problem (27). Then, the related uncertainty set, as described in Theorem 1, is given by*

$$\mathbb{U}_{\ell_2} = \{ \mathbf{\Delta} \in \mathbb{R}^{n \times p} : \|\mathbf{\Delta}\mathbf{y}\|_2 \leq \varphi \|\mathbf{y}\|_2 \text{ for all } \mathbf{y} \in \mathbb{R}^p \}$$

with

$$\varphi = \frac{\|\hat{\beta}^{\ell_2}\|_2}{\|\mathbf{W}\hat{\beta}^{\ell_2} - \mathbf{v}\|_2} \lambda.$$

The proof can be found in Appendix 3 of the supplemental material. Observe that the uncertainty set parameter φ has a closed-form solution when $\mathbf{\Delta} \in \mathbb{U}_{\ell_2}$. We use the expression to substitute φ in the error bound (23). This allows us to state an upper bound for the MSE (22) when min–max robustification is achieved via Ridge regression. We obtain

$$\operatorname{MSE} \left(\hat{\mu}_j^{RBP}(\ell_2) \right) \leq \left[\left(\frac{\|\hat{\beta}^{\ell_2}\|_2}{\|\mathbf{W}\hat{\beta}^{\ell_2} - \mathbf{v}\|_2} \lambda \right) \|\hat{\beta}^{\ell_2}\|_2 \right]^2 + \gamma_j \frac{\sigma^2}{n_j}. \tag{28}$$

Error bound for the LASSO

We have a single uncertainty set parameter φ and the optimization problem

$$\hat{\beta}^{\ell_1} = \operatorname{argmin}_{\beta \in \mathbb{R}^p} \left\| \mathbf{V}^{-1/2} (\mathbf{y} - \tilde{\mathbf{X}}\beta) \right\|_2^2 + \lambda \|\beta\|_1. \tag{29}$$

Under this premise, φ can be recovered as follows.

Proposition 5 *Let $\hat{\beta}^{\ell_1} \neq \mathbf{0}_p$ be an optimal solution to the optimization problem (29). Then, the related uncertainty set, as described in Theorem 1, is given by*

$$\mathbb{U}_{\ell_1} = \{ \mathbf{\Delta} \in \mathbb{R}^{n \times p} : \|\mathbf{\Delta}\mathbf{y}\|_2 \leq \varphi \|\mathbf{y}\|_1 \text{ for all } \mathbf{y} \in \mathbb{R}^p \}$$

with

$$\varphi = \frac{\lambda}{\|\mathbf{W}\hat{\beta}^{\ell_1} - \mathbf{v}\|_2}.$$

The proof can be found in Appendix 4 of the supplemental material. We see that the uncertainty set parameter has a closed-form solution when $\mathbf{\Delta} \in \mathbb{U}_{\ell_1}$. We use the

expression to substitute φ in the error bound (23). This allows us to state an upper bound for the MSE (22) when min-max robustification is achieved via the LASSO. We obtain

$$\text{MSE}(\hat{\mu}_j^{RBP}(\ell_1)) \leq \left[\left(\frac{\lambda}{\|\mathbf{W}\hat{\beta}^{\ell_1} - \mathbf{v}\|_2} \right) \|\hat{\beta}^{\ell_1}\|_1 \right]^2 + \gamma_j \frac{\sigma^2}{n_j}. \tag{30}$$

Error bound for the Elastic Net

We have two uncertainty set parameters φ_1, φ_2 and the optimization problem

$$\hat{\beta}^{EN} = \underset{\beta \in \mathbb{R}^p}{\text{argmin}} \left\| \mathbf{V}^{-1/2} (\mathbf{y} - \tilde{\mathbf{X}}\beta) \right\|_2^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2. \tag{31}$$

They are recovered as demonstrated hereafter.

Proposition 6 *Let $\hat{\beta}^{EN} \neq \mathbf{0}_p$ be an optimal solution to the optimization problem (31). Then, the related uncertainty set, as described in Theorem 1, is given by*

$$\mathbb{U}_{EN} = \{ \Delta \in \mathbb{R}^{n \times p} : \|\Delta\gamma\|_2 \leq \varphi_1 \|\gamma\|_1 + \varphi_2 \|\gamma\|_2 \text{ for all } \gamma \in \mathbb{R}^p \}$$

with φ_1, φ_2 being a solution of the system

$$\|\mathbf{W}\hat{\beta}^{EN} - \mathbf{v}\|_2^{-1} (\lambda_1 \mathbf{1}_p + \lambda_2 \hat{\beta}^{EN}) = \varphi_1 \mathbf{1}_p + \varphi_2 \|\hat{\beta}^{EN}\|_2^{-1} \hat{\beta}^{EN}.$$

The proof can be found in Appendix 5 of the supplemental material. Note that the term $\mathbf{1}_p$ marks a column vector of p ones. We see that the uncertainty set parameters φ_1, φ_2 do not have a closed-form solution. They can be quantified numerically, for instance by applying the Moore–Penrose inverse. For a given robust estimation problem with optimal solution $\hat{\beta}^{EN}$, let φ_1^*, φ_2^* be the solutions for the uncertainty set parameters. Plugging them into the MSE equation (22), we obtain the following result when min-max robustification is achieved via the Elastic Net:

$$\begin{aligned} \text{MSE}(\hat{\mu}_j^{RBP}(EN)) &\leq (\varphi_1^* \|\hat{\beta}^{EN}\|_1)^2 + 2\varphi_1^* \varphi_2^* \|\hat{\beta}^{EN}\|_1 \|\hat{\beta}^{EN}\|_2 \\ &\quad + (\varphi_2^* \|\hat{\beta}^{EN}\|_2)^2 + \gamma_j \frac{\sigma^2}{n_j}. \end{aligned} \tag{32}$$

3.2 Conservative Jackknife estimator

We now use the MSE bounds (28), (30), and (32) for the RBP to construct a conservative Jackknife estimator for the MSE of the REBP. For this, we rely on theoretical developments presented by Jiang et al. (2002). The term "conservative" stems from the fact that we do not use the actual MSE, but their upper bounds in accordance with

Theorem 1. With this, we do not obtain estimates of $MSE(\hat{\mu}_j^{REBP})$ in the classical sense. Instead, an upper bound for the measure is estimated under a given level of uncertainty resulting from unobservable measurement errors. We refer to this as pessimistic MSE (PMSE) estimation. A delete-1-Jackknife procedure is applied. In every iteration of the algorithm, a domain-specific subsample \mathcal{S}_j is deleted from the data base. The remaining observations are used to perform model parameter estimation. Based on the obtained estimates, predictions for μ_j in all domains $\mathcal{U}_1, \dots, \mathcal{U}_m$ are produced. The principle is repeated until all domain-specific subsamples have been deleted once. With this resampling scheme, we obtain an approximation to the prediction uncertainty resulting from model parameter estimation (Jiang et al. 2002; Burgard et al. 2020a). The deviation of the produced predictions from the original predictions based on all sample observations are quantified. Finally, in combination with the MSE bounds, this yields an estimate of the PMSE.

Let $B(\varphi_1, \dots, \varphi_d, h_1, \dots, h_d, \hat{\theta})$ denote a general MSE bound in the sense of Sect. 3.1, where the known model parameter estimates θ have been replaced by empirical estimates $\hat{\theta}$. Let $\hat{\theta}_{-k}$ be the estimate of θ without the observations from \mathcal{S}_k , where $\mathcal{S}_k \subset \mathcal{S}$. Likewise, denote $\hat{\mu}_j^{REBP}(\hat{\theta}_{-k})$ as the REBP of μ_j based on $\hat{\theta}_{-k}$, where $j, k \in \{1, \dots, m\}$. The delete-1-Jackknife procedure is performed as follows.

Algorithm 2 Delete-1-Jackknife

- 1: Obtain model parameter estimates $\hat{\theta}$ based on \mathcal{S} .
 - 2: Generate predictions $\hat{\mu}_1^{REBP}(\hat{\theta}), \dots, \hat{\mu}_m^{REBP}(\hat{\theta})$ according to (19).
 - 3: Calculate $B(\varphi_1, \dots, \varphi_d, h_1, \dots, h_d, \hat{\theta})$.
 - 4: **for** $k = 1, \dots, m$ **do**
 - 5: Obtain model parameter estimates $\hat{\theta}_{-k}$ based on $\mathcal{S} \setminus \mathcal{S}_k$.
 - 6: Generate predictions $\hat{\mu}_1^{REBP}(\hat{\theta}_{-k}), \dots, \hat{\mu}_m^{REBP}(\hat{\theta}_{-k})$ according to (19).
 - 7: Calculate $B(\varphi_1, \dots, \varphi_d, h_1, \dots, h_d, \hat{\theta}_{-k})$.
 - 8: **end**
-

After the algorithm is completed, the conservative Jackknife estimator for the REBP of μ_j is calculated according to

$$\widehat{PMSE}(\hat{\mu}_j^{REBP}) = B(\varphi_1, \dots, \varphi_d, h_1, \dots, h_d, \hat{\theta}) + \frac{m-1}{m} \sum_{k=1}^m \left(\hat{\mu}_j^{REBP}(\hat{\theta}_{-k}) - \hat{\mu}_j^{REBP}(\hat{\theta}) \right)^2 - \frac{m-1}{m} \sum_{k=1}^m \left(B(\varphi_1, \dots, \varphi_d, h_1, \dots, h_d, \hat{\theta}) - B(\varphi_1, \dots, \varphi_d, h_1, \dots, h_d, \hat{\theta}_{-k}) \right). \tag{33}$$

4 Consistency

Hereafter, we study conditions under which min-max robustification as described in Sect. 2.1 allows for consistency in model parameter estimation. We adapt the theoretical framework developed by Fan and Li (2001) as well as Ghosh and Thoresen (2018) for the asymptotic behavior of regularized regression with non-concave penalties. Based on the insights of Sect. 2, we introduce min–max robustification to their developments. For this purpose, let us state the optimization problem for model parameter estimation as follows:

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmin}} \underbrace{-L(\boldsymbol{\beta}, \boldsymbol{\eta}) + n \sum_{l=1}^d \lambda_n^l \sum_{k=1}^p h_k^l(\beta_k)}_{\mathcal{Q}_\theta}, \tag{34}$$

with $\Theta \subseteq \mathbb{R}^p \times (0, \infty)^{1+q^*}$ as parameter space, $\mathcal{Q}_\theta : \Theta \rightarrow \mathbb{R}$ as objective function, and

$$L(\boldsymbol{\beta}, \boldsymbol{\eta}) = -\frac{1}{2} \left[n \log(2\pi) + \log(|\mathbf{V}|) + (\mathbf{y} - \tilde{\mathbf{X}}\boldsymbol{\beta})' \mathbf{V}^{-1} (\mathbf{y} - \tilde{\mathbf{X}}\boldsymbol{\beta}) \right] \tag{35}$$

denoting the negative log-likelihood function of (16). The superscript $l = 1, \dots, d$ marks the index of regularization terms and $k = 1, \dots, p$ is the index of fixed effect coefficients. Thus, the term $n \sum_{l=1}^d \lambda_n^l \sum_{k=1}^p h_k^l(\beta_k)$ is a componentwise non-concave regularization with parameters $\lambda_n^1, \dots, \lambda_n^d$ depending on the sample size. For each parameter λ_n^l , we have a family of increasing, convex and non-concave functions h_1^l, \dots, h_p^l with $h_k^l(\beta_k) : \mathbb{R} \rightarrow \mathbb{R}_+$. This componentwise notation is required to establish consistency for a robust estimator $\hat{\boldsymbol{\beta}}$ that is potentially sparse. Further, following Fan and Li (2001) as well as Ghosh and Thoresen (2018), we let the regularization directly depend on n via multiplication. Note that this does not affect min–max robustification presented in Theorem 1. For a given regularized regression problem, we can substitute a predefined parameter value λ_l with an equivalent term $n\tilde{\lambda}_l$ by choosing $\tilde{\lambda}_l = \lambda_l/n$.

4.1 Asymptotic perturbation behavior

A necessary condition for consistency is that $\lambda_n^l \rightarrow 0$ as $n \rightarrow \infty$ for all $l = 1, \dots, d$. In the classical context of regularized regression, this is a reasonable assumption. Increasing the sample size solves typical issues for which regularization is applied, such as the model being not identifiable or rank deficiency of the design matrix. However, in the light of robustification, it also implies that the impact of the measurement errors on model parameter estimation has to approach zero. Let us restate the uncertainty set

from Theorem 1 in the asymptotic setting for $g = \ell_2$ according to

$$\mathbb{U}_n = \left\{ \Delta \in \mathbb{R}^{n \times p} : \frac{1}{n} \|\Delta \boldsymbol{\gamma}\|_2 \leq \frac{1}{n} \sum_{l=1}^d \varphi^l \sum_{k=1}^p h_k^l(\gamma_k) \text{ for all } \boldsymbol{\gamma} \in \mathbb{R}^p \right\}.$$

With this formulation, the mean impact of the measurement errors on model parameter estimation is constrained by the regularization. Yet, if we draw new observations, it is not guaranteed that the mean impact will approach zero for $n \rightarrow \infty$. This is demonstrated hereafter. Without loss of generality, assume that the sample size in the asymptotic setting is increased in terms of sequential draws denoted by $r = 1, 2, \dots$. We start with a set of $n_r^{init} > 0$ initial sample observations in the r -th draw with $r = 1$. The perturbations resembling the uncertainty in the corresponding covariate observations are represented by the perturbation matrix $\Delta_r^{init} \in \mathbb{R}^{n_r^{init} \times p}$. Next, we draw $n_r^{new} > 0$ new observations and pool them with the initial sample observations. The perturbation matrix for these new observations is $\Delta_r^{new} \in \mathbb{R}^{n_r^{new} \times p}$. In the next draw $r + 1 = 2$, the previously pooled observations represent the initial ones, such that $n_{r+1}^{init} = n_r^{init} + n_r^{new}$. This is repeated for $r \rightarrow \infty$, implying that $n_{r+1}^{init} \rightarrow \infty$. Note that the limit

$$\lim_{r \rightarrow \infty} \frac{1}{n_r^{init} + n_r^{new}} \left\| \left(\Delta_r^{init'}, \Delta_r^{new'} \right)' \boldsymbol{\gamma} \right\|_2 \tag{36}$$

is not necessarily zero, provided that $\boldsymbol{\gamma} \neq \mathbf{0}_p$. This, however, is imperative if $\lambda_n^l \rightarrow 0$ for all $l = 1, \dots, d$ while simultaneously ensuring a robust solution in the sense of Theorem 1. At this point, we can conclude that min–max robust model parameter estimation within the robust LMM is not consistent for arbitrary design matrix perturbations, since the impact of the measurement error does not vanish. In order to guarantee that the mean impact of the measurement errors approaches zero, we have to introduce further assumptions on the asymptotic perturbation behavior. These assumptions are with respect to the magnitude of the perturbations. We explicitly avoid assumptions regarding their distribution, as the main feature of min–max robustification is the absence of distribution assumptions. The required behavior in terms of sequential draws is characterized by the subsequent lemma.

Lemma 1 *Let $\Delta_r^{init} \in \mathbb{R}^{n_r^{init} \times p}$ be an initial perturbation matrix in the r -th draw. Assume that a new set of observations is drawn with perturbation matrix $\Delta_r^{new} \in \mathbb{R}^{n_r^{new} \times p}$. If for repeated draws $r = 1, 2, \dots$ and a given $\boldsymbol{\gamma} \in \mathbb{R}^p$ every new perturbation matrix satisfies*

$$\|\Delta_r^{new} \boldsymbol{\gamma}\|_2 < \frac{\sqrt{n_r^{new} (2n_r^{init} + n_r^{new})}}{n_r^{init}} \|\Delta_r^{init} \boldsymbol{\gamma}\|_2,$$

then

$$\lim_{r \rightarrow \infty} \frac{1}{n_r^{init} + n_r^{new}} \left\| \left(\Delta_r^{init'}, \Delta_r^{new'} \right)' \boldsymbol{\gamma} \right\|_2 = 0.$$

The proof can be found in Appendix 6 of the supplemental material. In practice, the behavior stated in Lemma 1 may be viewed as the measurement process becoming more accurate over time, or the number of contaminated observations rising at a smaller rate than the number of correct observations.

4.2 Asymptotic results

For illustrative purposes, let θ^* denote the true value of the model parameter vector. Consistency is studied by investigating the asymptotic behavior of $\|\hat{\theta} - \theta^*\|_2$ as $n \rightarrow \infty$. We consider a deterministic design matrix setting with a fixed number of covariates $(p + 1 + q^*) < n$. With this, we assume that model parameter estimation is a low-dimensional problem. The asymptotics of regularized regression are usually studied in high-dimensional settings. However, the focus of our contribution is on robust estimation rather than on high-dimensional inference. Therefore, we restrict the analysis to the simpler case of low-dimensional problems. See Schelldorfer et al. (2011), Shao and Deng (2012), as well as Loh and Wainwright (2012) for theoretical results on high-dimensional inference.

In what follows, we draw from the theoretical framework presented by Fan and Li (2001) as well as Ghosh and Thoresen (2018). Several assumptions are introduced that are required in order to establish consistency. For simplicity, and by the centering of the response observations as well as the standardization of the covariate values as displayed in Algorithm 1, we assume that the domain response vectors $y_j \in \mathbf{y}$ are iid. Thus, we can state the negative log-likelihood

$$-L(\boldsymbol{\beta}, \boldsymbol{\eta}) = L_n(\boldsymbol{\beta}, \boldsymbol{\eta}) = \sum_{j=1}^m \log [f(\mathbf{y}_j; \boldsymbol{\beta}, \boldsymbol{\eta})]. \tag{37}$$

See that $L_n(\boldsymbol{\beta}, \boldsymbol{\eta})$ is convex in $\boldsymbol{\beta}$ and non-convex in $\boldsymbol{\eta}$. Let $\lambda_1, \dots, \lambda_d \rightarrow 0$ as $n \rightarrow \infty$. Define

$$A_n := \max_{k=1, \dots, p} \left\{ \sum_{l=1}^d \lambda_n^l \frac{\partial h_k^l(|\beta_k^*|)}{\partial \beta_k} : \beta_k^* \neq 0 \right\},$$

$$B_n := \max_{k=1, \dots, p} \left\{ \sum_{l=1}^d \lambda_n^l \frac{\partial^2 h_k^l(|\beta_k^*|)}{\partial \beta_k^2} : \beta_k^* \neq 0 \right\} \tag{38}$$

as the maximum values of the first- and second-order derivatives with respect to all non-zero elements of the regression coefficient vector at the true point $\boldsymbol{\beta}^*$.

Assumption 1

The model is identifiable and the support of $f(\mathbf{y}; \boldsymbol{\beta}, \boldsymbol{\eta})$ is independent of the parameter $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\eta}')$. The density $f(\mathbf{y}; \boldsymbol{\beta}, \boldsymbol{\eta})$ has first- and second-order derivatives with

$$E \left(\frac{\partial \log(f)}{\partial (\boldsymbol{\beta}, \boldsymbol{\eta})} \right) = \mathbf{0}_{p+q^*+1} \tag{39}$$

and

$$\mathcal{I}(\boldsymbol{\beta}, \boldsymbol{\eta}) = \mathbb{E} \left[\left(\frac{\partial \log(f)}{\partial(\boldsymbol{\beta}, \boldsymbol{\eta})} \right)' \left(\frac{\partial \log(f)}{\partial(\boldsymbol{\beta}, \boldsymbol{\eta})} \right) \right] = \mathbb{E} \left[\left(\frac{\partial^2 \log(f)}{\partial(\boldsymbol{\beta}, \boldsymbol{\eta})^2} \right) \right]. \tag{40}$$

Assumption 2

The Fisher information matrix $\mathcal{I}(\boldsymbol{\beta}, \boldsymbol{\eta})$ is finite and positive definite at $(\boldsymbol{\beta}^*, \boldsymbol{\eta}^*)$.

Assumption 3

There exists an open subset of Θ containing $(\boldsymbol{\beta}^*, \boldsymbol{\eta}^*)$, on which $f(\mathbf{y}; \boldsymbol{\beta}, \boldsymbol{\eta})$ admits all its third-order partial derivatives for almost all \mathbf{y} that are uniformly bounded by some function with finite expectation under the true value of the full parameter vector.

Assumption 4

Regarding the regularization term, $B_n \rightarrow 0$ as $n \rightarrow \infty$.

Assumption 5

With respect to the perturbations, $n^{-1} \|\Delta \boldsymbol{\beta}\| \leq n^{-1} \sum_{l=1}^d \varphi^l \sum_{k=1}^p h_k^l(\beta_k)$ for any $\boldsymbol{\theta} \in \Theta$ with $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\eta}')$, and $\lim_{n \rightarrow \infty} \|\Delta \boldsymbol{\beta}\|_2 = 0$.

Assumptions 1 to 3 are basic regularity conditions on ML estimation problems and satisfied by the majority of common statistical models. Assumption 4 is a requirement for non-concave regularizations to ensure that the objective value difference between a local minimizer and the true parameter value approaches zero asymptotically. Since we have assumed that $\lambda \rightarrow 0$ as $n \rightarrow \infty$, this holds for all considered regularizations. Assumption 5 is a technical requirement that can be viewed as the number of contaminated observations rising at a slower rate than the correctly measured observations. Based on the presented system of assumptions, the following theorem can be stated.

Theorem 2 *Consider the model parameter estimation problem (34) under the Assumptions 1–3. If $\lambda_n^l \rightarrow 0$ for $n \rightarrow \infty$ and $l = 1, \dots, d$, $\sum_{l=1}^d \lambda_n^l \sum_{k=1}^p h_k^l(\beta_k)$ satisfies Assumption 4, and the perturbations obey Assumption 5, then there exists a local minimizer $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}', \hat{\boldsymbol{\eta}}')$ of \mathcal{Q}_θ , for which*

$$\left\| \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \right\|_2 = \mathcal{O}_p(n^{-1/2} + A_n), \quad \left\| \hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^* \right\|_2 = \mathcal{O}_p(n^{-1/2})$$

holds.

The proof can be found in Appendix 7 of the supplemental material.

5 Simulation

5.1 Set up

A Monte Carlo simulation with $R = 500$ iterations indexed by $r = 1, \dots, R$ is conducted. We use a deterministic design matrix setting and create a synthetic population \mathcal{U} of $N = 50\,000$ individuals in $m = 100$ equally sized domains with $N_j = 500$. A random subset $\mathcal{S} \subset \mathcal{U}$ of size $n = 500$ is drawn once such that $n_j = 5$ for $j = 1, \dots, m$. The response variable realizations are generated in each iteration individually. The overall specifications are

$$y_{ij} = \mathbf{x}_{ij}\boldsymbol{\beta} + z_j + e_{ij}, \quad \mathbf{x}_{ij} \stackrel{iid}{\sim} N(\boldsymbol{\mu}_X, \sigma_X^2 \mathbf{I}_3), \quad z_j \stackrel{iid}{\sim} N(0, \psi^2), \quad e_{ij} \stackrel{iid}{\sim} N(0, \sigma^2),$$

where $\boldsymbol{\mu}_X = (2, 2, 2)'$, $\sigma_X^2 = 1$, $\boldsymbol{\beta} = (2, 2, 2)'$, and $\psi^2 = \sigma^2 = 1$. Since we consider a deterministic design matrix setting, \mathbf{x}_{ij} is generated once for each $i \in \mathcal{U}$, and then held fixed over all Monte Carlo iterations. The random components z_j and e_{ij} are drawn from their respective distributions in each iteration individually. The objective is to estimate the domain mean of Y as given in (12) based on \mathcal{S} . For the sample observations, we simulate erroneous covariate observations in the sense that only $\tilde{\mathbf{x}}_{ij} = \mathbf{x}_{ij} + \mathbf{d}_{ij}$ is observed for all $i \in \mathcal{S}$ rather than \mathbf{x}_{ij} . Regarding the covariate measurement error vector $\mathbf{d}_{ij} = (d_{ij}^1, d_{ij}^2, d_{ij}^3)$, we once again follow the deterministic design matrix setting. Thus, we generate \mathbf{d}_{ij} once for each $i \in \mathcal{U}$, and then held fixed over all Monte Carlo iterations. However, we implement four different scenarios with respect to \mathbf{d}_{ij} to evaluate the benefits of min-max robustification under different data constellations. For $j = 1, \dots, m, i = 1, \dots, N_j$, the scenarios are

- **Scenario 1:** no measurement errors $\mathbf{d}_{ij} = \mathbf{0}_{1 \times 3}$.
- **Scenario 2:** weakly correlated symmetric measurement errors $\mathbf{d}_{ij} \stackrel{iid}{\sim} N(\mathbf{0}_3, \boldsymbol{\Sigma}_D^{[2]})$.
- **Scenario 3:** strongly correlated symmetric measurement errors $\mathbf{d}_{ij} \stackrel{iid}{\sim} N(\mathbf{0}_3, \boldsymbol{\Sigma}_D^{[3]})$.
- **Scenario 4:** weakly correlated asymmetric measurement errors $\mathbf{d}_{ij} \stackrel{iid}{\sim} F(\mathbf{0}_3)$.

Here, $\boldsymbol{\Sigma}_D^{[2]}$ and $\boldsymbol{\Sigma}_D^{[3]}$ are scenario-specific covariance matrices that are given by

$$\boldsymbol{\Sigma}_D^{[2]} = \begin{pmatrix} 0.500 & 0.051 & 0.091 \\ 0.051 & 0.500 & 0.059 \\ 0.091 & 0.059 & 0.500 \end{pmatrix}, \quad \boldsymbol{\Sigma}_D^{[3]} = \begin{pmatrix} 0.500 & 0.267 & 0.234 \\ 0.267 & 0.500 & 0.290 \\ 0.234 & 0.290 & 0.500 \end{pmatrix}$$

Further, $F(\mathbf{0}_3)$ represents some distribution with expectation zero. For scenario 4, a given measurement error realization $d_{ij}^k \in \mathbf{d}_{ij}$ follows a transformed χ^2 -distribution by setting

$$d_{ij}^k = 0.5z_{ij}^k - \bar{z}, \quad z_{ij}^k \stackrel{iid}{\sim} \chi^2(1.2), \quad \bar{z} = \frac{1}{3n} \sum_{j=1}^m \sum_{i \in \mathcal{S}_j} \sum_{k=1}^3 0.5z_{ij}^k, \quad k = 1, 2, 3.$$

Under each setting, we consider the following prediction and estimation methods:

- **EBLUP/ML:** empirical best linear unbiased predictor under the basic LMM, ML estimation (Pinheiro and Bates 2000).
- **L1.REBP/L1.ML:** REBP under the robust LMM, ℓ_1 -regularized ML estimation.
- **L2.REBP/L2.ML:** REBP under the robust LMM, ℓ_2^2 -regularized ML estimation.
- **EN.REBP/EN.ML:** REBP under the robust LMM, Elastic Net-regularized ML estimation.

We use several performance measures for the evaluation of domain mean prediction:

$$\begin{aligned} \text{Bias}(\hat{\mu}) &= \frac{1}{mR} \sum_{j=1}^m \sum_{r=1}^R (\hat{\mu}_j^{(r)} - \bar{Y}_j^{(r)}), & \text{MSE}(\hat{\mu}) &= \frac{1}{mR} \sum_{j=1}^m \sum_{r=1}^R (\hat{\mu}_j^{(r)} - \bar{Y}_j^{(r)})^2, \\ \text{RBias}(\hat{\mu}) &= \frac{1}{mR} \sum_{j=1}^m \sum_{r=1}^R \frac{\hat{\mu}_j^{(r)} - \bar{Y}_j^{(r)}}{R^{-1} \sum_{r=1}^R \bar{Y}_j^{(r)}}, & \text{RRMSE}(\hat{\mu}) &= \frac{1}{mR} \sum_{j=1}^m \sum_{r=1}^R \frac{\sqrt{(\hat{\mu}_j^{(r)} - \bar{Y}_j^{(r)})^2}}{R^{-1} \sum_{r=1}^R \bar{Y}_j^{(r)}}. \end{aligned}$$

For model parameter estimation, let $k = 1, \dots, p + q^* + 1$ be the index of all model parameters. For each $\hat{\theta}_k \in \hat{\theta}$, we consider the following performance measures:

$$\text{Bias}(\hat{\theta}_k) = \frac{1}{R} \sum_{r=1}^R (\hat{\theta}_k^{(r)} - \theta_k), \quad \text{MSE}(\hat{\theta}_k) = \frac{1}{R} \sum_{r=1}^R (\hat{\theta}_k^{(r)} - \theta_k)^2.$$

Regarding MSE estimation, define $\text{Mean}(\widehat{\text{PMSE}}(\hat{\mu})) = (mR)^{-1} \sum_{j=1}^m \sum_{r=1}^R \widehat{\text{PMSE}}(\hat{\mu}_j^{(r)})$. We look at the relative bias as well as the coefficient of variation (CV):

$$\begin{aligned} \text{RBias}(\widehat{\text{PMSE}}(\hat{\mu})) &= \frac{1}{mR} \sum_{j=1}^m \sum_{r=1}^R \frac{\widehat{\text{PMSE}}(\hat{\mu}_j^{(r)}) - \text{MSE}(\hat{\mu})}{\text{MSE}(\hat{\mu})}, \\ \text{CV}(\widehat{\text{PMSE}}(\hat{\mu})) &= \frac{\sqrt{(mR)^{-1} \sum_{j=1}^m \sum_{r=1}^R (\widehat{\text{PMSE}}(\hat{\mu}_j^{(r)}) - \text{Mean}(\widehat{\text{PMSE}}(\hat{\mu})))^2}}{\text{Mean}(\widehat{\text{PMSE}}(\hat{\mu}))} \end{aligned}$$

We use the R package *glmnet* for the optimization problem (17) to obtain fixed effect estimates. The tuning of the regularization parameters required for the optimization problem (17) is performed via standard cross validation using a parameter grid of 1000 candidate values. Further, we use the R package *nloptr* for the maximum likelihood problem (18) to obtain variance parameter estimates.

5.2 Results

We start with the results for domain mean prediction. They are summarized in Table 1 and further visualized in Fig. 1. In the latter, the densities of $(\hat{\mu}_j^{(r)} - \bar{Y}_j^{(r)})/\bar{Y}_j^{(r)}$ are plotted for the EBLUP and the L2.REBP. From Table 1, it can be seen that in Scenario 1 (absence of measurement errors) the unregularized EBLUP and the regularized predictors have very similar performance. This is due to the fact that the optimal regularization parameter found by cross validation is very close to zero in the absence of measurement errors. A different picture arises in the presence of measurement errors, thus in Scenario 2 to Scenario 4. Here, the regularized predictors are much more efficient than the unregularized EBLUP.

The efficiency advantage in terms of the MSE ranges from 37% to 47%. Observe that the advantage through min–max robustification is even evident for asymmetric errors. Thus underlines the theoretical result that we do not require any distribution assumptions on the error. Another interesting aspect is that min-max robustification

Table 1 Simulation results of domain mean prediction

Predictor	Scenario	MSE	RRMSE	Bias	RBias
EBLUP	1	0.16902	0.03428	0.00334	0.00028
L1.REBP	1	0.16857	0.03423	- 0.01145	- 0.00095
L2.REBP	1	0.16856	0.03423	- 0.01084	- 0.00090
EN.REBP	1	0.16857	0.03423	- 0.01119	- 0.00093
EBLUP	2	1.74700	0.11020	0.06182	0.00515
L1.REBP	2	0.94744	0.08115	- 0.53693	- 0.04477
L2.REBP	2	0.94211	0.08092	- 0.53179	- 0.04434
EN.REBP	2	0.94477	0.08104	- 0.53436	- 0.04455
EBLUP	3	2.03764	0.11901	- 0.18649	- 0.01555
L1.REBP	3	1.29266	0.09479	- 0.75264	- 0.06275
L2.REBP	3	1.28384	0.09447	- 0.74665	- 0.06225
EN.REBP	3	1.28824	0.09463	- 0.74964	- 0.06250
EBLUP	4	1.64817	0.10704	0.09823	0.00819
L1.REBP	4	0.87170	0.07784	- 0.47170	- 0.03933
L2.REBP	4	0.86729	0.07764	- 0.46687	- 0.03892
EN.REBP	4	0.86949	0.07774	- 0.46928	- 0.03913

leads to a bias in domain mean prediction. In general, this was expected, since it is well-known that regularization introduces bias to model parameter estimation (Hoerl and Kennard 1970). However, by looking at Fig. 1, we see that the bias increases in the presence of measurement errors. This is due to the fact that the optimal regularization parameter found by cross validation is larger in Scenario 2 to 4 compared to Scenario 1. This is an important aspect, as it implies that cross validation is sensible with respect to measurement errors. The robustness argument provided in Theorem 1 relies on the assumption that $\lambda_1, \dots, \lambda_d$ is chosen sufficiently high. Although in practice, we never have a guarantee that the assumption is satisfied, this observation suggests that cross validation is capable of finding $\lambda_1, \dots, \lambda_d$ that at least approximate the level of uncertainty in $\tilde{\mathbf{X}}$.

We continue with the results for model parameter estimation. They are summarized in Table 2 and further visualized in Fig. 2. For Table 2, note that since $\beta_1 = \beta_2 = \beta_3 = 2$, we pool the estimates $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$ for the calculation of the performance measures and summarize the results into the columns $\text{MSE}(\boldsymbol{\beta})$ and $\text{Bias}(\boldsymbol{\beta})$. For the variance parameters, we proceed accordingly. Since $\sigma^2 = \psi^2 = 1$, we pool the estimates and summarize the results into the columns $\text{MSE}(\boldsymbol{\eta})$ and $\text{Bias}(\boldsymbol{\eta})$. Likewise, in Fig. 2, the variance parameter deviation $\hat{\eta}_k^{(r)} - \eta_k$ is plotted for all considered methods and both parameters simultaneously. With respect to $\boldsymbol{\beta}$ -estimation, the unregularized approach is slightly more efficient than the regularized methods for Scenario 1. This is because regularization introduces bias to the estimates, as pointed out by Hoerl and Kennard (1970). In the presence of measurement errors, min-max robustification obtains much more efficient results. The advantage ranges from 36 to 72%. We further see that the additional noise in the covariate data affects the unregularized approach considerably,

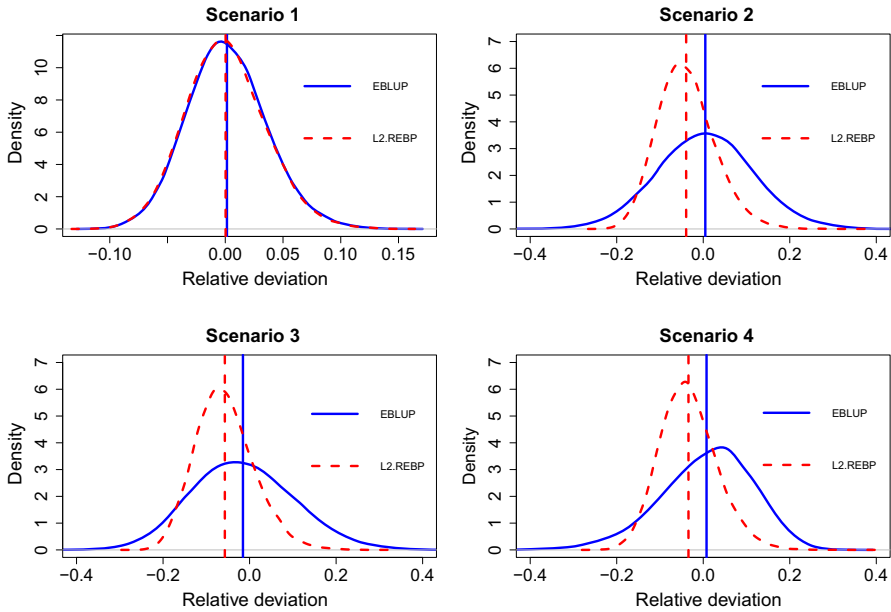


Fig. 1 Relative deviation of domain mean prediction for EBLUP and L2.REBLUP

which leads to a negative bias. Regularization, on the other hand, manages to reduce the bias in this setting as the estimates are less influenced by the contamination. For η -estimation, the benefits of min–max robustification are also visible. We see that the variance parameter estimates based on the unregularized β -estimates are severely biased upwards. This suggests that the additional noise in the design matrix is falsely interpreted as increased error term variance σ^2 . The results based on the robust β -estimates are less biased. By looking at Fig. 2, it further becomes evident that min–max robustification allows for stable variance parameter estimates. The overall spread of the estimates is much more narrow than for the unregularized approach. In terms of the MSE, the advantage of regularized estimation over unregularized estimation ranges from 15 to 31%.

We finally look at the results for (pessimistic) MSE estimation. They are presented in Table 3. For the absence of measurement errors, the resulting estimates are approximately unbiased. This makes sense given the fact that the optimal regularization parameter found by cross validation on correctly observed data is close to zero. In this setting, the modified Jackknife procedure summarized in Algorithm 2 basically reduces to a standard Jackknife, which is known to be suitable for MSE estimation in mixed models (Jiang et al. 2002). In Scenario 2 to 4, the proposed method leads to biased estimates of the true MSE. The relative overestimation ranges from 68 to 90%. However, this was expected in light of the theoretical results from Chapter 3. The obtained estimates are based on the MSE bounds (28), (30), and (32). These bounds were used in order to find upper limits of the true MSE given the fact that covariate observations are contaminated by unknown errors. If the bounds are plugged into

Table 2 Simulation results for model parameter estimation

Estimator	Scenario	MSE(β)	Bias(β)	MSE(η)	Bias(η)
EBLUP	1	0.0019	0.0001	0.0166	- 0.0086
L1.REBP	1	0.0022	- 0.0133	0.0216	- 0.0046
L2.REBP	1	0.0021	- 0.0126	0.0216	- 0.0047
EN.REBP	1	0.0022	- 0.0130	0.0216	- 0.0046
EBLUP	2	0.0454	- 0.1852	22.1121	3.8628
L1.REBP	2	0.0240	- 0.1213	18.8294	2.8012
L2.REBP	2	0.0236	- 0.1201	18.8518	2.8030
EN.REBP	2	0.0238	- 0.1207	18.8406	2.8021
EBLUP	3	0.1364	- 0.3291	49.3853	6.1500
L1.REBP	3	0.0391	- 0.1594	34.0506	3.8542
L2.REBP	3	0.0385	- 0.1581	34.0981	3.8571
EN.REBP	3	0.0388	- 0.1587	34.0743	3.8557
EBLUP	4	0.0412	- 0.1615	19.2715	3.5224
L1.REBP	4	0.0266	- 0.1099	16.3922	2.5893
L2.REBP	4	0.0263	- 0.1088	16.4106	2.5909
EN.REBP	4	0.0264	- 0.1093	16.4013	2.5901

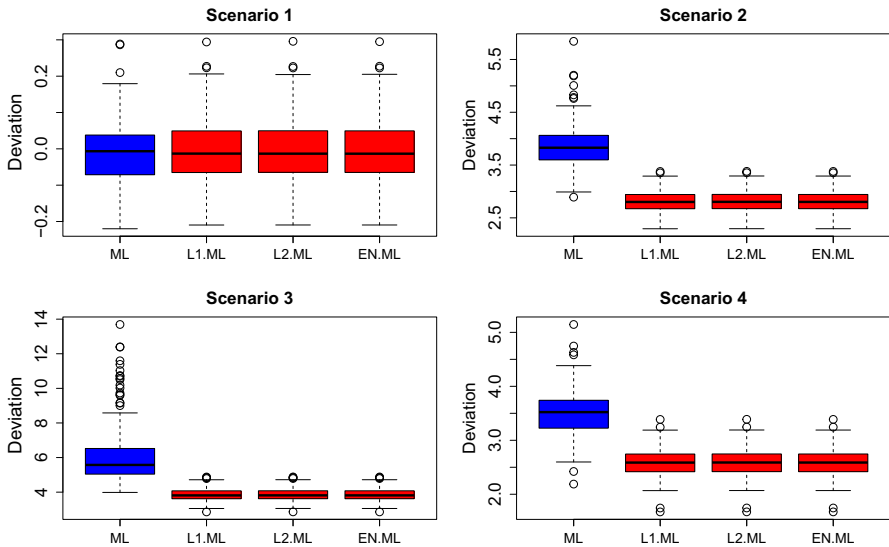


Fig. 2 Deviation of variance parameter estimation

Algorithm 2, the resulting estimator (33) will always overestimate the true MSE for $\lambda_1, \dots, \lambda_d$ chosen appropriately large. This is due to the nature of min-max robustification characterized in Proposition 1. It introduces design matrix perturbations that are maximal with respect to the underlying uncertainty set. Thus, the obtained MSE estimates are based on the premise that the measurement error associated with a given

Table 3 Simulation results for MSE estimation

Predictor	Scenario	Mean	RBias	CV
L1.REBP	1	0.16891	0.00202	0.05216
L2.REBP	1	0.16908	0.00309	0.05454
EN.REBP	1	0.16591	0.01577	0.07066
L1.REBP	2	1.75497	0.85233	1.09438
L2.REBP	2	1.75705	0.86500	1.09444
EN.REBP	2	1.75255	0.85501	1.09657
L1.REBP	3	2.17723	0.68429	1.13312
L2.REBP	3	2.18005	0.69807	1.13322
EN.REBP	3	2.17545	0.68869	1.13483
L1.REBP	4	1.64999	0.89284	1.06127
L2.REBP	4	1.65184	0.90460	1.06130
EN.REBP	4	1.64736	0.89464	1.06357

prediction is potentially maximal with respect to \mathbb{U} . If the distribution of the covariate measurement errors would be known, then it would be possible to find more accurate MSE bounds. See Loh and Wainwright (2012) for a corresponding theoretical analysis. However, we do not consider this case here, as our main focus is to achieve robustness without knowledge of the measurement error.

6 Conclusion

The presented paper addressed the connection between regularization and min-max robustification for linear mixed models in the presence of unobserved covariate measurement errors. It was demonstrated that min-max robustification represents a powerful tool to obtain reliable model parameter estimates and predictions when the data basis is contaminated by unknown errors. We showed that this approach to robust estimation is equivalent to regularized model parameter estimation for a broad range of penalties. These insights were subsequently used to construct a robust version of the basic LMM to perform regression analysis on contaminated data sets. We derived robust best predictors under the model and presented a novel Jackknife algorithm for conservative mean squared error estimation with respect to response predictions. The methodology allows for reliable uncertainty measurement and does not require any distribution assumptions regarding the error. In addition to that, we conducted an asymptotic analysis and proved that min-max robustification allows for consistency in model parameter estimation.

The theoretical findings of our study shed a new light on regularized regression in future research. The proposed min-max robustification marks a very attractive addition to big data analysis, where measurement errors tend to be uncontrollable. Indeed, regularized regression is already well-established tools for corresponding applications. However, standard applications are (i) high-dimensional inference, (ii) variable selection, and (iii) dealing with multicollinearity. With the presented equivalence, reg-

ularized regression has novel legitimacy in these contexts. Nevertheless, our results further suggest that it is also beneficial for standard applications. The methodology introduces an alternative concept of robustness that is relatively new to statistics. Accounting for general data uncertainty of a given magnitude (min–max robust) rather than measurement errors of a given distribution (min–min robust) marks a different paradigm that can enhance robust statistics in future applications. By the virtue of these properties, regularized regression can obtain reliable results for instance in survey analysis when sampled individuals provide inaccurate information, or in official statistics when indicators are based on estimates. Another big advantage of our method is its simplicity. In Theorem 1, we establish that regularized regression and min–max robustification are equivalent under the considered setting. This implies that we can obtain min–max robust estimation results by using well-known standard software packages, such as *glmnet*. Accordingly, min–max robustification can be broadly applied and is computationally efficient even for large data sets.

Nevertheless, there is still demand for future research. In Sect. 2.1, it was stated that the nature of the robustification effect is determined by the underlying uncertainty set, which is again determined by the regularization the researcher chooses. It is likely that in practice, there are situations where one regularization works better than the other, depending on the underlying data contamination. As of now, it remains an open question how an optimal regularization could be determined when the measurement errors are completely unknown.

Acknowledgements We thank two anonymous reviewers for their very constructive and supportive comments on our manuscript. Under this guidance, the legibility of the paper could be improved significantly.

Funding Open Access funding enabled and organized by Projekt DEAL. This research is supported by the Spanish Grant PGC2018-096840-B-I00 and by the grant "Algorithmic Optimization (ALOP) - graduate school 2126" funded by the German Research Foundation.

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

- Alfons A, Templ M, Filzmoser P (2013) Robust estimation of economic indicators from survey samples based on Pareto tail modelling. *J R Stat Soc C* 62(2):271–286
- Battese GE, Harter RM, Fuller WA (1988) An error-components model for prediction of county crop areas using survey and satellite data. *J Am Stat Assoc* 83(401):28–36
- Ben-Tal A, El Ghaoui L, Nemirovski A (2009) Robust optimization, vol 28. Princeton University Press, Princeton
- Bertsimas D, Copenhaver MS (2018) Characterization of the equivalence of robustification and regularization in linear and matrix regression. *Eur J Oper Res* 270(3):931–942
- Bertsimas D, Brown D, Caramanis C (2011) Theory and applications of robust optimization. *SIAM Rev* 53(3):464–501

- Bertsimas D, Copenhaver MS, Mazumder R (2017). The trimmed lasso: sparsity and robustness. [arXiv:1708.04527](https://arxiv.org/abs/1708.04527)
- Bottou L, Curtis FE, Nocedal J (2018) Optimization methods for large-scale machine learning. *SIAM Rev* 60(2):223–311
- Burgard J, Krause J, Kreber D (2019) Regularized area-level modelling for robust small area estimation in the presence of unknown covariate measurement errors. *Res Pap Econ* 4:19
- Burgard J, Esteban M, Morales D, Pérez A (2020a) A Fay–Herriot model when auxiliary variables are measured with error. *TEST* 29(1):166–195
- Burgard J, Esteban M, Morales D, Pérez A (2020b) Small area estimation under a measurement error bivariate Fay–Herriot model. *Stat Methods Appl* 5:2. <https://doi.org/10.1007/s10260-020-00515-9>
- Davalos S (2017) Big data has a big role in biostatistics with big challenges and big expectations. *Biostat Biometr Open Access J* 1(3):1–2
- El Ghaoui L, Lebret H (1997) Robust solutions to least-squared problems with uncertain data. *SIAM J Matrix Anal Appl* 18(4):1035–1064
- Fan J, Li R (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *J Am Stat Assoc* 96(456):1348–1360
- Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate gradient descent. *J Stat Softw* 33(1):1–22
- Ghosh A, Thoresen M (2018) Non-concave penalization in linear mixed-effect models and regularized selection of fixed effects. *Adv Stat Anal* 102(2):179–210
- Hoerl A, Kennard R (1970) Ridge regression: biased estimation for nonorthogonal problems. *Techometrics* 12(1):55–67
- Huber PJ (1973) Robust regression: asymptotics, conjectures and Monte Carlo. *Ann Stat* 1:799–821
- Jiang J, Lahiri P, Wan S-M (2002) A unified Jackknife theory for empirical best prediction with M-estimation. *Ann Stat* 30(6):1782–1810
- Li J, Zheng M (2007) Robust estimation of multivariate regression model. *Stat Pap* 50:81–100
- Lindstrom M, Bates D (1988) Newton–Raphson and EM algorithms for linear mixed-effects models for repeated measures data. *J Am Stat Assoc* 84(404):1014–1022
- Loh P-L, Wainwright MJ (2012) High-dimensional regression with noisy and missing covariates: provable guarantees with nonconvexity. *Ann Stat* 40(3):1637–1664
- Markovsky I, Huffel SV (2007) Overview of total least-squares methods. *Signal Process* 87:2283–2302
- Mitra A, Alam K (1980) Measurement error in regression analysis. *Commun Stat Theory Methods* 9(7):717–723
- Neykov NM, Filzmoser P, Neytchev PN (2014) Ultrahigh dimensional variable selection through the penalized maximum trimmed likelihood estimator. *Stat Pap* 55:187–207
- Norouzrad M, Arashi M (2019) Preliminary test and stein-type shrinkage ridge estimators in robust regression. *Stat Pap* 60:1849–1882
- Pinheiro J, Bates D (2000) *Mixed-effects models in S and S-Plus*. Springer series in statistic and computing. Springer, New York
- Rosenbaum M, Tsybakov A (2010) Sparse recovery under matrix uncertainty. *Ann Stat* 38(5):2620–2651
- Rousseeuw PJ, Leroy AM (2003) *Robust regression and outlier detection*. Wiley, Hoboken
- Schelldorfer J, Bühlmann P, van de Geer S (2011) Estimation for high-dimensional linear mixed-effects models using l1-penalization. *Scand J Stat* 38:197–214
- Schmid T, Münnich RT (2014) Spatial robust small area estimation. *Stat Pap* 55:653–670
- Searle S, Casella G, McCulloch C (1992) *Variance components*. Wiley, New York
- Shao J, Deng X (2012) Estimation in high-dimensional linear models with deterministic design matrices. *Ann Stat* 40(2):812–831
- Sørensen O, Frigessi A, Thoresen M (2015) Measurement error in lasso: impact and likelihood bias correction. *Stat Sin* 25:809–829
- Tibshirani R (1996) Regression shrinkage and selection via the lasso. *J R Stat Soc B* 58(1):267–288
- Tseng P, Yun S (2009) A coordinate gradient descent method for nonsmooth separable minimization. *Math Program* 117(1–2):387–402
- White E (2011) Measurement error in biomarkers: sources, assessment, and impact on studies. *IARC Sci Publ* 163:143–161
- Yamada K, Takayasu H, Takayasu M (2018) Estimation of economic indicator announced by government from social big data. *Entropy* 20(11):852–864

- Yang H, Li N, Yang J (2018) A robust and efficient estimation and variable selection method for partially linear models with large-dimensional covariates. *Stat Pap* 57:2: <https://doi.org/10.1007/s00362-018-1013-1>
- Zhang C, Xiang Y (2015) On the oracle property of adaptive group lasso in high-dimensional linear models. *Stat Pap* 57:249–265
- Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. *J R Stat Soc B* 67(2):301–320

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