Universität Trier

Regularized Area-level Modelling for Robust Small Area Estimation in the Presence of Unknown Covariate Measurement Errors

Jan Pablo Burgard
Joscha Krause
Dennis Kreber

Research Papers in Economics
No. 4/19

Regularized area-level modelling for robust small area estimation in the presence of unknown covariate measurement errors

Jan Pablo Burgard, Joscha Krause, Dennis Kreber

Department of Economic and Social Statistics, Trier University
Research Training Group on Algorithmic Optimization, Trier University

Abstract

An approach to model-based small area estimation under covariate measurement errors is presented. Using a min-max approach, we proof that regularized regression coefficient estimation is equivalent to robust optimization under additive noise. Applying this equivalence, the Fay-Herriot model is extended by ℓ_1-norm, squared ℓ_2-norm and elastic net regularizations as robustification against design matrix perturbations. This allows for reliable area-statistic estimates without distributive information about the measurement errors. A best predictor and a Jackknife estimator of the mean squared error are presented. The methodology is evaluated in a simulation study under multiple measurement error scenarios to support the theoretical findings. A comparison to other robust small area approaches is conducted. An empirical application to poverty mapping in the US is provided. Estimated economic figures from the US Census Bureau and crime records from the Uniform Crime Reporting Program are used to model the number of citizens below the federal poverty threshold.

Keywords: min-max, pathwise coordinate descent, regularized least squares, robust optimization

1 Introduction

Small area estimation (SAE) is widely used to obtain reliable estimates of aggregate-specific quantities (area-statistics) from small samples. Model-based SAE methods use regression models to combine data from multiple areas. The objective is to increase estimation efficiency relative to a direct estimator that only uses information from one area at a time. A famous corresponding approach is the Fay-Herriot model introduced by Fay and Herriot (1979). It uses aggregated auxiliary information on the area-level as covariates for model parameter estimation. It is thus commonly referred to as *area-level model* and has been frequently applied in empirical studies over time (see e.g. Slud and Maiti, 2011; Xie et al., 2007; You and Zhou, 2011).

The efficiency gain of the empirical best linear unbiased predictor (EBLUP) under the Fay-Herriot model relative to a direct estimator is determined by the explanatory power of the underlying regression model. It establishes a linear relation between the area-statistic of interest and the aggregated auxiliary information. However, even if the proposed linear relation is generally valid, the corresponding regression model may lack in sufficient explanatory power, or may give false implications regarding the area-statistic. That is, if the covariates are subject to measurement errors. In the Fay-Herriot model, a direct estimator of the area-statistic is regressed on the aggregated auxiliary information. Thus, the approach implicitly allows for some random sampling error on the response variable of the underlying regression model. However, the covariates are assumed to be measured correctly. A violation of this assumption leads to considerably diminished area-statistic estimates, as the Fay-Herriot EBLUP is a convex linear combination of the direct estimates and the predictions from the regression model. Accordingly, if the covariates are perturbed by measurement errors, adjustments are required in order to obtain reliable results.

Several methods have been proposed to treat contaminated observations in model-based SAE. A common approach is using M-estimators (Huber, 1973) for model parameter estimation, as e.g. demonstrated by Sinha and Rao (2009). The basic idea is to reduce the influence of individual observations by applying a sophisticated weighting scheme (e.g. using influence functions). However, this method is primarily suitable for treating distributive outliers of the response variable, but not for dealing with noise in the design matrix. A different approach that explicitly accounts for covariate measurement errors in the Fay-Herriot model was proposed by Ybarra and Lohr (2008). Here, the perturbed covariate values are treated as estimators of the real covariate values. Area-statistic estimates are then derived by accounting for additional uncertainty resulting from the design matrix. A more general approach to treat covariate measurement errors in regression analysis was introduced by Loh and Wainwright (2012). They propose a correction term to a ℓ_1-regularized likelihood function in order to ensure better estimation bounds for the regression coefficient estimates in the presence of measurement errors. However, both Ybarra and Lohr (2008) as well as Loh and Wainwright (2012) require the covariance matrix of the measurement error distribution to be known. This can be an overly restrictive assumption, depending on the empirical application.

We propose a robust extension to the Fay-Herriot model that does not require distribu-

tive information about the measurement errors. For this, we extend theoretical findings provided by Bertsimas and Copenhaver (2018). They used a min-max approach to show that regularized regression coefficient estimation is equivalent to robust optimization under additive noise when the loss function is a seminorm and the regularization is a norm. However, the Fay-Herriot model (and many other regression models) does not fit naturally in this setting, as its loss function is not a seminorm, but a function of a seminorm. In addition to that, many common regularizations, such as the ridge penalty (Hoerl and Kennard, 1970) or the elastic net (Zou and Hastie, 2005), are not norms, but functions of norms. In order to use the characterization of Bertsimas and Copenhaver (2018) for a broader range of models, we generalize their results and proof that a corresponding equivalence holds for strictly monotonously increasing, bijective functions of seminorms and norms as well. In the light of this equivalence, we robustify model parameter estimation in the Fay-Herriot model against unknown design matrix perturbations by applying ℓ_1-norm, squared ℓ_2-norm and elastic net regularizations. Along with robustness, the regularized estimation approach also provides other advantages. Since regularization is often used in the context of high-dimensional inference, it allows for efficient model parameter estimates when the number of observations is small. This is particularly attractive for the SAE setting which is usually characterized by small samples. Further, if the regularization of choice is sparsity inducing, it even allows for automatic variable selection while model parameter estimation.

Regularization parameter tuning is done by way of k-fold cross validation. Regression coefficient estimation is performed via regularized least squared using a modification of the pathwise coordinate descent algorithm proposed by Friedman et al. (2007). The model variance parameter of the Fay-Herriot model is estimated from adjusted maximum likelihood according to Li and Lahiri (2010). A best predictor (BP) under covariate measurement errors is derived. Thereafter, a conservative Jackknife estimator for the mean squared error (MSE) is presented using insights from Jiang et al. (2002). A simulation study is conducted where the regularized predictors are tested against the original Fay-Heriot EBLUP as well as the approaches of Ybarra and Lohr (2008) and Loh and Wainwright (2012) under multiple measurement error scenarios. In addition to that, an empirical application on poverty mapping in the US is provided. We use estimated economic figures provided by the US Census Bureau (US Census Bureau, 2016a,b) and crime records obtained from the Uniform Crime Reporting Program (Uniform Crime Reporting (UCR) Program, 2016) from 2015 to model the number of people with an annual income below 100% of the federal poverty threshhold on the state-level. The estimated values of the auxiliary variables are treated as covariates with measurement error.

The remainder of the paper is organized as follows. In Chapter 2, the area-level model under covariate measurement errors is presented. This includes a description of the Fay-Herriot model and a corresponding extension to measurement errors. Further, it is shown how the model can be robustified by applying regularization. In Chapter 3, model parameter estimation and MSE estimation are presented. Chapter 4 contains the simulation study. In Chapter 5, the empirical application is provided. Chapter 6 closes with an outlook and some conclusive remarks.

2 Regularized area-level model

2.1 Area-level model under covariate measurement errors

The original Fay-Herriot model is described first, then an extension to covariate measurement errors is presented. Let $U = \bigcup_{i=1}^{m} U_i$ be a finite population of size N that is segmented into m pairwise disjoint areas U_i of size N_i with $\sum_{i=1}^{m} N_i = N$. For simplicity, we only refer to areas with their corresponding index $i = 1, ..., m$. Assume a random sample $S \subset U$ of size n to be drawn such that there are m area-specific subsamples $S_i \subset U_i$ of size $n_i > 0$ with $S = \bigcup_{i=1}^{m} S_i$ and $\sum_{i=1}^{m} n_i = n$. Note that in the SAE context, n_i is usually small. Let $\theta_i \in \mathbb{R}$ denote an unknown statistic of interest within area i, for example the area-specific mean of some random variable. Let $\widehat{\theta}_i^{dir} \in \mathbb{R}$ be a direct estimator of θ_i that is available for all $i = 1, ..., m$. It is assumed to be design-unbiased, hence $E(\widehat{\theta}_i^{dir}) = \theta_i$, and obtained from only using the sample information within S_i. Due to the small area-specific sample size n_i, its variance $Var(\widehat{\theta}_i^{dir})$ is too large in order to draw reliable conclusions on θ_i. Thus, the objective is to find a better estimator of θ_i, denoted by $\widehat{\theta}_i \in \mathbb{R}$, for all $i = 1, ..., m$. Fay and Herriot (1979) proposed a very influential model, the so-called Fay-Herriot model, to obtain an improved estimator for θ_i by using suitable auxiliary information. The model consists of two components. The first component (sampling model) states that due to the design-unbiasedness of $\widehat{\theta}_i^{dir}$, the direct estimator is equal to the unknown θ_i plus some random sampling error

$$\widehat{\theta}_i^{dir} = \theta_i + e_i, \quad e_i \overset{ind}{\sim} N(0, D_i), \quad \forall\ i = 1, ..., m, \tag{1}$$

with e_i as sampling error and $D_i = Var(\widehat{\theta}_i^{dir} | \theta_i)$ as sampling variance in area i. Within this paper, we assume D_i to be known for all areas. In practise, it is usually obtained from some generalized variance function. The second component (linking model) treats θ_i as random and establishes a linear relation to some area-level auxiliary information

$$\theta_i = \mathbf{x}_i' \boldsymbol{\beta} + v_i, \quad v_i \overset{iid}{\sim} N(0, A), \quad \forall\ i = 1, ..., m, \tag{2}$$

with $\mathbf{x}_i \in \mathbb{R}^p$ as vector of auxiliary information and $\boldsymbol{\beta} \in \mathbb{R}^p$ as vector of regression coefficients. v_i denotes an area-specific random effect with unknown model variance $A \geq 0$ and $e_1, ..., e_m, v_1, ..., v_m$ stochastically independent. The marginal distribution of the direct estimator under the model thus is

$$\widehat{\theta}_i^{dir} \overset{ind}{\sim} N(\mathbf{x}_i' \boldsymbol{\beta}, D_i + A), \quad \forall\ i = 1, ..., m, \tag{3}$$

or, in matrix notation, $\widehat{\boldsymbol{\theta}}^{dir} \overset{ind}{\sim} MVN(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}(A))$, with $\widehat{\boldsymbol{\theta}}^{dir} = (\widehat{\theta}_1^{dir}, ..., \widehat{\theta}_m^{dir})'$ as response vector, $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_m)'$ as design matrix, and $\boldsymbol{\Sigma}(A) = diag(A + D_1, ..., A + D_m)$ as covariance matrix of the variance components. The basic idea of the Fay-Herriot model is to improve the direct estimator $\widehat{\theta}_i^{dir}$ by exploiting the functional relation between θ_i and $\mathbf{x}_i$. In order to determine the functional relation, the unknown model parameters A and $\boldsymbol{\beta}$ have to be estimated. This is usually performed iteratively by finding some initial model variance

estimate $\widehat{A}$ first, and then obtain the regression coefficient estimates $\widehat{\boldsymbol{\beta}}$ conditionally on $\widehat{A}$ according to

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \ \left|\left|\boldsymbol{\Sigma}(\widehat{A})^{-1/2}\left(\widehat{\boldsymbol{\theta}}^{dir} - \mathbf{X}\boldsymbol{\beta}\right)\right|\right|_2^2, \quad \boldsymbol{\Sigma}(\widehat{A}) = diag(D_1 + \widehat{A}, ..., D_m + \widehat{A}). \tag{4}$$

Afterwards, the model variance estimate $\widehat{A}$ is updated conditionally on $\widehat{\boldsymbol{\beta}}$ using maximum likelihood (ML) or restricted maximum likelihood (REML) approaches. See Li and Lahiri (2010) or Yoshimori and Lahiri (2014) for further details. The conditional estimation steps are repeated until convergence. Note that estimation is performed by using the sample and auxiliary information from all m areas simultaneously. Accordingly, the estimation of θ_i is improved through considering more information relative to the direct estimator $\widehat{\theta}_i^{dir}$, that only considers information from area i. This often referred to as *borrowing strength*. The final estimator $\widehat{\theta}_i^{FH}$ under the Fay-Herriot model is then obtained from a convex linear combination of $\widehat{\theta}_i^{dir}$ and the regression-synthethic component $\mathbf{x}_i'\widehat{\boldsymbol{\beta}}$ of the model (Molina et al., 2015):

$$\widehat{\theta}_i^{FH} = \widehat{\gamma}_i\widehat{\theta}_i^{dir} + (1 - \widehat{\gamma}_i)\mathbf{x}_i'\widehat{\boldsymbol{\beta}}, \quad \forall\ i = 1, ..., m, \tag{5}$$

with $\widehat{\gamma}_i = \widehat{A}/(\widehat{A} + D_i)$ as area-specific shrinkage factor that is determined by the relation between the variance parameters of the two model components. Here, the term shrinkage relates to $\widehat{\theta}_i^{FH}$ being shrunken towards $\widehat{\theta}_i^{dir}$ as a result of variance weighting. It is not to be confused with the shrinkage of some $\widehat{\beta}_j \in \widehat{\boldsymbol{\beta}}$ towards zero due to regularization, which is described in the next subsection. If $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\boldsymbol{\Sigma}(\widehat{A})^{-1}\boldsymbol{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}(\widehat{A})^{-1}\widehat{\boldsymbol{\theta}}^{dir}$, and $\widehat{A}$ is a consistent estimator of A, then $\widehat{\theta}_i^{FH}$ is the EBLUP for θ_i. For further details on the Fay-Herriot model, we refer to Rao and Molina (2015).

We now present the area-level model under covariate measurement errors and derive a corresponding BP. Consider the sampling model (1) and the linking model (2) from the original Fay-Herriot model. In order to account for the measurement errors, an additional error model is required. It can be stated as

$$\tilde{\mathbf{x}}_i = \mathbf{x}_i + \boldsymbol{\Delta}_i, \quad \forall\ i = 1, ..., m, \tag{6}$$

where $\tilde{\mathbf{x}}_i \in \mathbb{R}^p$ denotes the impaired covariate vector resulting from an unknown area-specific error vector $\boldsymbol{\Delta}_i \in \mathbb{R}^p$ that is added to $\mathbf{x}_i$. From (1), (2) and (6), the area-level model under measurement errors is formulated according to

$$\widehat{\theta}_i^{dir} = (\mathbf{x}_i + \boldsymbol{\Delta}_i)'\boldsymbol{\beta} + v_i + e_i, \quad \forall\ i = 1, ..., m. \tag{7}$$

Note that no distributive assumption regarding the measurement errors is made. We treat $\boldsymbol{\Delta}_i$ as a fixed unobservable perturbation of $\mathbf{x}_i$. Under the model (7), the conditional distributions of the direct estimator $\widehat{\theta}_i^{dir}$ are given by

$$\begin{aligned} \widehat{\theta}_i^{dir}|\mathbf{x}_i, v_i &\sim N((\mathbf{x}_i + \boldsymbol{\Delta}_i)'\boldsymbol{\beta} + v_i, D_i) \\ \widehat{\theta}_i^{dir}|\mathbf{x}_i &\sim N((\mathbf{x}_i + \boldsymbol{\Delta}_i)'\boldsymbol{\beta} + v_i, D_i + A). \end{aligned} \tag{8}$$

Assuming the model variance A and the sampling variances D_i to be known, for the conditional distribution of the random effect

$$f(v_i|\widehat{\theta}_i^{dir}, \mathbf{x}_i) \propto f(v_i) f(\widehat{\theta}_i^{dir}|\mathbf{x}_i, v_i) \tag{9}$$

holds, where

$$\begin{aligned}
f(v_i) f(\widehat{\theta}_i^{dir}|\mathbf{x}_i, v_i) &= \frac{1}{\sqrt{2\pi A}} exp\left(-\frac{v_i^2}{2A}\right) \frac{1}{\sqrt{2\pi D_i}} exp\left(-\frac{(\widehat{\theta}_i^{dir} - (\mathbf{x}_i + \boldsymbol{\Delta}_i)'\boldsymbol{\beta} - v_i)^2}{2D_i}\right) \\
&\propto exp\left(-\frac{v_i^2}{2A}\right) exp\left(-\frac{v_i^2 - 2v_i\left(\widehat{\theta}_i^{dir} - (\mathbf{x}_i + \boldsymbol{\Delta}_i)'\boldsymbol{\beta}\right)}{2D_i}\right) \\
&= exp\left(-\frac{v_i^2}{2}\left(\frac{1}{D_i} + \frac{1}{A}\right) + \frac{\widehat{\theta}_i^{dir} - (\mathbf{x}_i + \boldsymbol{\Delta}_i)'\boldsymbol{\beta}}{D_i} v_i\right) \\
&= exp\left(-\frac{v_i^2}{2}\frac{1}{\frac{D_i \cdot A}{D_i + A}} + \frac{1}{\frac{D_i \cdot A}{D_i + A}} \frac{A\left(\widehat{\theta}_i^{dir} - (\mathbf{x}_i + \boldsymbol{\Delta}_i)'\boldsymbol{\beta}\right)}{D_i + A} v_i\right).
\end{aligned}$$

Accordingly, the conditional distribution is a univariate normal

$$v_i|\widehat{\theta}_i^{dir}, \mathbf{x}_i \sim N\left(\frac{A\left(\widehat{\theta}_i^{dir} - (\mathbf{x}_i + \boldsymbol{\Delta}_i)'\boldsymbol{\beta}\right)}{D_i + A}, \frac{D_i \cdot A}{D_i + A}\right). \tag{10}$$

Finally, the BP under the model is the conditional expectation $E(\theta_i|\mathbf{x}_i, \widehat{\theta}_i^{dir})$, which can be expressed as

$$\begin{aligned}
\widehat{\theta}_i^{BP} &= \mathbf{x}_i'\boldsymbol{\beta} + E\left(\boldsymbol{\Delta}_i|\mathbf{x}_i, \widehat{\theta}_i^{dir}\right) + E\left(v_i|\mathbf{x}_i, \widehat{\theta}_i^{dir}\right) \\
&= \mathbf{x}_i'\boldsymbol{\beta} + \boldsymbol{\Delta}_i'\boldsymbol{\beta} + \frac{A}{A + D_i}\left(\widehat{\theta}_i^{dir} - \mathbf{x}_i'\boldsymbol{\beta} - \boldsymbol{\Delta}_i'\boldsymbol{\beta}\right) \\
&= \frac{A}{A + D_i}\widehat{\theta}_i^{dir} + \frac{D_i}{A + D_i}\left(\mathbf{x}_i'\boldsymbol{\beta} + \boldsymbol{\Delta}_i'\boldsymbol{\beta}\right) \\
&= \gamma_i \widehat{\theta}_i^{dir} + (1 - \gamma_i)\left(\mathbf{x}_i'\boldsymbol{\beta} + \boldsymbol{\Delta}_i'\boldsymbol{\beta}\right) \\
&= \gamma_i \widehat{\theta}_i^{dir} + (1 - \gamma_i)\tilde{\mathbf{x}}_i'\boldsymbol{\beta}.
\end{aligned} \tag{11}$$

2.2 Robustification against covariate measurement errors

Hereafter, we show analytically how model parameter estimation in the presented area-level model under covariate measurement errors is related to regularized model parameter estimation in the original Fay-Herriot model. Recall that in the latter model the auxiliary information is assumed to be measured without error. We robustify the regression coefficient estimates $\widehat{\boldsymbol{\beta}}$ in the Fay-Herriot model against design matrix perturbations. The term

"robustness" is not always connoted consistently and therefore many approaches exist that account for the effects of measurement interference. Bertsimas et al. (2017) single out two general approaches to robustification in regression, an optimistic and a pessimistic perspective, which they call the *min-min* and *min-max* approach. For a function $g : \mathbb{R}^n \to \mathbb{R}$, a set $\mathcal{U} \subseteq \mathbb{R}^{n\times p}$, a design matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$ and a response vector $\boldsymbol{y} \in \mathbb{R}^n$, the *min-min* approach is formulated by the optimization problem

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

while the *min-max* approach is characterized by the problem

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

In both variants the design matrix is perturbed to account for some measurement errors. The *min-min* approach is mainly used in *robust statistics*, where the concern is to robustify against distributive 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) and *total least squares* (Markovsky and Huffel, 2007). In contrast, the *min-max* method mainly stems from *robust optimization*, which aims at finding solutions that are still "good" or feasible under some uncertainty. Here, deterministic assumptions about the set $\mathcal{U}$ are made. The set $\mathcal{U}$ is then called the *uncertainty set* and is chosen in accordance to how the user believes the additive error might be structured. This robustification viewpoint is for example given by Bertsimas and Copenhaver (2018), Ben-Tal et al. (2009), and El Ghaoui and Lebret (1997). In light of the Fay-Harriot model, we assume to have no distributive information about the errors, and at most we might only be able to guess the severance of the noise. Due to this lack of information we regard the disturbance of $\mathbf{X}$ pessimistically, i.e., we use the *min-max* approach to introduce robustness to our estimate. That is, we are looking at the optimization model

$$\min_{\boldsymbol{\beta}} \max_{\boldsymbol{\Delta}\in\mathcal{U}} \left\| \boldsymbol{\Sigma}(\widehat{A})^{-1/2} \left(\widehat{\boldsymbol{\theta}}^{dir} - (\mathbf{X} + \boldsymbol{\Delta})\,\boldsymbol{\beta} \right) \right\|_2^2.$$

Since it is not obviously clear how to efficiently solve such a min-max problem, we next present how this problem is connected to regularized regression problems in the form

$$\min_{\boldsymbol{\beta}} g(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda h(\boldsymbol{\beta}), \tag{12}$$

where $\lambda > 0$ is a regularization parameter. From an optimization stand point, problems of the class (12) can be handled much better and therefore solved more efficiently. However, it is uncommon to regard (12) as a robustification. Typically, regression models are extended by some form of regularization to induce some shrinkage on the coefficients in order to conduct a model selection or to deal with multicollinearity. However, rarely are they considered means of robustification. We take a look at the regularization from a different

point of view, albeit an unconventional one, that is, we regard regularization as a form of robustification. In this sense, the following result by Bertsimas and Copenhaver (2018) is particularly helpful, in that it connects the min-max method with a regularization problem.

Proposition 1 (Bertsimas and Copenhaver (2018))**.** *If $g : \mathbb{R}^n \to \mathbb{R}$ is a seminorm which is not identically zero and $h : \mathbb{R}^p \to \mathbb{R}$ is a norm, then for any $\boldsymbol{z} \in \mathbb{R}^n$ and $\boldsymbol{\beta} \in \mathbb{R}^p$*

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

where

$$\mathcal{U} = \left\{ \boldsymbol{\Delta} : \max_{\boldsymbol{\gamma} \in \mathbb{R}^p} \frac{g(\boldsymbol{\Delta}\boldsymbol{\gamma})}{h(\boldsymbol{\gamma})} \leq \lambda \right\}.$$

Clearly, the Proposition directly implies that

$$\min_{\boldsymbol{\beta}} \max_{\boldsymbol{\Delta} \in \mathcal{U}} g((\mathbf{X} + \boldsymbol{\Delta})\boldsymbol{\beta} - \boldsymbol{y}) = g(\mathbf{X}\boldsymbol{\beta} - \boldsymbol{y}) + \lambda h(\boldsymbol{\beta})$$

for g, h and $\mathcal{U}$ as in Proposition 1. 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 constraints the uncertainty set $\mathcal{U}$, which defines a set of perturbations for the design matrix. In other words, the regularization controls the magnitude of noise, which can be added to $\mathbf{X}$. 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, which is added to $\mathbf{X}$. However, many regression methods are formulated using the squared norm or a mix of squared and non-squared norms. For instance, ridge regression (Hoerl and Kennard, 1970) is posed as the optimization problem

$$\min_{\boldsymbol{\beta}} \|\mathbf{X}\boldsymbol{\beta} - \boldsymbol{y}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_2^2$$

with both, the deviation and regularization, being squared. On the other hand, we have LASSO (Tibshirani, 1996), which is defined by

$$\min_{\boldsymbol{\beta}} \|\mathbf{X}\boldsymbol{\beta} - \boldsymbol{y}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1.$$

Here, the deviation is squared while the regularization term is not. 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. However, it is claimed that ridge regression and LASSO correspond to the specific cases $g = h = \ell_2$ and $g = \ell_2, h = \ell_1$. In the following we propose a generalization of the described issue and transfer it to the robustness framework presented in Proposition 1.

Lemma 1. *Let $g : \mathbb{R}^n \to \mathbb{R}$ and $h_1, h_2, \ldots, h_d : \mathbb{R}^n \to \mathbb{R}$ be convex functions. If*

$$\hat{\boldsymbol{z}} \in \operatorname*{argmin}_{\boldsymbol{z} \in \mathbb{R}^n} g(\boldsymbol{z}) + \sum_{i=1}^{d} \lambda_i h_i(\boldsymbol{z}) \tag{13}$$

for the parameters $\lambda_1, \ldots, \lambda_d > 0$, then there exist $c_1, \ldots, c_d > 0$ such that

$$\begin{aligned} \hat{\boldsymbol{z}} \in \underset{\boldsymbol{z} \in \mathbb{R}^n}{\operatorname{argmin}} \quad & g(\boldsymbol{z}) \\ \text{s.t.} \quad & h_i(\boldsymbol{z}) \leq c_i \qquad \text{for all } 1 \leq i \leq d \end{aligned} \tag{14}$$

and vice versa, if there is a $\boldsymbol{z}$ such that $h_i(\boldsymbol{z}) < c_i$ for all $1 \leq i \leq d$ then for given $c_1, \ldots, c_d > 0$ there exist $\lambda_1, \ldots, \lambda_d > 0$ such that $\hat{\boldsymbol{z}}$ is an optimal solution for both problems.

Proof. Assume (13) holds. We then define $c_i = h_i(\hat{\boldsymbol{z}})$ for all $1 \leq i \leq d$. Now assume that $\hat{\boldsymbol{z}}$ is not an optimal solution of (14) and instead $\boldsymbol{z}^*$ provides a better objective value, i.e., $g(\hat{\boldsymbol{z}}) > g(\boldsymbol{z}^*)$, while satisfying $h_i(\boldsymbol{z}^*) \leq c_i$ for all $1 \leq i \leq d$. This would imply that

$$g(\boldsymbol{z}^*) + \sum_{i=1}^{d} \lambda_i h_i(\boldsymbol{z^*}) < g(\hat{\boldsymbol{z}}) + \sum_{i=1}^{d} \lambda_i h_i(\boldsymbol{z^*}) \leq g(\hat{\boldsymbol{z}}) + \sum_{i=1}^{d} \lambda_i h_i(\hat{\boldsymbol{z}})$$

in contradiction to $\boldsymbol{z}$ being an optimal solution of (13). Therefore, (14) must hold.

Now assume that $\hat{\boldsymbol{z}}$ is an optimal solution of the constrained optimization problem, i.e., (14) holds. We use Lagrange duality to prove that (13) holds as well. Note that, the Slater conditions are satisfied due to $g, h_1, \ldots, h_d$ being convex and because there is a $\boldsymbol{z}$ such that $h_i(\boldsymbol{z}) < c_i$ for all $1 \leq i \leq d$. Thus, strong duality holds. It follows that

$$\max_{\boldsymbol{\lambda} \geq \boldsymbol{0}} \min_{\boldsymbol{z} \in \mathbb{R}^n} g(\boldsymbol{z}) + \sum_{i=1}^{d} \lambda_i (h_i(\boldsymbol{z}) - c_i) \tag{15}$$

is equivalent to (14), that is, the objective values of (14) and (15) are identical. Let $\hat{\boldsymbol{\lambda}}$ be an optimal solution of (15), then due to complementary slackness (see for example Boyd and Vandenberghe, 2009, p. 242)

$$g(\hat{\boldsymbol{z}}) = g(\hat{\boldsymbol{z}}) + \sum_{i=1}^{d} \hat{\lambda}_i (h_i(\hat{\boldsymbol{z}}) - c_i) = \min_{\boldsymbol{z} \in \mathbb{R}^n} g(\boldsymbol{z}) + \sum_{i=1}^{d} \hat{\lambda}_i (h_i(\boldsymbol{z}) - c_i)$$

holds, which proves the conjecture. □

Proposition 2. *Let $g : \mathbb{R}^n \to \mathbb{R}$ be a seminorm which is not identically zero, let $h_1, h_2, \ldots h_d : \mathbb{R}^p \to \mathbb{R}$ be norms and $f, f_1, f_2, \ldots, f_d : \mathbb{R}_+ \to \mathbb{R}_+$ be strictly monotonously increasing, bijective functions, then there exist $\mu_1, \ldots, \mu_d > 0$ such that*

$$\underset{\boldsymbol{\beta}}{\operatorname{argmin}} \max_{\boldsymbol{\Delta} \in \mathcal{U}} g(\boldsymbol{y} - (\mathbf{X} + \boldsymbol{\Delta})\boldsymbol{\beta}) = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\ f(g(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})) + \sum_{i=1}^{d} \lambda_i f_i(h_i(\boldsymbol{\beta}))$$

where

$$\mathcal{U} = \left\{ \boldsymbol{\Delta} : g(\boldsymbol{\Delta}\boldsymbol{\gamma}) \leq \sum_{i=1}^{d} \mu_i h_i(\boldsymbol{\gamma}) \text{ for all } \boldsymbol{\gamma} \in \mathbb{R}^p \right\}$$

Proof. We first look at the right-hand side minimization problem

$$M \coloneqq \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\ f(g(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})) + \sum_{i=1}^{d} \lambda_i f_i(h_i(\boldsymbol{\beta})).$$

Lemma 1 yields that there exist $c_1, \ldots, c_d > 0$ such that

$$\begin{aligned} M = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \quad & f(g(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})) \\ \text{s.t.} \quad & f_i(h_i(\boldsymbol{\beta})) \leq c_i \quad \text{for all } 1 \leq i \leq d. \end{aligned}$$

Since $f_1, \ldots, f_d$ are bijective and monotonously increasing, it follows that

$$\begin{aligned} M = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \quad & g(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}) \\ \text{s.t.} \quad & h_i(\boldsymbol{\beta}) \leq f_i^{-1}(c_i) \quad \text{for all } 1 \leq i \leq d. \end{aligned}$$

We once again apply Lemma 1 and get that there are $\mu_1, \ldots, \mu_d > 0$ such that

$$M = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\ g(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}) + \sum_{i=1}^{d} \mu_i h_i(\boldsymbol{\beta})$$

It is easy to see that $\sum_{i=1}^{d} \mu_i h_i$ is a norm and thus by Proposition 1 we get that

$$M = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \max_{\boldsymbol{\Delta} \in \mathcal{U}} g(\boldsymbol{y} + (\mathbf{X} + \boldsymbol{\Delta})\boldsymbol{\beta})$$

with $\mathcal{U} = \left\{ \boldsymbol{\Delta} : g(\boldsymbol{\Delta}\boldsymbol{\gamma}) \leq \sum_{i=1}^{d} \mu_i h_i(\boldsymbol{\gamma}) \text{ for all } \boldsymbol{\gamma} \in \mathbb{R}^p \right\}$. □

The Proposition enables us to regard more sophisticated regularizations in light of robustification. Unfortunately, the direct one-to-one relation of the regularization parameter and the uncertainty set is lost, when generalizing Proposition 1. However, in practice the optimal regularization parameter is usually not known a priori and is obtained by conducting a cross validation. Hence, we argue that a direct one-to-one connection is not necessarily required, even though it would certainly paint a clearer picture.

2.3 Regularizations

After pointing out the relation between regularization and robustification in the Fay-Herriot model, we briefly describe the regularizations considered for this study.

ℓ_1-regularization

The first regularization is the ℓ_1-norm, which is famously used for the LASSO introduced by Tibshirani (1996) for linear regression. Within the Fay-Herriot model, including an

ℓ_1-norm regularization changes the weighted minimization problem (4) for obtaining $\widehat{\boldsymbol{\beta}}$ to

$$\widehat{\boldsymbol{\beta}}_{\ell_1} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left|\left|\boldsymbol{\Sigma}(\widehat{A})^{-1/2}\left(\widehat{\boldsymbol{\theta}}^{dir} - \mathbf{X}\boldsymbol{\beta}\right)\right|\right|_2^2 + \lambda||\boldsymbol{\beta}||_1. \tag{16}$$

Applying Proposition 2 with $g(\boldsymbol{z}) \coloneqq \|\boldsymbol{z}\|_2$, $h(\boldsymbol{z}) = \|\boldsymbol{z}\|_1$, $f(z) = z^2$ and $f_1(z) = z$ yields the equivalence

$$\widehat{\boldsymbol{\beta}}_{\ell_1} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \max_{\boldsymbol{\Delta}\in\mathcal{U}_{\ell_1}} \left\|\boldsymbol{\Sigma}(\widehat{A})^{-1/2}\left(\widehat{\boldsymbol{\theta}}^{dir} - \mathbf{X}\boldsymbol{\beta}\right) + \boldsymbol{\Delta}\boldsymbol{\beta}\right\|_2 \tag{17}$$

for some $\mu > 0$ and $\mathcal{U}_{\ell_1} = \{\boldsymbol{\Delta} : \|\boldsymbol{\Delta}\boldsymbol{\gamma}\|_2 \leq \mu\|\boldsymbol{\gamma}\|_1 \text{ for all } \boldsymbol{\gamma} \in \mathbb{R}^p\}$. From the formulation (17), it is evident that the coefficients are robustified against noise by introduction of the perturbation $\boldsymbol{\Delta}$, which is maximized in regard to the fitted coefficients and is constrained by the uncertainty set $\mathcal{U}_{\ell_1}$. The following result by Bertsimas and Copenhaver (2018) improves the interpretability of the uncertainty set $\mathcal{U}_{\ell_1}$.

Proposition 3 (Bertsimas and Copenhaver (2018))**.** *Let be $p \in [1, \infty]$, let $\|\boldsymbol{\beta}\|_0$ be the number of non-zero entries of $\boldsymbol{\beta}$ and let $\boldsymbol{\Delta}_i$ be the i-th column of $\boldsymbol{\Delta}$. If*

$$\mathcal{U}' = \{\boldsymbol{\Delta} : \|\boldsymbol{\Delta}\boldsymbol{\beta}\|_2 \leq \mu\|\boldsymbol{\beta}\|_0 \quad \forall\|\boldsymbol{\beta}\|_p \leq 1\}$$

and

$$\mathcal{U}'' = \{\boldsymbol{\Delta} : \|\boldsymbol{\Delta}_i\|_2 \leq \mu \quad \forall i\}$$

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

Thus, in case of the ℓ_1-regularization the noise is constrained column-wise by the parameter μ. However, to our knowledge it remains an open question why a specific uncertainty set, or accordingly a specific regularization, is an appropriate choice. Therefore, it makes sense to take other well-known properties of a regularization into consideration. On that note, including the ℓ_1-norm in the minimization problem induces a sparse solution for $\widehat{\boldsymbol{\beta}}$. As a result, some elements $\widehat{\beta}_j \in \widehat{\boldsymbol{\beta}}$ that are irrelevant for the functional description of $\widehat{\boldsymbol{\theta}}^{dir}$ are set exactly to zero in the estimation process. This implies an automatic variable selection, which makes the ℓ_1-norm regularization applicable to a broad range of high-dimensional regression problems. However, note that it is known to produce instable results in the presence of strongly correlated covariates (Zou and Hastie, 2005; Friedman et al., 2010).

$\boldsymbol{\ell_2}$-regularization

The second regularization is the squared ℓ_2-norm, which is famously used for ridge regression proposed by Hoerl and Kennard (1970). The corresponding weighted minimization problem to determine $\widehat{\boldsymbol{\beta}}$ can be stated as

$$\widehat{\boldsymbol{\beta}}_{\ell_2} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left|\left|\boldsymbol{\Sigma}(\widehat{A})^{-1/2}\left(\widehat{\boldsymbol{\theta}}^{dir} - \mathbf{X}\boldsymbol{\beta}\right)\right|\right|_2^2 + \lambda||\boldsymbol{\beta}||_2^2. \tag{18}$$

Note that using the squared ℓ_2-norm ensures separability in the coordinate descent algorithm used for model parameter estimation in Chapter 3.1. Thus, once again we cannot

use Proposition 1 directly and have to apply Proposition 2. When setting $g(\boldsymbol{z}) = \|\boldsymbol{z}\|_2$, $h(\boldsymbol{z}) = \|\boldsymbol{z}\|_2$, $f(z) = z^2$ and $f_1(z) = z^2$ we obtain

$$\widehat{\boldsymbol{\beta}}_{\ell_2} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \max_{\boldsymbol{\Delta} \in \mathcal{U}_{\ell_2}} \left\| \boldsymbol{\Sigma}(\widehat{A})^{-1/2} \left(\widehat{\boldsymbol{\theta}}^{dir} - \mathbf{X}\boldsymbol{\beta} \right) + \boldsymbol{\Delta}\boldsymbol{\beta} \right\|_2$$

for some $\mu > 0$ and $\mathcal{U}_{\ell_2} = \{\boldsymbol{\Delta} : \|\boldsymbol{\Delta}\boldsymbol{\gamma}\|_2 \leq \mu \|\boldsymbol{\gamma}\|_2 \text{ for all } \boldsymbol{\gamma} \in \mathbb{R}^p\}$. Clearly, $\mathcal{U}_{\ell_2}$ is equal to the set $\{\boldsymbol{\Delta} : \sigma_{\max}(\boldsymbol{\Delta}) \leq \mu\}$ where $\sigma_{\max}(\boldsymbol{\Delta})$ is the maximum singular value of the matrix $\boldsymbol{\Delta}$. Whereas, the ℓ_1 regularization induced bounds on the individual columns of the perturbation, the ℓ_2 regularization enforces a coherent bound of the whole noise matrix. In addition to the robustness effect, including the ℓ_2-norm in the minimization problem induces a dense and smooth solution for $\widehat{\boldsymbol{\beta}}$. All elements $\widehat{\beta}_j \in \widehat{\boldsymbol{\beta}}$ remain non-zero in the estimation process, whereas their individual contributions to the description of θ_i are equalized to some extent, depending on the value of λ. If $\lambda \to \infty$, then $\widehat{\beta}_1 = ... = \widehat{\beta}_p$. The ℓ_2-norm regularization has shown to produce stable results in the presence of correlated covariates. However, as $\widehat{\beta}_j \neq 0 \; \forall j = 1, ..., p$, no automatic variable selection is conducted.

Elastic net
The third regularization is a linear combination of ℓ_1- and ℓ_2-norm, which is used for the elastic net introduced by Zou and Hastie (2005). The corresponding weighted minimzation problem is given by

$$\widehat{\boldsymbol{\beta}}_{en} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\| \left\| \boldsymbol{\Sigma}(\widehat{A})^{-1/2} \left(\widehat{\boldsymbol{\theta}}^{dir} - \mathbf{X}\boldsymbol{\beta} \right) \right\|_2^2 + \lambda \left[\alpha ||\boldsymbol{\beta}||_1 + (1-\alpha) ||\boldsymbol{\beta}||_2^2 \right], \tag{19}$$

where $\alpha \in [0, 1]$ is a hyperparameter controlling the influence of the ℓ_1- and ℓ_2-norm in the regularization. Note that for $\alpha = 1$, the elastic net reduces to the LASSO, and for $\alpha = 0$, it is equivalent to the ridge penalty. By Proposition 2 where $g(\boldsymbol{z}) = \|\boldsymbol{z}\|_2$, $h_1(\boldsymbol{z}) = \|\boldsymbol{z}\|_1$, $h_2(\boldsymbol{z}) = \|\boldsymbol{z}\|_2$, $f(z) = z^2$, $f_1(z) = z$ and $f_2(z) = z^2$ we obtain

$$\widehat{\boldsymbol{\beta}}_{en} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \max_{\boldsymbol{\Delta} \in \mathcal{U}_{en}} \left\| \boldsymbol{\Sigma}(\widehat{A})^{-1/2} \left(\widehat{\boldsymbol{\theta}}^{dir} - \mathbf{X}\boldsymbol{\beta} \right) + \boldsymbol{\Delta}\boldsymbol{\beta} \right\|_2$$

for some $\mu_1, \mu_2 > 0$ and $\mathcal{U}_{en} = \{\boldsymbol{\Delta} : \|\boldsymbol{\Delta}\boldsymbol{\gamma}\|_2 \leq \mu_1 \|\boldsymbol{\gamma}\|_1 + \mu_2 \|\boldsymbol{\gamma}\|_2 \text{ for all } \boldsymbol{\gamma} \in \mathbb{R}^p\}$. Unfortunately, it is less apparent how to interpret $\mathcal{U}_{en}$.

3 Estimation

3.1 Model parameter estimation

Hereafter, we apply the theoretical findings from Chapter 2 and present some details on regularized model parameter estimation. For this, a value for the regularization parameter λ must be chosen. Remember that λ has implications regarding the level of noise that can be added to the design matrix $\mathbf{X}$. Following Friedman et al. (2010), we use k-fold cross validation under minimization of the squared prediction error in order to determine λ. Then, assuming the sampling variances D_i to be known for $i = 1, ..., m$, and letting

$r = 1, ...$ denote the index of iterations until convergence, model parameter estimation is performed according to the following procedure:

Algorithm 1 Model Parameter Estimation

1: find an initial estimate $\widehat{A}^{init}$ and set $\widehat{A}^{init} := \widehat{A}^{r-1}$
2: **while** not converged **do**
3: estimate $\widehat{\boldsymbol{\beta}}^r = \widehat{\boldsymbol{\beta}}^r(\widehat{A}^{r-1})$ conditionally on $\widehat{A}^{r-1}$ under a given regularization
4: update the model variance estimate $\widehat{A}^r = \widehat{A}^r(\widehat{\boldsymbol{\beta}}^r)$ conditionally on $\widehat{\boldsymbol{\beta}}^r$
5: set $\widehat{A}^r := \widehat{A}^{r-1}$
6: check convergence
7: **end while**
8: **return** $\widehat{\boldsymbol{\beta}} := \widehat{\boldsymbol{\beta}}^r$ and $\widehat{A} = \widehat{A}^r$

An initial estimate $\widehat{A}^{init}$ of A is chosen. Then the regression coefficients $\boldsymbol{\beta}^r$ are estimated given $\widehat{A}^{init}$. Afterwards, the initial estimate is updated by a new estimate $\widehat{A}^r$ conditionally on the obtained estimates for $\boldsymbol{\beta}^r$. The procedure is repeated until convergence. In order to include regularization in the estimation process, let

$$\mathcal{Q}(\alpha, \boldsymbol{\beta}, \lambda) = \left|\left|\boldsymbol{\Sigma}(\widehat{A})^{-1/2}\left(\widehat{\boldsymbol{\theta}}^{dir} - \mathbf{X}\boldsymbol{\beta}\right)\right|\right|_2^2 + \lambda\left[\alpha||\boldsymbol{\beta}||_1 + (1-\alpha)||\boldsymbol{\beta}||_2^2\right] \tag{20}$$

denote the objective function to be minimized for regression coefficient estimation under elastic net regularization. As $\mathcal{Q}(\cdot)$ contains the ℓ_1-norm and ℓ_2-norm as special cases, all regularizations discussed in Chapter 2.3 can be described accordingly. To minimize the function, the pathwise coordinate descent algorithm described by Friedman et al. (2007) as well as Friedman et al. (2010) is applied. Coordinate descent implies that the loss-function is partially minimized with respect to a single $\tilde{\beta}_j \in \tilde{\boldsymbol{\beta}}$ in each coordinate descent step while the remaining $\tilde{\beta}_k$ with $k \neq j$ are kept fixed. This requires separability of the regression coefficients in the objective function. If $\tilde{\beta}_j \neq 0$, the gradient at $\beta_j = \tilde{\beta}_j$ can be computed according to

$$\left.\frac{\partial Q(\cdot)}{\partial \beta_j}\right|_{\boldsymbol{\beta}=\tilde{\boldsymbol{\beta}}} = -2\left(\boldsymbol{\Sigma}(\widehat{A})^{-1/2}\mathbf{x}^j\right)'\left[\boldsymbol{\Sigma}(\widehat{A})^{-1/2}\left(\widehat{\boldsymbol{\theta}}^{dir} - \mathbf{X}\boldsymbol{\beta}\right)\right] + \alpha\lambda + 2\lambda(1-\alpha)\beta_j, \tag{21}$$

where $\mathbf{x}^j$ corresponds to the j-th column of the design matrix $\mathbf{X}$. The coordinate descent update is then given by

$$\tilde{\beta}_j \leftarrow \frac{S\left(2\left(\boldsymbol{\Sigma}(\widehat{A})^{-1/2}\mathbf{x}^j\right)'\left[\boldsymbol{\Sigma}(\widehat{A})^{-1/2}\left(\widehat{\boldsymbol{\theta}}^{dir} - \widehat{\boldsymbol{\theta}}^{dir}_{(j)}\right)\right], \lambda\alpha\right)}{1 + \lambda(1-\alpha)}, \tag{22}$$

where $\widehat{\boldsymbol{\theta}}^{dir} - \widehat{\boldsymbol{\theta}}^{dir}_{(j)}$ is the partial residual resulting from regularized weighted partial least

squares excluding the contribution of the j-th covariate and the corresponding regression coefficient β_j. $S(\cdot)$ is the soft-thresholding operator with value

$$sign(z,\zeta)(|z|-\zeta)_+ = \begin{cases} z-\zeta & \text{if } z>0 \text{ and } \zeta<|z| \\ z+\zeta & \text{if } z<0 \text{ and } \zeta<|z| \\ 0 & \text{if } \zeta \geq |z| \end{cases} . \tag{23}$$

The β_j are successively updated in that manner until convergence. Further details on the algorithm can be retrieved from Friedman et al. (2007) as well as Friedman et al. (2010). For estimating the model variance parameter A, adjusted maximum likelihood approach proposed by Li and Lahiri (2010) is used. Let

$$\mathcal{L}(A) = c|\boldsymbol{\Sigma}(A)|^{-1/2} exp\left[-\frac{1}{2}(\widehat{\boldsymbol{\theta}}^{dir})'\left(\boldsymbol{\Sigma}(A)^{-1} - \boldsymbol{\Sigma}(A)^{-1}\mathbf{X}(\mathbf{X}'\boldsymbol{\Sigma}(A)^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}(A)^{-1}\right)\widehat{\boldsymbol{\theta}}^{dir}\right]$$

be the likelihood function for of the variance parameter, where c is a generic constant independent from the other likelihood components. The corresponding adjusted maximum likelihood estimate is then obtained from

$$\widehat{A} = \underset{A}{\operatorname{argmax}} \left\{A \cdot \mathcal{L}(A)\right\}. \tag{24}$$

We use standard constrained optimization techniques to solve the upper problem. See Brent (2002) for further details. Adjusting the likelihood function $\mathcal{L}(A)$ by multiplying with a candidate value for A avoids a common problem of the Fay-Herriot model, which is obtaining $\widehat{A} = 0$ as model variance parameter estimate. In such a case, for each shrinkage factor $\gamma_i = 1$ holds. Then the predictor (5) collapses to a synthetic prediction from the underlying linking model, which is often undesirable in the SAE context, as the information regarding the design-unbiased direct estimates is partially ignored.

3.2 Mean squared error estimation

Next, we elaborate on MSE estimation for the empirical best predictor (EBP) under the model in the presence of unknown covariate measurement errors, which is obtained from replacing A and $\boldsymbol{\beta}$ by consistent estimates $\widehat{A}$ and $\widehat{\boldsymbol{\beta}}$ in (11). The MSE of the original Fay-Herriot EBLUP is usually decomposed into three components (Rao and Molina, 2015):

$$MSE\left(\widehat{\theta}_i^{FH}\right) = E\left[\left(\widehat{\theta}_i^{FH} - \theta_i\right)^2\right] = g_{i1} + g_{i2} + g_{i3}, \tag{25}$$

where g_{i1} is due to the estimation of the random effect v_i, g_{i2} stems from the estimation of the regression coefficients $\boldsymbol{\beta}$ and g_{i3} results from the general uncertainty of the chosen estimation method. In the presence of covariate measurement errors, we also have to consider the additional uncertainty resulting from the design matrix perturbations. However, remember that no distributive information regarding the errors is available. Further, the regularized regression coefficient estimates are biased and don't have a closed-form solution

(except for the ℓ_2-regularization). Therefore, it is not clear how an analytic quantification of the MSE components can be derived. Instead, we follow a different strategy for MSE estimation. We first derive the conditional MSE of the BP under known model parameters $\boldsymbol{\beta}$, A, hence $MSE(\widehat{\theta}_i^{BP}|\mathbf{x}_i)$. It can be characterized by

$$\begin{aligned}
MSE\left(\widehat{\theta}_i^{BP}\middle|\,\mathbf{x}_i\right) &= E\left[\left(\widehat{\theta}_i^{BP}-\theta_i\right)^2\right] \\
&= E\left[\left(\gamma_i\widehat{\theta}_i^{dir}+(1-\gamma_i)(\mathbf{x}_i+\boldsymbol{\Delta}_i)'\boldsymbol{\beta}-\mathbf{x}_i'\boldsymbol{\beta}-v_i\right)^2\right] \\
&= E\left[\left(\gamma_i\widehat{\theta}_i^{dir}+\mathbf{x}_i'\boldsymbol{\beta}+\boldsymbol{\Delta}_i'\boldsymbol{\beta}-\gamma_i\mathbf{x}_i'\boldsymbol{\beta}-\gamma_i\boldsymbol{\Delta}_i'\boldsymbol{\beta}-\mathbf{x}_i'\boldsymbol{\beta}-v_i\right)^2\right] \\
&= E\left[\left(\gamma_i\widehat{\theta}_i^{dir}+\boldsymbol{\Delta}_i'\boldsymbol{\beta}-\gamma_i\mathbf{x}_i'\boldsymbol{\beta}-\gamma_i\boldsymbol{\Delta}_i'\boldsymbol{\beta}-v_i\right)^2\right] \\
&= E\left[\left(\gamma_i\left(\widehat{\theta}_i^{dir}-\mathbf{x}_i'\boldsymbol{\beta}-v_i\right)+(1-\gamma_i)\boldsymbol{\Delta}_i'\boldsymbol{\beta}-(1-\gamma_i)v_i\right)^2\right] \\
&= E\left[\left(\gamma_i e_i+(1-\gamma_i)^2(\boldsymbol{\Delta}_i'\boldsymbol{\beta}-v_i\right)^2\right].
\end{aligned}$$

Recall that e_i, v_i are assumed to be independent for $i = 1, ..., m$. As no distributive information regarding the measurement errors is available, the term $\boldsymbol{\Delta}_i'\boldsymbol{\beta}$ is treated as an unknown constant. It follows

$$\begin{aligned}
MSE\left(\widehat{\theta}_i^{BP}\middle|\,\mathbf{x}_i\right) &= \gamma_i^2 E\left[e_i^2\right]+(1-\gamma_i)^2 E\left[\left(\boldsymbol{\Delta}_i'\boldsymbol{\beta}-v_i\right)^2\right] \\
&= \gamma_i^2 D_i+(1-\gamma_i)^2 E\left[\left(\boldsymbol{\Delta}_i'\boldsymbol{\beta}\right)^2-2\boldsymbol{\Delta}_i\boldsymbol{\beta}v_i+v_i^2\right] \\
&= \gamma_i^2 D_i+(1-\gamma_i)^2 A+(1-\gamma_i)^2(\boldsymbol{\Delta}_i'\boldsymbol{\beta})^2.
\end{aligned} \tag{26}$$

Next, as the term $\boldsymbol{\Delta}_i'\boldsymbol{\beta}$ cannot be quantified, we subsitute it by an upper boundary. Remember that the design matrix perturbations are limited by the regularization parameter. For the ℓ_1-regularization the noise $\boldsymbol{\Delta} = (\boldsymbol{\Delta}_1, ..., \boldsymbol{\Delta}_m)'$ is constrained column-wise in terms of the maximum ℓ_2-norm, whereas for the ℓ_2-regularization the maximum singular value of $\boldsymbol{\Delta}$ is restricted. Therefore, we can provide a conservative estimate of the conditional MSE by concluding

$$MSE(\widehat{\theta}_i^{BP}|\mathbf{X}) \leq \gamma_i^2 D_i+(1-\gamma_i)^2 A+(1-\gamma_i)^2(\boldsymbol{\lambda}'\boldsymbol{\beta})^2, \tag{27}$$

where $\boldsymbol{\lambda} \in \mathbb{R}_+^p$ is the expanded vector of the regularization parameter obtained from k-fold cross validation. However, note that (27) corresponds to a worst-case quantification of the MSE. It implies that for each area the biggest possible error is considered by assuming the entire magnitude of feasible noise to be concentrated in the current observation. Further, it implies that the measurement errors in the covariates all have the same direction. Clearly, when estimating the MSE over all $i = 1, ..., m$, the actual noise in the design matrix and the resulting prediction error are then vastly exaggerated. As a result, the corresponding

MSE estimates are highly inefficient. In order to obtain more efficient MSE estimates, some minor assumptions regarding the measurement errors are required. Due to the fact that the absolute errors are bounded by the regularization parameter, assuming that the error $(m)^{-1}\sum_{i=1}^{m}\boldsymbol{\Delta}_i'\boldsymbol{\beta}=0$ we can state, that the maximal error δ_i in area i is $\delta_i = 0.5\cdot\lambda$. Further, by assuming the noise to be spreaded homogeneously over all areas, one obtains

$$\widehat{MSE}(\widehat{\theta}_i^{BP}|\mathbf{X}) \leq \gamma_i^2 D_i + (1-\gamma_i)^2 A + \frac{1}{4m}(1-\gamma_i)^2(\boldsymbol{\lambda}'\boldsymbol{\beta})^2. \tag{28}$$

Note that (28) is less conservative than (27), as it excludes the extreme case of one area containing all the noise. In practise, this assumption could be justified empirically by testing for distributional outliers. It is likely that if all feasible noise is concentrated within one or a small fraction of areas, it is detectable in advance. After quantifying the conditional MSE, we follow the argumentation of Jiang et al. (2002) and apply a Jackknife procedure to account for the uncertainty resulting from the unknown model parameters. For this, let $\widehat{\varphi} := (\widehat{A}, \widehat{\boldsymbol{\beta}})$ and define

$$\widehat{\theta}_i^{EBP} = \widehat{\theta}_i^{EBP}(\widehat{\varphi}) = \widehat{\gamma}_i\widehat{\theta}_i^{dir} + (1-\widehat{\gamma}_i)\tilde{\mathbf{x}}_i'\widehat{\boldsymbol{\beta}} \tag{29}$$

as the EBP under regularization using the information from all areas. Further, let

$$b_i(\widehat{\varphi}) = \widehat{\gamma}_i{}^2 D_i + (1-\widehat{\gamma}_i)^2\widehat{A} + \frac{1}{m}(1-\widehat{\gamma}_i)^2(\boldsymbol{\lambda}'\widehat{\boldsymbol{\beta}})^2. \tag{30}$$

Then the following algorithm is applied:

Algorithm 2 Jackknife for MSE Estimation

1: **for** $j = 1, ..., m$ **do**
2: delete area j from the data set
3: estimate $\widehat{A}_{-j}$ and $\widehat{\boldsymbol{\beta}}_{-j}$ from the remaining areas
4: quantify $\widehat{\theta}_i^{EBP}(\widehat{\varphi}_{-j})$ for all $i = 1, ..., m$ according to (29)
5: quantify $b_i(\widehat{\varphi}_{-j})$ for all $i = 1, ..., m$ according to (30)
6: **end**

The Jackknife estimator for the unconditional MSE of the EBP is finally given by

$$\widehat{MSE}(\widehat{\theta}_i^{EBP}) = b_i(\widehat{\varphi}_i) - \frac{m-1}{m}\sum_{j=1}^{m}[b_i(\widehat{\varphi}_{-j}) - b_i(\widehat{\varphi})] + \frac{m-1}{m}\sum_{j=1}^{m}\left[\widehat{\theta}_i^{EBP}(\widehat{\varphi}_{-j}) - \widehat{\theta}_i^{EBP}(\widehat{\varphi})\right]^2. \tag{31}$$

4 Simulation

4.1 Set up

In order to support the analytical findings numerically, a Monte Carlo simulation with $R = 1000$ iterations ($r = 1, ..., R$) is conducted. For this, a synthetic population of $m = 50$ areas is generated. The area-statistic of interest is generated according to

$$\theta_i^r = (\mathbf{x}_i^r)'\boldsymbol{\beta} + v_i^r, \quad v_i^r \sim N(0, 30^2), \quad \forall\ i = 1, ..., m,$$

where $\mathbf{x}_i^r \in \mathbb{R}^7$ and drawn independently in each iteration from a multivariate normal. The direct estimator $\widehat{\theta}_i^{dir,r}$ for θ_i^r is simulated in each iteration by assigning a random heteroscedastic random error to the area-statistic

$$\widehat{\theta}_i^{dir,r} = \theta_i^r + e_i^r, \quad e_i^r \sim N(0, D_i^r), \quad \forall\ i = 1, ..., m,$$

where D_i^r is drawn in each iteration from $unif(a = 250^2, b = 350^2)$. The predictions are obtained from the following approaches:

- FH: original Fay-Herriot EBLUP (Fay and Herriot, 1979),
- YL: measurement error Fay-Herriot EBLUP (Ybarra and Lohr, 2008),
- CLASSO: corrected ℓ_1-regularized area-level predictor,
- L1: ℓ_1-regularized area-level predictor,
- L2: ℓ_2-regularized area-level predictor,
- EN: elastic net-regularized area-level predictor.

Note that CLASSO represents modification of the corrected LASSO proposed by Loh and Wainwright (2012) that makes it applicable to the Fay-Herriot model. We adopt the suggested correction term including the covariance matrix of the measurement errors into the optimization problem (16). Within the simulation study, normal distributed errors are added to $\mathbf{x}_i$ in each iteration. The errors are generated for every area individually from an area-specific covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\Delta}_i^r)$ in order to meet the distributive assumptions required by Ybarra and Lohr (2008). However, generating the measurement errors in this manner violates the distributive assumptions required by Loh and Wainwright (2012). This has to be considered when looking at the results in the next subsection. Still, as our methodology explicitly doesn't demand any knowledge of the error distribution, and in the light of the Fay-Herriot model, it seems more natural that the measurement error distribution varies across areas. In practise, the covariates are typically aggregated auxiliary information from disjoint geographic, administrative, or contextual units that differ systematically. A total of 8 measurement error scenarios is considered:

- Scenario 1: no errors, true model known,
- Scenario 2: correlated errors in all areas, true model known,

- Scenario 3: uncorrelated errors in 25 areas, true model known,
- Scenario 4: correlated errors in 25 areas, true model known,
- Scenario 5: no errors, true model unknown,
- Scenario 6: correlated errors in all areas, true model unknown,
- Scenario 7: uncorrelated errors in 25 areas, true model unknown,
- Scenario 8: correlated errors in 25 areas, true model unknown.

More details on the measurement error scenarios can be retrieved from the appendix. For those scenarios, in which the true model is unknown, variable selection is performed from a set of 15 potential covariates (7 are relevant) via the Akaike information criterion (Akaike, 1974). However, note that no variable selection is required for L1, EN and CLASSO, as the ℓ_1-norm included in the weighted minimization problem induces a sparse solution for $\widehat{\boldsymbol{\beta}}$. The performance of the predictors is evaluated in terms of the relative root mean squared error (RRMSE):

$$RRMSE(\widehat{\theta}_i^{EBP}) = \frac{\sqrt{1/R \sum_{r=1}^{R} \left(\widehat{\theta}_i^{EBP,r} - \theta_i^r\right)^2}}{\sum_{i=1}^{r} \theta_i^r}.$$

We further look at the relative bias

$$RBias(\widehat{\theta}_i^{EBP}) = \frac{1/R \sum_{r=1}^{R} \left(\widehat{\theta}_i^{EBP,r} - \theta_i^r\right)}{\sum_{r=1}^{R} \theta_i^R}$$

in order to analyze the predictors' behavior over all scenarios in greater detail. The MSE estimation is evaluated by comparing $\widehat{MSE}(\widehat{\theta}_i^{EBP,r})$ with its corresponding Monte Carlo approximation

$$\widehat{MSE}^{MC}(\widehat{\theta}_i^{EBP}) = 1/R \sum_{r=1}^{R} (\widehat{\theta}_i^{EBP,r} - \theta_i^r)^2,$$

with $r = 1, ..., R$ as index of the Monte Carlo iterations. We further look at the coverage rates of 95% prediction intervals for the true value of the area-statistic. They are constructed from the MSE estimates according to

$$\text{Coverage rate} = \frac{\sum_{r=1}^{R} \sum_{i=1}^{m} \mathbb{1}_{PI_{\theta_i^r}}(\theta_i^r)}{R \cdot m} \cdot 100\%,$$

where the 95% prediction interval is defined by

$$PI_{\theta_i^r} = \widehat{\theta}_i^{EBP,r} \pm 1.96 \cdot \sqrt{\widehat{MSE}(\widehat{\theta}_i^{EBP,r})} \cdot \sqrt{1 + 1/m}.$$

4.2 Results

We start by looking at the RRMSE of the area-statistic estimates over all areas. The corresponding results are displayed in Table 1. The performance of the direct estimator (Direct) is included additionally to the model predictors as reference for the influence of the sampling variance implemented in the simulation. Note that the direct estimator does not use any auxiliary information and is thus not affected by covariate measurement errors or variable selection.

Scen	Direct	FH	YL	CLASSO	L1	L2	EN
1	0.07724	0.03207	0.03224	0.03193	0.03128	0.03006	0.03119
2	0.07724	0.04394	0.04272	0.04298	0.04333	0.04216	0.04318
3	0.07724	0.04369	0.04344	0.04312	0.04312	0.04166	0.04287
4	0.07724	0.03754	0.03703	0.03690	0.03646	0.03516	0.03633
5	0.07724	0.04175	0.04215	0.04031	0.03831	0.03868	0.03802
6	0.07724	0.05234	0.05163	0.04921	0.04753	0.04695	0.04705
7	0.07724	0.05429	0.05289	0.05023	0.04867	0.04825	0.04817
8	0.07724	0.04658	0.04655	0.04414	0.04211	0.04208	0.04174

Table 1: Relative root mean squared error over all areas

One can see that the regularized predictors L1, L2 and EN outperform the unregularized predictors FH and YL in terms of the RRMSE in all scenarios. Accordingly, their results are more efficient, which supports the theoretical findings presented in Chapter 2.2. The YL is more efficient than the FH for all scenarios that include measurement errors, hence 2 to 4 and 6 to 8. This is consistent with thel results of Ybarra and Lohr (2008). The CLASSO also outperforms the unregularized predictors in all scenarios as well. Additionally, it is more efficient than L1 and EN in scenario 2. Here, the additional knowledge about the measurement error distribution improves the estimates further. For the other scenarios, it delivers slightly worse results than the other regularized predictors. However, note that with the area-specific measurement error covariance matrices, a distributive assumption of the CLASSO is violated. Another interesting aspect is that the performance difference between the regularized and unregularized predictors increases in the higher-dimensional scenarios 5 to 8, where variable selection is conducted. While the unregularized predictors show a considerable increase in the RRMSE relative to the lower-dimensional scenarios 1 to 4, the regularized predictors suffer only a small efficiency loss. This suggests that identifying the correct model in the presence of covariate measurement errors is methodological problem that can be solved with regularization.

Figure 1 shows the relative deviation of the area-statistic estimates from their true value in percent for scenario 7. The results of Direct, FH and EN are displayed. When comparing the Direct with the FH, it becomes evident that even without regularization the underlying regression model leads to an efficiency gain despite measurement errors. The probability mass of the FH is more concentrated around the point of zero deviation than the Direct. Accordingly, it has a visibly smaller variance. However, the FH is outperformed by the EN. Its advantage in terms of robustness against the additional uncertainty resulting from

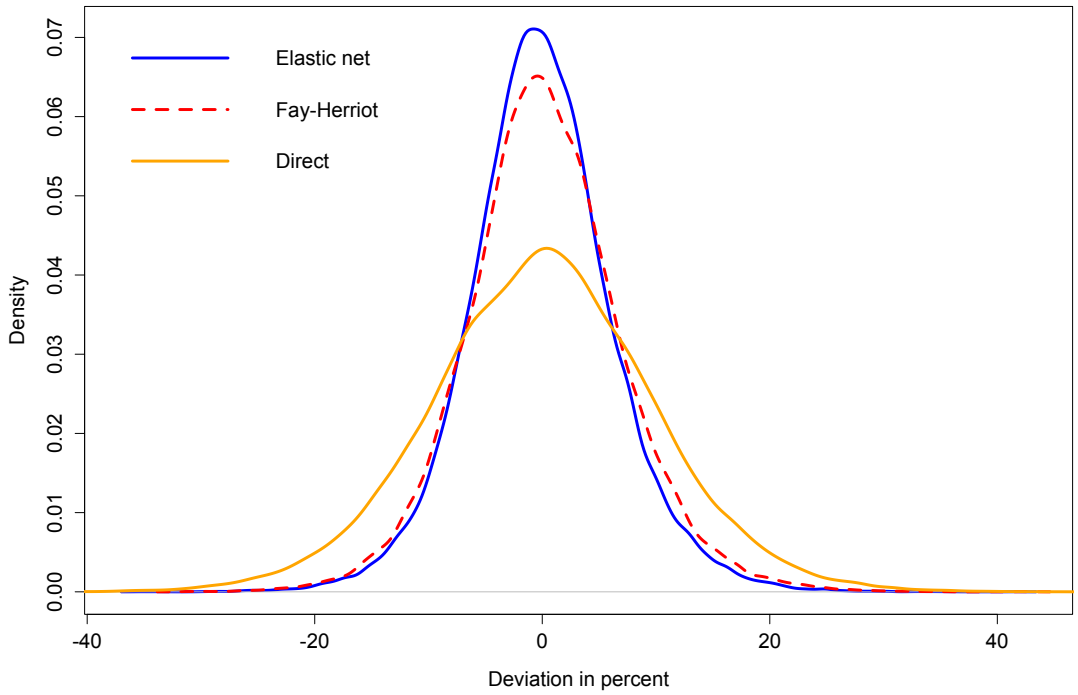

Figure 1: Distribution of the relative deviation, scenario 7

the covariates leads to even more efficient estimates. This underlines the effectiveness of regularization in the estimation process.

Scen	Direct	FH	YL	CLASSO	L1	L2	EN
1	-0.00033	-0.00035	-0.00040	-0.00085	-0.00031	0.00057	-0.00015
2	-0.00033	-0.00167	-0.00177	-0.00147	0.00116	0.00231	0.00134
3	-0.00033	-0.00189	-0.00233	-0.00156	0.00202	0.00317	0.00225
4	-0.00033	-0.00098	-0.00091	-0.00110	0.00041	0.00145	0.00059
5	-0.00033	-0.00051	-0.00053	-0.00105	0.00005	-0.00041	-0.00001
6	-0.00033	-0.00186	-0.00190	-0.00168	0.00058	-0.00032	0.00048
7	-0.00033	-0.00221	-0.00264	-0.00192	0.00081	-0.00029	0.00062
8	-0.00033	-0.00108	-0.00099	-0.00131	0.00024	-0.00035	0.00017

Table 2: Relative bias over all areas

Table 2 displays the relative bias of the area-statistic estimates over all areas. As can be seen, the relative bias is below 0.4% for all predictors over all scenarios, which implies that the bias is generally small. However, the impact of the measurement errors on the bias is very evident. While all predictors have a relatively small bias in scenarios 1 and 5, it increases in the other scenarios. In the scenario 6 to 8, where variable selection under measurement errors is conducted for the FH and YL, the difference in bias is especially evident. The regularized predictors L1, L2 and EN, however, are more robust in that

regard. Their bias increases only slightly (or not at all). This further suggests that the regularized solutions for the model parameter estimates are still decent in the presence of measurement errors.

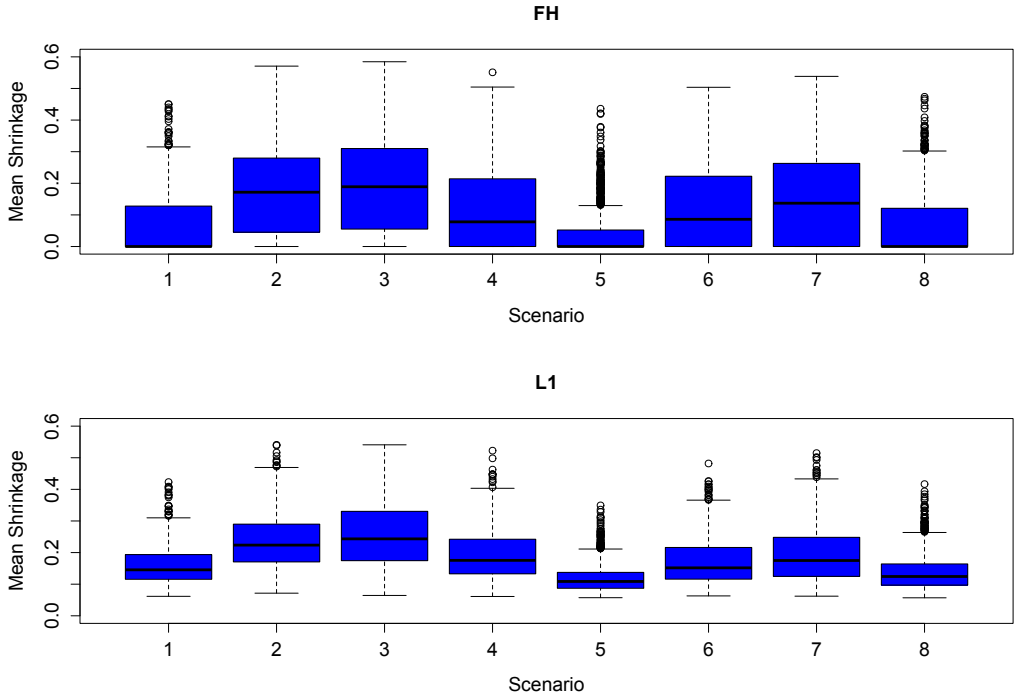

Figure 2: Distribution of the mean shrinkage factor over all areas

Another interesting aspect is the shrinkage behaviour of the predictors within the area-level model. In Figure 2, the distribution of the mean shrinkage factors over all areas per iteration are displayed. The results of the FH and the L1 over all scenarios are included. One can see that the unregularized FH shows a much more volatile shrinkage behaviour compared to the regularized L1. The boxes as well as the whiskers that correspond to the FH are much larger than those of the L1. This suggests that the different realizations of the covariate measurement errors throughout the simulation lead to a very unstable weighting between the direct component and the regression-synthetic component of the FH. The L1, on the other hand, sustains a similar weighting scheme over the iterations. However, note that its general level of shrinkage is higher than for the FH. Accordingly, it puts more emphasize on the direct component, which makes sense in the presence of covariate measurement errors as the underlying regression model is associated with additional uncertainty.

In the following, the results of the MSE estimation are presented. In Table 3, the MSE estimates (Est) of the simulation are displayed next to the Monte Carlo approximation (MC) of the MSE. As can be seen, the MSE is overestimated in all scenarios. For the L1, the relative overestimation ranges from 8.17% to 36.32% with a mean of 23.13%. For the L2, it is from 29.36% to 76.90% with a mean of 50.54%. For the EN, it is from 8.66% to 42.29% with a mean of 25.47%. This tendency of overestimation was generally expected. Using

the expanded regularization parameter vector as substitute for the unknown area-specific covariate perturbations marks a conservative approach to MSE estimation, even with the proposed more practicable approximation. The MSE estimates for the L2 are the most inefficient, as the largest singular values of the design matrix perturbations implemented in the simulation study are always larger than their maximum column-wise ℓ_2-norm. Accordingly, the optimal regularization parameter obtained from cross validation is always the largest for the L2. By squaring the necessary term $\boldsymbol{\lambda}'\boldsymbol{\beta}$, the effect is even stronger, which leads in large MSE estimates.

Scen	L1 MC	L1 Est	L2 MC	L2 Est	EN MC	EN Est
1	15 259	20 253	14 027	22 121	15 154	20 196
2	28 428	30 749	26 851	34 733	28 216	30 658
3	31 732	36 890	30 105	40 273	31 421	36 756
4	21 252	24 672	19 811	27 698	21 086	24 609
5	22 610	30 823	22 861	40 441	22 242	31 647
6	34 064	41 553	33 141	51 387	33 363	42 120
7	37 680	49 211	36 691	57 459	36 875	49 505
8	27 476	34 956	27 211	45 119	26 974	35 461

Table 3: MSE estimation for $m = 50$

However, the MSE estimates become more efficient if the number of areas is increased. Therefore, we conducted an additional simulation run with $m = 90$ areas. The corresponding results are displayed in Table 4. As can be seen, the overestimation tendency decreases as more areas are used for model parameter estimation. L1 overestimates in a range from 4.27% to 27.61% with a mean of 13.57%. For the L2, it is from 9.00% to 35.37% with a mean of 21.89%. The EN overestimates in a range from 4.31% to 27.78% with a mean of 14.64%. This increase in efficiency is mainly due to a decrease in variation resulting from the Jackknife resamples. The corresponding model parameter estimates are more robust against the deletion of a single area when the total number of areas is large.

Scen	L1 MC	L1 Est	L2 MC	L2 Est	EN MC	EN Est
1	8 433	10 761	7 982	10 806	8 403	10738
2	24 640	25 692	24 043	26 207	24 599	25 659
3	19 741	21 324	19 183	21 753	19 686	21 279
4	12 570	13 607	12 077	13 758	12 534	13 573
5	13 048	16 118	14 615	19 726	13 177	16 266
6	27 572	30 591	27 931	34 215	27 457	30 683
7	23 851	26 988	24 662	31 623	23 829	28 948
8	16 958	20 043	18 263	23 206	17 014	20 156

Table 4: MSE estimation for $m = 90$

Based on the MSE estimation for $m = 50$, we construct prediction intervals for the true value of the area-statistic and check their coverage rates within the simulation. The corresponding results can be found in Table 5. While all approaches deliver stable coverage rates

for the scenarios 1 to 5, the FH fails to sustain the 95% level for the higher-dimensional scenarios that include measurement errors. It cannot account for the additional uncertainty resulting from the design matrix perturbations. On the other hand, the prediction interval resulting from the regularized estimation procedure remain stable. The L2 shows the highest coverage rates. All intervals significantly more than the aimed 95% level. This is due to the relatively strong overestimation of the MSE displayed in Table 3.

Scen	FH	L1	L2	EN
1	98.69%	96.45%	97.53%	96.49%
2	95.29%	93.58%	95.55%	93.63%
3	95.53%	95.27%	96.19%	95.35%
4	96.69%	94.81%	96.17%	94.91%
5	94.44%	96.28%	98.32%	96.75%
6	90.07%	94.64%	97.72%	95.05%
7	90.78%	95.86%	97.73%	96.14%
8	92.34%	95.25%	97.86%	95.70%

Table 5: Prediction interval coverage rates

5 Empirical application

After the simulation study, the methodology is applied empirically to poverty mapping in the US. We use estimated income-related figures from the US Census Bureau (US Census Bureau, 2016b) and estimated crime records from the Uniform Crime Reporting (UCR) Program (Uniform Crime Reporting (UCR) Program, 2016) as auxiliary information to quantify of the number of people below 100% of the federal poverty threshold per state. We take existing state-level estimates of the corresponding statistics obtained from the Current Population Survey (US Census Bureau, 2016a), which is conducted in a collaboration of the US Census Bureau and the Bureau of Labor Statistics. It encloses roughly 60 000 households. We robustify and improve these estimates by applying our regularized estimation approach. Since the original estimates are published with the corresponding standard errors, we don't have to use a generalized variance function in order to quantify the sampling variances. Further, as the auxiliary information values we use are also estimated, the design matrix created from them in the estimation process is associated with uncertainty. Therefore, the data situation fits naturally in our framework. All numbers correspond to the report year 2015. A list of all considered variables can be found in the appendix.

We use the elastic net regularization for poverty mapping. This choice is made for two reasons. First, some variables within the considered data sets may not be relevant for the functional description of the area-statistic of interest. As the elastic net is a sparsity inducing regularization, its application leads to an automatic variable selection. Second, within the set of relevant variables, there may be grouping structures. The elastic net has been found to perform better than the ℓ_1-regularization in terms of variable selection

and prediction in the presence of strong correlation within the covariates (Zou and Hastie, 2005). As pointed out in Chapter 3, the regularization parameter λ is obtained from k-fold cross validation. Applying the robustified area-level modelling in the described manner yields the following results.

State	Model	Former	FH	Former Var	Model MSE	FH MSE
Alabama	710	784	704	3 600	356	365
Alaska	62	65	65	36	38	65
Arizona	1 076	1 156	1 110	7 056	592	731
Arkansas	487	475	494	729	196	331
California	5 439	5 441	5 396	44 521	5 886	39 518
Colorado	570	538	525	4 761	363	850
Connecticut	379	324	341	1 849	173	310
Delaware	107	106	103	144	45	114
Distr. Columbia	120	113	112	64	128	89
Florida	3 017	3 253	3 117	30 976	2 972	5 714
Georgia	1 805	1 833	1 888	12 321	1 781	1 744
Hawaii	159	151	147	256	115	203
Idaho	197	204	187	441	83	107
Illinois	1 491	1 380	1 498	10 816	1 551	2 560
Indiana	780	880	784	5 184	290	654
Iowa	333	321	301	1 521	151	166
Kansas	333	404	329	1 764	123	244
Kentucky	760	851	758	3 249	651	509
Louisiana	871	854	869	3 249	429	717
Maine	138	165	170	400	67	188
Maryland	568	566	552	4 900	1 112	2 144
Massachusetts	808	782	757	4 225	617	1 007
Michigan	1 233	1 259	1 160	9 604	859	5 655
Minnesota	498	428	492	3 364	356	425
Mississippi	538	563	548	1 089	277	579
Missouri	686	582	622	5 041	525	1 166
Montana	117	121	121	144	63	136
Nebraska	204	192	206	324	76	94
Nevada	382	371	398	1 296	441	872
New Hampshire	86	94	94	169	56	141
New Jersey	929	1 004	896	9 025	968	863
New Mexico	426	400	425	784	222	340
New York	2 609	2 791	2 747	24 336	14 334	11 737
North Carolina	1 374	1 509	1 452	9 025	2 087	3 509
North Dakota	83	82	82	144	49	121
Ohio	1 541	1 550	1 690	11 025	721	3 068
Oklahoma	556	551	526	2 025	309	301
Oregon	506	478	503	4 761	231	443
Pennsylvania	1 483	1 542	1 470	11 449	2 226	1 631

Rhode Island	122	123	122	256	70	110
South Carolina	834	683	800	3 136	536	685
South Dakota	117	118	115	100	43	92
Tennessee	982	973	1 008	4 900	438	407
Texas	4 075	4 036	4 261	42 849	3 842	5 623
Utah	239	279	245	900	245	339
Vermont	60	65	62	64	37	73
Virginia	1 040	894	923	6 561	2 284	1 718
Washington	855	819	824	2 916	1 457	1 402
West Virginia	258	261	234	2 116	141	235
Wisconsin	714	654	682	4 624	439	339
Wyoming	56	56	65	64	31	69

Table 6: Number of people below 100% of the federal poverty threshold in thousands

Table 6 shows the number of people below 100% of the federal poverty threshold per state in thousands. The *Model* column corresponds to the model estimates from regularized estimation approach, while the *Former* column displays to the original estimates provided by the US Census Bureau. The *FH* column shows the results of the original unregularized Fay-Herriot EBLUP. The *Former Var* column corresponds to the variance of the originally published estimates of the US Census Bureau. The columns *Model MSE* and *FH MSE* display the estimated MSEs of the respective model estimates. As can be seen, there are some differences between the model estimates and the original estimates. While the original nationwide estimate is 43 124 000, the regularized approach obtains 42 818 000 and the Fay-Herriot EBLUP estimates 42 980 000. This marks a difference of -0.71% and -0.33%, respectively. On the level of the federal states, the differences are partially larger. However, given the relatively large variances of the original estimates, no model estimate is implausible. The advantage of the model estimates is that they are more efficient in terms of the MSE. While the average variance of the original estimates is 5 964, the average MSE of the regularized estimates is 1 002. The average MSE of the Fay-Herriot EBLUP is 1971. Accordingly, the regularized modelling approach allows for a decisive efficiency gain compared to both the original estimates as well as the results of the unregularized Fay-Herriot approach. Another interesting aspect is the difference in the shrinkage behaviour between the regularized estimates and the Fay-Herriot results. As already pointed out in the simulation study, the Fay-Herriot tends to shrink less towards the original estimates than the regularized approach. It puts more emphasize on the model component, which can be problematic in the presence of uncertainty in the design matrix.

Figure 3 is a map of the estimated percentage of people below 100% of the federal poverty threshold. It is obtained from dividing the regularized model estimates by the population size, which is retrieved from the US Census Bureau. As can be seen, the highest poverty rates are located in the south and in the south-east. The highest estimated percentage is in New Mexico with approximately 20.48%, followed by Louisiana (18.66%), Mississippi (18.00%), the District of Columbia (17.90%), Georgia (17.70%) and Kentucky (17.18%). The lowest estimated percentage is in New Hampshire with approximately 6.47%, followed

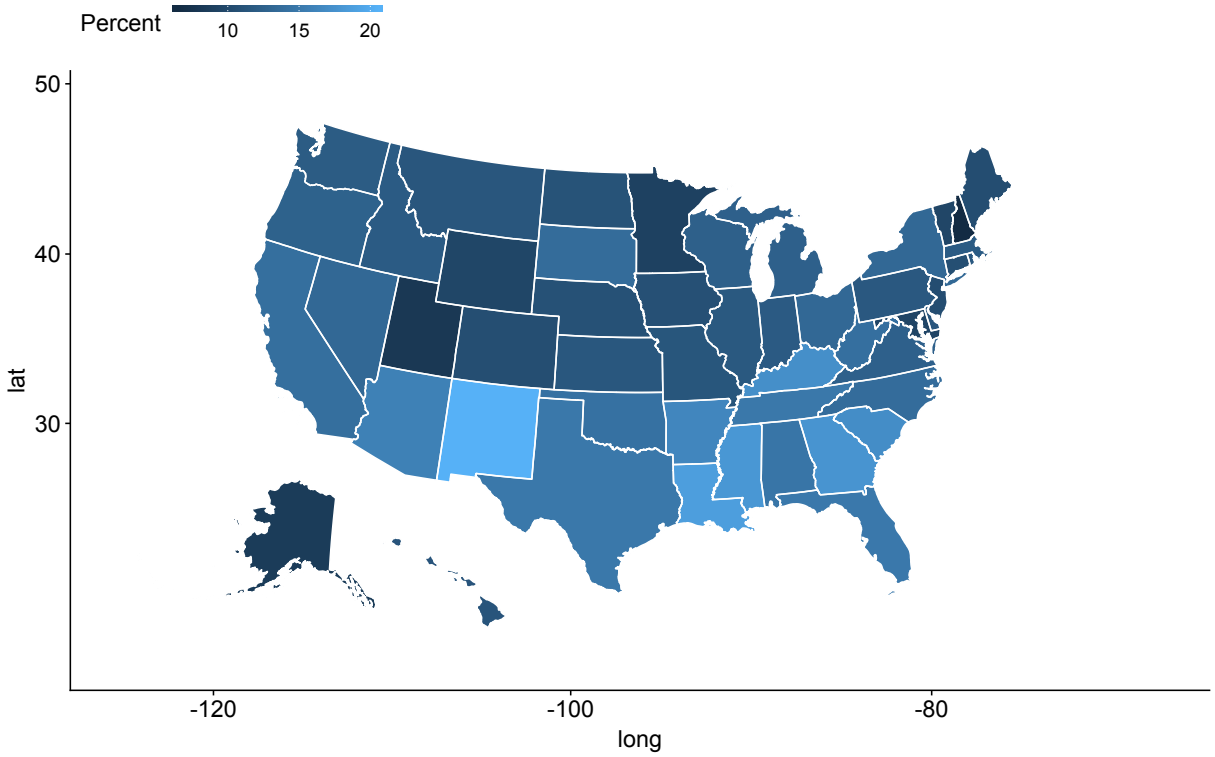

Figure 3: Percentage of people below 100% of the federal poverty threshold

by Utah (8.00%), Alaska (8.40%), Minnesota (9.08%), Maryland (9.55%) and Wyoming (9.58%).

6 Conclusion

A robust extension to the Fay-Herriot model under covariate measurement errors was proposed. We showed that regularized regression coefficient estimation is equivalent to robust optimization under additive noise when loss and regularization are strictly monotonously increasing, bijective functions of seminorms and norms. Applying this equivalence, we derived a model parameter estimation procedure that is easy to implement, delivers efficient results in the presence of design matrix perturbations, and does not require distributive information about the measurement error. It further allows for stable area-statistic estimates from small samples and, depending on the choice of the regularization, even performs model parameter estimation and variable selection simultaneously. Therefore, it is particularly relevant for a variety of SAE applications where the available auxiliary information is subject to uncertainty. Due to its general formulation, the presented equivalence has implications for a much broader range of regression problems, for example in linear mixed model theory or time series analysis. However, as our focus is on SAE, we limit the discus-

sion on future research to that field, even though several aspects can be translated beyond its scope.

An important question is how to optimally choose the regularization. Usually, the regularization is determined by observable distributive characteristics of the covariates (e.g. grouping structures or multicollinearity), or depending on the researchers ideas towards the true structure of the regression coefficients (e.g. sparsity). But in the light of the theoretical findings, when applying regularization as robustification against covariate measurement errors, one has to further consider the potential structure of the noise in the design matrix. As this noise is, however, unobservable, it is currently not obvious how to consider it for the optimal regularization choice. A possible approach could be to perform k-fold cross validation on multiple grids, where each grid corresponds to a regularization parameter for a specific regularization. Another question is the efficiency of the MSE estimation when the number of areas is small. Especially for the squared ℓ_2-norm, the proposed Jackknife estimator overestimates the actual MSE considerably. Therefore, it should be investigated how the MSE estimation procedure can be adjusted depending on the regularization.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control 19*(6), 716–723.

Ben-Tal, A., L. El Ghaoui, and A. Nemirovski (2009). *Robust optimization*, Volume 28. Princeton University Press.

Bertsimas, D. and M. S. Copenhaver (2018). Characterization of the equivalence of robustification and regularization in linear and matrix regression. *European Journal of Operational Research 270*(3), 931–942.

Bertsimas, D., M. S. Copenhaver, and R. Mazumder (2017). The trimmed lasso: Sparsity and robustness. *arXiv preprint arXiv:1708.04527*.

Boyd, S. and L. Vandenberghe (2009). *Convex optimization*. Cambridge university press.

Brent, R. P. (2002). *Algorithms for minimization without derivatives*. Dover Books on Mathematics, Prentice-Hall series in automatic computation. Courier Corporation.

El Ghaoui, L. and H. Lebret (1997). Robust solutions to least-squared problems with uncertain data. *SIAM Journal on matrix analysis and applications 18*(4), 1035–1064.

Fay, R. E. and R. A. Herriot (1979). Estimates of income for small places: An application of james-stein procedures to census data. *Journal of the American Statistical Association 74*(366), 269–277.

Friedman, J., T. Hastie, H. Höfling, and R. Tibshirani (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics 1*(2), 302–332.

Friedman, J., T. Hastie, H. Höfling, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software 33*(1).

Hoerl, A. and R. Kennard (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Techometrics 12*(1), 55–67.

Huber, P. J. (1973). Robust regression: Asymptotics, conjectures and monte carlo. *The Annals of Statistics 1*, 799–821.

Jiang, J., P. Lahiri, and S.-M. Wan (2002). A unified jackknife theory for empirical best prediction with m-estimation. *The Annals of Statistics 30*(6), 1782–1810.

Li, H. and P. Lahiri (2010). Adjusted maximum likelihood method for solving small area estimation problems. *Journal of Multivariate Analysis 101*, 882–892.

Loh, P.-L. and M. J. Wainwright (2012). High-dimensional regression with noisy and missing covariates: Provable guarantees with nonconvexity. *The Annals of Statistics 40*(3), 1637–1664.

Markovsky, I. and S. V. Huffel (2007). Overview of total least-squares methods. *Signal Processing 87*, 2283–2302.

Molina, I., J. N. K. Rao, and G. S. Datta (2015). Small area estimation under a fay-herriot model with preliminary testing for the presence of random area effects. *Survey Methodology 41*(1), 1–19.

Rao, J. N. K. and I. Molina (2015). *Small Area Estimation* (2 ed.). Wiley Series in Survey Methodology. New Jersey: John Wiley & Sons.

Rousseeuw, P. J. and A. M. Leroy (2003). *Robust regression and outlier detection*. John wiley & sons.

Sinha, S. K. and J. N. K. Rao (2009). Robust small area estimation. *The Canadian Journal of Statistics 37*, 381–399.

Slud, E. V. and T. Maiti (2011). Small-area estimation based on survey data from a left-censored fay-herriot model. *Journal of Statistical Planning and Inference 141*(11), 3520–3535.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological) 58*(1), 267–288.

Uniform Crime Reporting (UCR) Program (2016). Crime in the united states by state, 2015. Online. URL: https://ucr.fbi.gov/crime-in-the-u.s/2015/crime-in-the-u.s.-2015/tables/table-5, retrieved on 11/29/2018.

US Census Bureau (2016a). Current population survey (cps). Online. URL: https://www.census.gov/programs-surveys/cps.html, retrieved on 11/29/2018.

US Census Bureau (2016b). Pov-46. poverty status by state. Online. URL: https://www.census.gov/data/tables/time-series/demo/income-poverty/cps-pov/pov-46.2016.html, retrieved on 11/29/2018.

Xie, D., T. E. Raghunathan, and J. M. Lepkowski (2007). Estimation of the proportion of overweight individuals in small areas - a robust extension of the fay-herriot model. *Statistics in Medicine 26*(13), 2699–2715.

Ybarra, L. M. R. and S. L. Lohr (2008). Small area estimation when auxiliary information is measured with error. *Biometrika 95*, 919–931.

Yoshimori, M. and P. Lahiri (2014). A new adjusted maximum likelihood method for the fay-herriot small area model. *Journal of Multivariate Analysis 124*, 281–294.

You, Y. S. and Q. Zhou (2011). Hierarchical bayes small area estimation under a spatial model with application to health survey data. *Survey Methodology 37*(1), 25–36.

Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B (Methodological) 67*(2), 301–320.

Appendix

Data for the empirical application

The following variables were considered on the state-level for the empirical application of the methodology in Chapter 5.

Variable	**Source**
Median household income	US Census Bureau
Percentage of households with income below 10.000$	US Census Bureau
Percentage of households on the Supplemental Nutrition Assistance Program	US Census Bureau
Unemployment rate	US Census Bureau
Number of people below 50% of the federal poverty threshold	US Census Bureau
Number of people below 100% of the federal poverty threshold	US Census Bureau
Number of violent crimes per 100.000	UCR Program
Number of murders and nonnegligent manslaughters per 100.000	UCR Program
Number of rapes per 100.000, legacy definition	UCR Program
Number of rapes per 100.000, revised definition	UCR Program
Number of robberies per 100.000	UCR Program
Number of aggravated assaults per 100.000	UCR Program
Number of property crimes per 100.000	UCR Program
Number of burglaries per 100.000	UCR Program
Number of larceny-thefts per 100.000	UCR Program
Number of motor vehicle thefts per 100.000	UCR Program

Measurement error scenarios

The design matrix perturbations are drawn for each iteration of the simulation individually from multivariate normal distributions with fixed area-specific covariance matrices $\mathbf{\Sigma}_{\mathbf{\Delta}_1^r}, ..., \mathbf{\Sigma}_{\mathbf{\Delta}_m^r}$. Note that these measurement error covariance matrices differ per scenario. They are generated according to the following procedure.

Scenario	Generation	Areas
1	$\mathbf{\Sigma}(\mathbf{\Delta}_i^r) = \mathbf{0}_{7,7}$	$i = 1, ..., m$
2	$\mathbf{\Sigma}(\mathbf{\Delta}_i^r) \in \mathbb{R}_+^{7\times7}$ positive-definite, $\sigma_{j,l}^r \sim unif(a = 40, b = 60)$	$i = 1, ..., m$
3	$\mathbf{\Sigma}(\mathbf{\Delta}_i^r) = diag(\sigma_1^r, ..., \sigma_7^r)$, $\sigma_j^r \sim unif(a = 40, b = 60)$	$i = 1, ..., (m/2)$
3	$\mathbf{\Sigma}(\mathbf{\Delta}_i^r) = \mathbf{0}_{7,7}$	$i = (m/2 + 1), ..., m$
4	$\mathbf{\Sigma}(\mathbf{\Delta}_i^r) \in \mathbb{R}_+^{7\times7}$ positive-definite, $\sigma_{j,l}^r \sim unif(a = 40, b = 60)$	$i = 1, ..., (m/2)$
4	$\mathbf{\Sigma}(\mathbf{\Delta}_i^r) = \mathbf{0}_{7,7}$	$i = (m/2 + 1), ..., m$
5	$\mathbf{\Sigma}(\mathbf{\Delta}_i^r) = \mathbf{0}_{15,15}$	$i = 1, ..., m$
6	$\mathbf{\Sigma}(\mathbf{\Delta}_i^r) \in \mathbb{R}_+^{15\times15}$ positive-definite, $\sigma_{j,l}^r \sim unif(a = 40, b = 60)$	$i = 1, ..., m$
7	$\mathbf{\Sigma}(\mathbf{\Delta}_i^r) = diag(\sigma_1^r, ..., \sigma_{15}^r)$, $\sigma_j^r \sim unif(a = 40, b = 60)$	$i = 1, ..., (m/2)$
7	$\mathbf{\Sigma}(\mathbf{\Delta}_i^r) = \mathbf{0}_{15,15}$	$i = (m/2 + 1), ..., m$
8	$\mathbf{\Sigma}(\mathbf{\Delta}_i^r) \in \mathbb{R}_+^{15\times15}$ positive-definite, $\sigma_{j,l}^r \sim unif(a = 40, b = 60)$	$i = 1, ..., (m/2)$
8	$\mathbf{\Sigma}(\mathbf{\Delta}_i^r) = \mathbf{0}_{15,15}$	$i = (m/2 + 1), ..., m$