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

Jan Pablo Burgard
Joscha Krause
Dennis Kreber
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 $\ell_1$-norm, squared $\ell_2$-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 $\ell_1$-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 $\ell_1$-norm, squared $\ell_2$-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-Herriot 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 threshold 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 $\hat{\theta}_i^{\text{dir}} \in \mathbb{R}$ be a direct estimator of $\theta_i$ that is available for all $i = 1, ..., m$. It is assumed to be design-unbiased, hence $E(\hat{\theta}_i^{\text{dir}}) = \theta_i$, and obtained from only using the sample information within $S_i$. Due to the small area-specific sample size $n_i$, its variance $\text{Var}(\hat{\theta}_i^{\text{dir}})$ is too large in order to draw reliable conclusions on $\theta_i$. Thus, the objective is to find a better estimator of $\theta_i$, denoted by $\hat{\theta}_i \in \mathbb{R}$, 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 $\theta_i$ by using suitable auxiliary information. The model consists of two components. The first component (sampling model) states that due to the design-unbiasedness of $\hat{\theta}_i^{\text{dir}}$, the direct estimator is equal to the unknown $\theta_i$ plus some random sampling error

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

with $e_i$ as sampling error and $D_i = \text{Var}(\hat{\theta}_i^{\text{dir}}|\theta_i)$ 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 $\theta_i$ as random and establishes a linear relation to some area-level auxiliary information

$$\theta_i = x_i' \beta + v_i, \quad v_i \sim N(0, A), \quad \forall \ i = 1, ..., m, \quad (2)$$

with $x_i \in \mathbb{R}^p$ as vector of auxiliary information and $\beta \in \mathbb{R}^p$ 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

$$\hat{\theta}_i^{\text{dir\, ind}} \sim N(x_i' \beta, D_i + A), \quad \forall \ i = 1, ..., m, \quad (3)$$

or, in matrix notation, $\hat{\theta}^{\text{dir\, ind}} \sim MVN(X \beta, \Sigma(A))$, with $\hat{\theta}^{\text{dir}} = (\hat{\theta}_1^{\text{dir}}, ..., \hat{\theta}_m^{\text{dir}})'$ as response vector, $X = (x_1, ..., x_m)'$ as design matrix, and $\Sigma(A) = \text{diag}(A + D_1, ..., A + D_m)$ as covariance matrix of the variance components. The basic idea of the Fay-Herriot model is to improve the direct estimator $\hat{\theta}_i^{\text{dir}}$ by exploiting the functional relation between $\theta_i$ and $x_i$. In order to determine the functional relation, the unknown model parameters $A$ and $\beta$ have to be estimated. This is usually performed iteratively by finding some initial model variance.
estimate $\hat{A}$ first, and then obtain the regression coefficient estimates $\hat{\beta}$ conditionally on $\hat{A}$ according to

$$\hat{\beta} = \arg\min_{\beta} \left\| \Sigma(\hat{A})^{-1/2} \left( \hat{\theta}^{\text{dir}} - X\beta \right) \right\|^2, \quad \Sigma(\hat{A}) = \text{diag}(D_1 + \hat{A}, ..., D_m + \hat{A}).$$  \quad \text{(4)}$$

Afterwards, the model variance estimate $\hat{A}$ is updated conditionally on $\hat{\beta}$ 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 $\theta_i$ is improved through considering more information relative to the direct estimator $\hat{\theta}_{i,\text{dir}}$, that only considers information from area $i$. This often referred to as borrowing strength.

The final estimator $\hat{\theta}_{i,\text{FH}}$ under the Fay-Herriot model is then obtained from a convex linear combination of $\hat{\theta}_{i,\text{dir}}$ and the regression-synthetic component $x_i'\hat{\beta}$ of the model (Molina et al., 2015):

$$\hat{\theta}_{i,\text{FH}} = \gamma_i\hat{\theta}_{i,\text{dir}} + (1 - \gamma_i)x_i'\hat{\beta}, \quad \forall \ i = 1, ..., m,$$

with $\gamma_i = \hat{A}/(\hat{A} + D_i)$ 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 $\hat{\theta}_{i,\text{FH}}$ being shrunken towards $\hat{\theta}_{i,\text{dir}}$ as a result of variance weighting. It is not to be confused with the shrinkage of some $\beta_j \in \hat{\beta}$ towards zero due to regularization, which is described in the next subsection. If $\hat{\beta} = (X'\Sigma(\hat{A})^{-1}X)^{-1}X'\Sigma(\hat{A})^{-1}\hat{\theta}^{\text{dir}}$, and $\hat{A}$ is a consistent estimator of $A$, then $\hat{\theta}_{i,\text{FH}}$ is the EBLUP for $\theta_i$. 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

$$\tilde{x}_i = x_i + \Delta_i, \quad \forall \ i = 1, ..., m,$$

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

$$\hat{\theta}_{i,\text{dir}} = (x_i + \Delta_i)'\beta + v_i + e_i, \quad \forall \ i = 1, ..., m.$$  \quad \text{(7)}$$

Note that no distributive assumption regarding the measurement errors is made. We treat $\Delta_i$ as a fixed unobservable perturbation of $x_i$. Under the model (7), the conditional distributions of the direct estimator $\hat{\theta}_{i,\text{dir}}$ are given by

$$\hat{\theta}_{i,\text{dir}} | x_i, v_i \sim N((x_i + \Delta_i)'\beta + v_i, D_i)$$

$$\hat{\theta}_{i,\text{dir}} | x_i \sim N((x_i + \Delta_i)'\beta + v_i, D_i + A).$$  \quad \text{(8)}$$
Assuming the model variance $A$ and the sampling variances $D_i$ to be known, for the conditional distribution of the random effect

$$f(v_i|\hat{\theta}_i^{\text{dir}}, x_i) \propto f(v_i)f(\hat{\theta}_i^{\text{dir}}|x_i, v_i)$$

holds, where

$$f(v_i)f(\hat{\theta}_i^{\text{dir}}|x_i, v_i) = \frac{1}{\sqrt{2\pi}A} \exp\left(-\frac{v_i^2}{2A}\right) \frac{1}{\sqrt{2\pi}D_i} \exp\left(-\frac{(\hat{\theta}_i^{\text{dir}} - (x_i + \Delta_i)'\beta - v_i)^2}{2D_i}\right)$$

$$\propto \exp\left(-\frac{v_i^2}{2A}\right) \exp\left(-\frac{v_i^2}{2D_i} - 2v_i\left(\hat{\theta}_i^{\text{dir}} - (x_i + \Delta_i)'\beta\right)\right)$$

$$= \exp\left(-\frac{v_i^2}{2D_i} \left(1 - \frac{1}{D_i}\right) + \frac{\hat{\theta}_i^{\text{dir}} - (x_i + \Delta_i)'\beta}{D_i} v_i\right)$$

$$= \exp\left(-\frac{v_i^2}{2D_i A(D_i+A)} + \frac{1 + \frac{1}{D_i} A}{D_i A(D_i+A)} \frac{\hat{\theta}_i^{\text{dir}} - (x_i + \Delta_i)'\beta}{D_i A(D_i+A)} v_i\right).$$

Accordingly, the conditional distribution is a univariate normal

$$v_i|\hat{\theta}_i^{\text{dir}}, x_i \sim N\left(A\left(\hat{\theta}_i^{\text{dir}} - (x_i + \Delta_i)'\beta\right), \frac{D_i A(D_i+A)}{D_i A(D_i+A)}\right).$$

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

$$\hat{\theta}_i^{BP} = x'_i\beta + E\left(\Delta_i|x_i, \hat{\theta}_i^{\text{dir}}\right) + E\left(v_i|x_i, \hat{\theta}_i^{\text{dir}}\right)$$

$$= x'_i\beta + \Delta'_i\beta + \frac{A}{A + D_i} \left(\hat{\theta}_i^{\text{dir}} - x'_i\beta - \Delta'_i\beta\right)$$

$$= \frac{A}{A + D_i} \hat{\theta}_i^{\text{dir}} + \frac{D_i}{A + D_i} (x'_i\beta + \Delta'_i\beta)$$

$$= \gamma_i \hat{\theta}_i^{\text{dir}} + (1 - \gamma_i) (x'_i\beta + \Delta'_i\beta)$$

$$= \gamma_i \hat{\theta}_i^{\text{dir}} + (1 - \gamma_i) \tilde{x}'_i\beta.$$

### 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 $\hat{\beta}$ 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 $g: \mathbb{R}^n \rightarrow \mathbb{R}$, a set $\mathcal{U} \subseteq \mathbb{R}^{n \times p}$, a design matrix $X \in \mathbb{R}^{n \times p}$ and a response vector $y \in \mathbb{R}^n$, the min-min approach is formulated by the optimization problem

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

while the min-max approach is characterized by the problem

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

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 $X$ pessimistically, i.e., we use the min-max approach to introduce robustness to our estimate. That is, we are looking at the optimization model

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

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

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

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 $g : \mathbb{R}^n \rightarrow \mathbb{R}$ is a seminorm which is not identically zero and $h : \mathbb{R}^p \rightarrow \mathbb{R}$ is a norm, then for any $z \in \mathbb{R}^n$ and $\beta \in \mathbb{R}^p$

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

where

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

Clearly, the Proposition directly implies that

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

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 $\lambda$ directly constrains 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 $X$. Under this interference, $\beta$ 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 $\beta$ while the other player tries to maximize the deviation by controlling the noise, which is added to $X$. 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

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

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

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

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 $\|\cdot\|_2^2$ 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 $g : \mathbb{R}^n \rightarrow \mathbb{R}$ and $h_1, h_2, \ldots, h_d : \mathbb{R}^n \rightarrow \mathbb{R}$ be convex functions. If

$$
\hat{z} \in \arg\min_{z \in \mathbb{R}^n} g(z) + \sum_{i=1}^d \lambda_i h_i(z)
$$

is the solution of the optimization problem of Proposition 1 with $(\lambda_1, \ldots, \lambda_d)$, then the solutions of the optimization problems

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

and

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

correspond to the specific cases $g = h = \ell_2$ and $g = \ell_2, h = \ell_1$, respectively.
for the parameters $\lambda_1, \ldots, \lambda_d > 0$, then there exist $c_1, \ldots, c_d > 0$ such that

$$
\hat{z} \in \arg\min_{z \in \mathbb{R}^n} g(z) \quad \text{s.t.} \quad h_i(z) \leq c_i \quad \text{for all } 1 \leq i \leq d
$$

(14)

and vice versa, if there is a $z$ such that $h_i(z) < c_i$ 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 $\hat{z}$ is an optimal solution for both problems.

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

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

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

Now assume that $\hat{z}$ 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 $z$ such that $h_i(z) < c_i$ for all $1 \leq i \leq d$. Thus, strong duality holds. It follows that

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

(15)

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

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

holds, which proves the conjecture. □

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

$$
\arg\min_{\beta} \max_{\Delta \in \mathcal{U}} g(y - (X + \Delta)\beta) = \arg\min_{\beta} f(g(y - X\beta)) + \sum_{i=1}^d \lambda_i f_i(h_i(\beta))
$$

where

$$
\mathcal{U} = \left\{ \Delta : g(\Delta \gamma) \leq \sum_{i=1}^d \mu_i h_i(\gamma) \text{ for all } \gamma \in \mathbb{R}^p \right\}
$$
**Proof.** We first look at the right-hand side minimization problem

\[ M := \arg\min_{\beta} f(g(y - X\beta)) + \sum_{i=1}^{d} \lambda_i f_i(h_i(\beta)). \]

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

\[ M = \arg\min_{\beta} f(g(y - X\beta)) \]

\[ \text{s.t. } f_i(h_i(\beta)) \leq c_i \text{ for all } 1 \leq i \leq d. \]

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

\[ M = \arg\min_{\beta} g(y - X\beta) \]

\[ \text{s.t. } h_i(\beta) \leq f_i^{-1}(c_i) \text{ for all } 1 \leq i \leq d. \]

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

\[ M = \arg\min_{\beta} g(y - X\beta) + \sum_{i=1}^{d} \mu_i h_i(\beta) \]

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

\[ M = \arg\min_{\beta} \max_{\Delta \in \mathcal{U}} g(y + (X + \Delta)\beta) \]

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

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.

**\( \ell_1 \)-regularization**

The first regularization is the \( \ell_1 \)-norm, which is famously used for the LASSO introduced by Tibshirani (1996) for linear regression. Within the Fay-Herriot model, including an
\( \ell_1 \)-norm regularization changes the weighted minimization problem (4) for obtaining \( \hat{\beta} \) to
\[
\hat{\beta}_{\ell_1} = \arg\min_{\beta} \left| \left| \left| \Sigma(A)^{-1/2} \left( \hat{\theta}^{\text{dir}} - X\beta \right) \right| \right|_2^2 + \lambda \left| \right| \beta \left| \right|_1 \right. ,
\]
(16)

Applying Proposition 2 with \( g(z) := \|z\|_2 \), \( h(z) = \|z\|_1 \), \( f(z) = z^2 \) and \( f_1(z) = z \) yields the equivalence
\[
\hat{\beta}_{\ell_1} = \arg\min_{\beta} \max_{\Delta \in U_{\ell_1}} \left| \left| \Sigma(A)^{-1/2} \left( \hat{\theta}^{\text{dir}} - X\beta \right) + \Delta \beta \right| \right|_2
\]
for some \( \mu > 0 \) and \( U_{\ell_1} = \{ \Delta : \|\Delta \gamma\|_2 \leq \mu \|\gamma\|_1 \text{ for all } \gamma \in \mathbb{R}^p \} \). From the formulation (17), it is evident that the coefficients are robustified against noise by introduction of the perturbation \( \Delta \), which is maximized in regard to the fitted coefficients and is constrained by the uncertainty set \( U_{\ell_1} \). The following result by Bertsimas and Copenhaver (2018) improves the interpretability of the uncertainty set \( U_{\ell_1} \).

**Proposition 3** (Bertsimas and Copenhaver (2018)). Let be \( p \in [1, \infty] \), let \( \|\beta\|_0 \) be the number of non-zero entries of \( \beta \) and let \( \Delta_i \) be the \( i \)-th column of \( \Delta \). If
\[
U' = \{ \Delta : \|\Delta \beta\|_2 \leq \mu \|\beta\|_0 \text{ } \forall \|\beta\|_p \leq 1 \}
\]
and
\[
U'' = \{ \Delta : \|\Delta_i\|_2 \leq \mu \text{ } \forall i \}
\]
then \( U_{\ell_1} = U' = U'' \).

Thus, in case of the \( \ell_1 \)-regularization the noise is constrained column-wise by the parameter \( \mu \). 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 \( \ell_1 \)-norm in the minimization problem induces a sparse solution for \( \hat{\beta} \). As a result, some elements \( \hat{\beta}_j \in \hat{\beta} \) that are irrelevant for the functional description of \( \hat{\theta}^{\text{dir}} \) are set exactly to zero in the estimation process. This implies an automatic variable selection, which makes the \( \ell_1 \)-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).

**\( \ell_2 \)-regularization**

The second regularization is the squared \( \ell_2 \)-norm, which is famously used for ridge regression proposed by Hoerl and Kennard (1970). The corresponding weighted minimization problem to determine \( \hat{\beta} \) can be stated as
\[
\hat{\beta}_{\ell_2} = \arg\min_{\beta} \left| \left| \Sigma(A)^{-1/2} \left( \hat{\theta}^{\text{dir}} - X\beta \right) \right| \right|_2^2 + \lambda \left| \right| \beta \left| \right|_2^2
\]
(18)

Note that using the squared \( \ell_2 \)-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 \( g(z) = \|z\|_2 \), \( h(z) = \|z\|_2 \), \( f(z) = z^2 \) and \( f_1(z) = z^2 \) we obtain

\[
\hat{\beta}_{\ell_2} = \arg\min_{\beta} \max_{\Delta \in \mathcal{U}_{\ell_2}} \left\| \Sigma(\hat{A})^{-1/2} \left( \hat{\theta}^{dir} - X\beta \right) + \Delta \beta \right\|_2
\]

for some \( \mu > 0 \) and \( \mathcal{U}_{\ell_2} = \{ \Delta : \|\Delta \gamma\|_2 \leq \mu \|\gamma\|_2 \text{ for all } \gamma \in \mathbb{R}^p \} \). Clearly, \( \mathcal{U}_{\ell_2} \) is equal to the set \( \{ \Delta : \sigma_{\max}(\Delta) \leq \mu \} \) where \( \sigma_{\max}(\Delta) \) is the maximum singular value of the matrix \( \Delta \). Whereas, the \( \ell_1 \) regularization induced bounds on the individual columns of the perturbation, the \( \ell_2 \) regularization enforces a coherent bound of the whole noise matrix. In addition to the robustness effect, including the \( \ell_2 \)-norm in the minimization problem induces a dense and smooth solution for \( \hat{\beta} \). All elements \( \hat{\beta}_j \in \hat{\beta} \) remain non-zero in the estimation process, whereas their individual contributions to the description of \( \theta_i \) are equalized to some extent, depending on the value of \( \lambda \). If \( \lambda \to \infty \), then \( \hat{\beta}_1 = \ldots = \hat{\beta}_p \). The \( \ell_2 \)-norm regularization has shown to produce stable results in the presence of correlated covariates. However, as \( \hat{\beta}_j \neq 0 \ \forall j = 1, \ldots, p \), no automatic variable selection is conducted.

**Elastic net**

The third regularization is a linear combination of \( \ell_1 \)- and \( \ell_2 \)-norm, which is used for the elastic net introduced by Zou and Hastie (2005). The corresponding weighted minimization problem is given by

\[
\hat{\beta}_{en} = \arg\min_{\beta} \left\| \Sigma(\hat{A})^{-1/2} \left( \hat{\theta}^{dir} - X\beta \right) \right\|_2^2 + \lambda \left[ \alpha \|\beta\|_1 + (1 - \alpha)\|\beta\|_2^2 \right], \tag{19}
\]

where \( \alpha \in [0, 1] \) is a hyperparameter controlling the influence of the \( \ell_1 \)- and \( \ell_2 \)-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 \( g(z) = \|z\|_2 \), \( h_1(z) = \|z\|_1 \), \( h_2(z) = \|z\|_2 \), \( f(z) = z^2 \), \( f_1(z) = z \) and \( f_2(z) = z^2 \) we obtain

\[
\hat{\beta}_{en} = \arg\min_{\beta} \max_{\Delta \in \mathcal{U}_{en}} \left\| \Sigma(\hat{A})^{-1/2} \left( \hat{\theta}^{dir} - X\beta \right) + \Delta \beta \right\|_2
\]

for some \( \mu_1, \mu_2 > 0 \) and \( \mathcal{U}_{en} = \{ \Delta : \|\Delta \gamma\|_2 \leq \mu_1 \|\gamma\|_1 + \mu_2 \|\gamma\|_2 \text{ for all } \gamma \in \mathbb{R}^p \} \). 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 \( \lambda \) must be chosen. Remember that \( \lambda \) has implications regarding the level of noise that can be added to the design matrix \( X \). Following Friedman et al. (2010), we use \( k \)-fold cross validation under minimization of the squared prediction error in order to determine \( \lambda \). Then, assuming the sampling variances \( D_i \) to be known for \( i = 1, \ldots, m \), and letting
$r = 1, \ldots$ 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 $\hat{A}^{init}$ and set $\hat{A}^{init} := \hat{A}^{-1}$
2. **while** not converged **do**
   3. estimate $\tilde{\beta}^r = \tilde{\beta}^r (\hat{A}^{-1})$ conditionally on $\hat{A}^{-1}$ under a given regularization
   4. update the model variance estimate $\hat{A}^r = \hat{A}^r (\tilde{\beta}^r)$ conditionally on $\tilde{\beta}^r$
   5. set $\hat{A}^r := \hat{A}^{-1}$
   6. check convergence
3. **end while**
4. return $\tilde{\beta} := \tilde{\beta}^r$ and $\hat{A} = \hat{A}^r$

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

$$Q(\alpha, \beta, \lambda) = \left\| \Sigma(\hat{A})^{-1/2} \left( \tilde{\theta}^{dir} - X\beta \right) \right\|_2^2 + \lambda \left[ \alpha \|\beta\|_1 + (1 - \alpha)\|\beta\|_2^2 \right]$$

(20)

denote the objective function to be minimized for regression coefficient estimation under elastic net regularization. As $Q(\cdot)$ contains the $\ell_1$-norm and $\ell_2$-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 $\tilde{\beta}_j \in \tilde{\beta}$ 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 $\beta_j \neq 0$, the gradient at $\beta_j = \tilde{\beta}_j$ can be computed according to

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

(21)

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

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

(22)

where $\tilde{\theta}^{dir} - \tilde{\theta}^{dir}_{(j)}$ is the partial residual resulting from regularized weighted partial least
squares excluding the contribution of the $j$-th covariate and the corresponding regression coefficient $\beta_j$. $S(\cdot)$ is the soft-thresholding operator with value

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

(23)

The $\beta_j$ 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

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

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

$$
\hat{A} = \arg\max_A \{ A \cdot \mathcal{L}(A) \}.
$$

(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 $\hat{A} = 0$ 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 $\beta$ by consistent estimates $\hat{A}$ and $\hat{\beta}$ in (11). The MSE of the original Fay-Herriot EBLUP is usually decomposed into three components (Rao and Molina, 2015):

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

(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 $\beta$ 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 \( \ell_2 \)-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 \( \beta, A \), hence \( MSE(\hat{\beta}^{BP}_i|x_i) \). It can be characterized by

\[
MSE \left( \hat{\beta}^{BP}_i \left| x_i \right. \right) = E \left[ (\hat{\beta}^{BP}_i - \theta_i)^2 \right] \\
= E \left[ (\gamma_i \tilde{\theta}_i^{dir} + (1 - \gamma_i)(x_i + \Delta_i)'\beta - x'_i\beta - v_i)^2 \right] \\
= E \left[ (\gamma_i \tilde{\theta}_i^{dir} + x'_i\beta + \Delta'_i\beta - \gamma_i x'_i\beta - \gamma_i \Delta'_i\beta - x'_i\beta - v_i)^2 \right] \\
= E \left[ (\gamma_i \tilde{\theta}_i^{dir} + \Delta'_i\beta - \gamma_i x'_i\beta - \gamma_i \Delta'_i\beta - v_i)^2 \right] \\
= E \left[ (\gamma_i (\tilde{\theta}_i^{dir} - x'_i\beta - v_i) + (1 - \gamma_i) \Delta'_i\beta - (1 - \gamma_i)v_i)^2 \right] \\
= E \left[ (\gamma_i e_i + (1 - \gamma_i)^2(\Delta'_i\beta - v_i)^2 \right].
\]

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 \( \Delta'_i\beta \) is treated as an unknown constant. It follows

\[
MSE \left( \hat{\beta}^{BP}_i \left| x_i \right. \right) = \gamma_i^2 E \left[ e_i^2 \right] + (1 - \gamma_i)^2 E \left[ (\Delta'_i\beta - v_i)^2 \right] \\
= \gamma_i^2 D_i + (1 - \gamma_i)^2 E \left[ (\Delta'_i\beta)^2 - 2 \Delta_i \beta v_i + v_i^2 \right] \\
= \gamma_i^2 D_i + (1 - \gamma_i)^2 A + (1 - \gamma_i)^2(\Delta'_i\beta)^2. \tag{26}
\]

Next, as the term \( \Delta'_i\beta \) cannot be quantified, we substitute it by an upper boundary. Remember that the design matrix perturbations are limited by the regularization parameter. For the \( \ell_1 \)-regularization the noise \( \Delta = (\Delta_1, ..., \Delta_m)' \) is constrained column-wise in terms of the maximum \( \ell_2 \)-norm, whereas for the \( \ell_2 \)-regularization the maximum singular value of \( \Delta \) is restricted. Therefore, we can provide a conservative estimate of the conditional MSE by concluding

\[
MSE(\hat{\beta}^{BP}_i|X) \leq \gamma_i^2 D_i + (1 - \gamma_i)^2 A + (1 - \gamma_i)^2(\lambda'\beta)^2, \tag{27}
\]

where \( \lambda \in \mathbb{R}_+^p \) 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 \( (m)^{-1} \sum_{i=1}^{m} \Delta_i' \beta = 0 \) we can state, that the maximal error \( \delta_i \) in area \( i \) is \( \delta_i = 0.5 \cdot \lambda \). Further, by assuming the noise to be spreaded homogeneously over all areas, one obtains

\[
\hat{MSE}(\hat{\theta}_{BP}^i|X) \leq \gamma_i^2 D_i + (1 - \gamma_i)^2 A + \frac{1}{4m} (1 - \gamma_i)^2 (X' \beta)^2. \tag{28}
\]

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 \( \hat{\varphi} := (\hat{A}, \hat{\beta}) \) and define

\[
\hat{\theta}_{EBP}^i = \hat{\theta}_{EBP}^i(\hat{\varphi}) = \gamma_i \hat{\theta}_{i}^{dir} + (1 - \gamma_i) \hat{x}'_i \hat{\beta} \tag{29}
\]

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

\[
b_i(\hat{\varphi}) = \gamma_i^2 D_i + (1 - \gamma_i)^2 A + \frac{1}{m} (1 - \gamma_i)^2 (X' \beta)^2. \tag{30}
\]

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 \( \hat{A}_{-j} \) and \( \hat{\beta}_{-j} \) from the remaining areas
4: quantify \( \hat{\theta}_{EBP}^i(\hat{\varphi}_{-j}) \) for all \( i = 1, ..., m \) according to (29)
5: quantify \( b_i(\hat{\varphi}_{-j}) \) for all \( i = 1, ..., m \) according to (30)
6: end

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

\[
\hat{MSE}(\hat{\theta}_{EBP}^i) = b_i(\hat{\varphi}) - \frac{m-1}{m} \sum_{j=1}^{m} [b_i(\hat{\varphi}_{-j}) - b_i(\hat{\varphi})] + \frac{m-1}{m} \sum_{j=1}^{m} [\hat{\theta}_{EBP}^i(\hat{\varphi}_{-j}) - \hat{\theta}_{EBP}^i(\hat{\varphi})]^2. \tag{31}
\]
4 Simulation

4.1 Set up

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

$$\theta^r_i = (x^r_i)' \beta + v^r_i, \quad v^r_i \sim N(0, 30^2), \quad \forall \ i = 1, \ldots, m,$$

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

$$\hat{\theta}^\text{dir,}^r_i = \theta^r_i + e^r_i, \quad e^r_i \sim N(0, D^r_i), \quad \forall \ i = 1, \ldots, m,$$

where $D^r_i$ is drawn in each iteration from $\text{unif}(a = 250^2, b = 350^2)$. 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 $\ell_1$-regularized area-level predictor,
- L1: $\ell_1$-regularized area-level predictor,
- L2: $\ell_2$-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 $x^r_i$ in each iteration. The errors are generated for every area individually from an area-specific covariance matrix $\Sigma(\Delta^r_i)$ 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 \( \ell_1 \)-norm included in the weighted minimization problem induces a sparse solution for \( \hat{\beta} \).

The performance of the predictors is evaluated in terms of the relative root mean squared error (RRMSE):

\[
RRMSE(\hat{\theta}^{EBP}_i) = \sqrt{\frac{1}{R} \sum_{r=1}^{R} \left( \hat{\theta}^{EBP,r}_i - \theta^r_i \right)^2} / \sum_{r=1}^{R} \theta^r_i.
\]

We further look at the relative bias

\[
RBias(\hat{\theta}^{EBP}_i) = \frac{1}{R} \sum_{r=1}^{R} \left( \hat{\theta}^{EBP,r}_i - \theta^r_i \right) / \sum_{r=1}^{R} \theta^r_i
\]

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

\[
\overline{MSE}^{MC}(\hat{\theta}^{EBP}_i) = \frac{1}{R} \sum_{r=1}^{R} (\hat{\theta}^{EBP,r}_i - \theta^r_i)^2,
\]

with \( r = 1, \ldots, R \) 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

\[
\text{Coverage rate} = \frac{\sum_{r=1}^{R} \sum_{i=1}^{m} \mathbb{1}_{P_{I_{\theta^r}}(\theta^r_i)}(\theta^r_i)}{R \cdot m} \cdot 100\%,
\]

where the 95% prediction interval is defined by

\[
P_{I_{\theta^r}} = \hat{\theta}^{EBP,r}_i \pm 1.96 \cdot \sqrt{\overline{MSE}(\hat{\theta}^{EBP,r}_i)} \cdot \sqrt{1 + 1/m}.
\]
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.

<table>
<thead>
<tr>
<th>Scen</th>
<th>Direct</th>
<th>FH</th>
<th>YL</th>
<th>CLASSO</th>
<th>L1</th>
<th>L2</th>
<th>EN</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.07724</td>
<td>0.03207</td>
<td>0.03224</td>
<td>0.03193</td>
<td>0.03006</td>
<td>0.03119</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.07724</td>
<td>0.04394</td>
<td>0.04344</td>
<td>0.04312</td>
<td>0.04166</td>
<td>0.04287</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.07724</td>
<td>0.03754</td>
<td>0.03703</td>
<td>0.03690</td>
<td>0.03516</td>
<td>0.03633</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.07724</td>
<td>0.04175</td>
<td>0.04215</td>
<td>0.04031</td>
<td>0.03831</td>
<td>0.03802</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.07724</td>
<td>0.04175</td>
<td>0.04215</td>
<td>0.04031</td>
<td>0.03831</td>
<td>0.03802</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.07724</td>
<td>0.05234</td>
<td>0.05163</td>
<td>0.04921</td>
<td>0.04753</td>
<td>0.04705</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0.07724</td>
<td>0.05429</td>
<td>0.05289</td>
<td>0.05023</td>
<td>0.04867</td>
<td>0.04817</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0.07724</td>
<td>0.04658</td>
<td>0.04655</td>
<td>0.04414</td>
<td>0.04211</td>
<td>0.04208</td>
<td></td>
</tr>
</tbody>
</table>

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

Table 2: Relative bias over all areas

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

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 $\ell_2$-norm. Accordingly, the optimal regularization parameter obtained from cross validation is always the largest for the L2. By squaring the necessary term $X'\beta$, the effect is even stronger, which leads in large MSE estimates.

<table>
<thead>
<tr>
<th>Scen</th>
<th>L1 MC</th>
<th>L1 Est</th>
<th>L2 MC</th>
<th>L2 Est</th>
<th>EN MC</th>
<th>EN Est</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>15259</td>
<td>20253</td>
<td>14027</td>
<td>22121</td>
<td>15154</td>
<td>20196</td>
</tr>
<tr>
<td>2</td>
<td>28428</td>
<td>30749</td>
<td>26851</td>
<td>34733</td>
<td>28216</td>
<td>30658</td>
</tr>
<tr>
<td>3</td>
<td>31732</td>
<td>36890</td>
<td>30105</td>
<td>40273</td>
<td>31421</td>
<td>36756</td>
</tr>
<tr>
<td>4</td>
<td>21252</td>
<td>24672</td>
<td>19811</td>
<td>27698</td>
<td>21086</td>
<td>24609</td>
</tr>
<tr>
<td>5</td>
<td>22610</td>
<td>30823</td>
<td>22861</td>
<td>40441</td>
<td>22242</td>
<td>31647</td>
</tr>
<tr>
<td>6</td>
<td>34064</td>
<td>41553</td>
<td>33141</td>
<td>51387</td>
<td>33363</td>
<td>42120</td>
</tr>
<tr>
<td>7</td>
<td>37680</td>
<td>49211</td>
<td>36691</td>
<td>57459</td>
<td>36875</td>
<td>49505</td>
</tr>
<tr>
<td>8</td>
<td>27476</td>
<td>34956</td>
<td>27211</td>
<td>45119</td>
<td>26974</td>
<td>35461</td>
</tr>
</tbody>
</table>

**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.

<table>
<thead>
<tr>
<th>Scen</th>
<th>L1 MC</th>
<th>L1 Est</th>
<th>L2 MC</th>
<th>L2 Est</th>
<th>EN MC</th>
<th>EN Est</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8433</td>
<td>10761</td>
<td>7982</td>
<td>10806</td>
<td>8403</td>
<td>10738</td>
</tr>
<tr>
<td>2</td>
<td>24640</td>
<td>25692</td>
<td>24043</td>
<td>26207</td>
<td>24599</td>
<td>25659</td>
</tr>
<tr>
<td>3</td>
<td>19741</td>
<td>21324</td>
<td>19183</td>
<td>21753</td>
<td>19686</td>
<td>21279</td>
</tr>
<tr>
<td>4</td>
<td>12570</td>
<td>13607</td>
<td>12077</td>
<td>13758</td>
<td>12534</td>
<td>13573</td>
</tr>
<tr>
<td>5</td>
<td>13048</td>
<td>16118</td>
<td>14615</td>
<td>19726</td>
<td>13177</td>
<td>16266</td>
</tr>
<tr>
<td>6</td>
<td>27572</td>
<td>30591</td>
<td>27931</td>
<td>34215</td>
<td>27457</td>
<td>30683</td>
</tr>
<tr>
<td>7</td>
<td>23851</td>
<td>26988</td>
<td>24662</td>
<td>31623</td>
<td>23829</td>
<td>28948</td>
</tr>
<tr>
<td>8</td>
<td>16958</td>
<td>20043</td>
<td>18263</td>
<td>23206</td>
<td>17014</td>
<td>20156</td>
</tr>
</tbody>
</table>

**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.

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

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 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 $\ell_1$-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 $\lambda$ is obtained from $k$-fold cross validation. Applying the robustified area-level modelling in the described manner yields the following results.

<table>
<thead>
<tr>
<th>State</th>
<th>Model</th>
<th>Former FH</th>
<th>Former Var</th>
<th>Model MSE</th>
<th>FH MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alabama</td>
<td>710</td>
<td>784</td>
<td>704</td>
<td>3 600</td>
<td>356</td>
</tr>
<tr>
<td>Alaska</td>
<td>62</td>
<td>65</td>
<td>65</td>
<td>36</td>
<td>38</td>
</tr>
<tr>
<td>Arizona</td>
<td>1 076</td>
<td>1 156</td>
<td>1 110</td>
<td>7 056</td>
<td>592</td>
</tr>
<tr>
<td>Arkansas</td>
<td>487</td>
<td>475</td>
<td>494</td>
<td>729</td>
<td>196</td>
</tr>
<tr>
<td>California</td>
<td>5 439</td>
<td>5 441</td>
<td>5 396</td>
<td>44 521</td>
<td>5 886</td>
</tr>
<tr>
<td>Colorado</td>
<td>570</td>
<td>538</td>
<td>525</td>
<td>4 761</td>
<td>363</td>
</tr>
<tr>
<td>Connecticut</td>
<td>379</td>
<td>324</td>
<td>341</td>
<td>1 849</td>
<td>173</td>
</tr>
<tr>
<td>Delaware</td>
<td>107</td>
<td>106</td>
<td>103</td>
<td>144</td>
<td>45</td>
</tr>
<tr>
<td>District Columbia</td>
<td>120</td>
<td>113</td>
<td>112</td>
<td>64</td>
<td>128</td>
</tr>
<tr>
<td>Florida</td>
<td>3 017</td>
<td>3 253</td>
<td>3 117</td>
<td>30 976</td>
<td>2 972</td>
</tr>
<tr>
<td>Georgia</td>
<td>1 805</td>
<td>1 833</td>
<td>1 888</td>
<td>12 321</td>
<td>1 781</td>
</tr>
<tr>
<td>Hawaii</td>
<td>159</td>
<td>151</td>
<td>147</td>
<td>256</td>
<td>115</td>
</tr>
<tr>
<td>Idaho</td>
<td>197</td>
<td>204</td>
<td>187</td>
<td>441</td>
<td>83</td>
</tr>
<tr>
<td>Illinois</td>
<td>1 491</td>
<td>1 380</td>
<td>1 498</td>
<td>10 816</td>
<td>1 551</td>
</tr>
<tr>
<td>Indiana</td>
<td>780</td>
<td>880</td>
<td>784</td>
<td>5 184</td>
<td>290</td>
</tr>
<tr>
<td>Iowa</td>
<td>333</td>
<td>321</td>
<td>301</td>
<td>1 521</td>
<td>151</td>
</tr>
<tr>
<td>Kansas</td>
<td>333</td>
<td>404</td>
<td>329</td>
<td>1 764</td>
<td>123</td>
</tr>
<tr>
<td>Kentucky</td>
<td>760</td>
<td>851</td>
<td>758</td>
<td>3 249</td>
<td>651</td>
</tr>
<tr>
<td>Louisiana</td>
<td>871</td>
<td>854</td>
<td>869</td>
<td>3 249</td>
<td>429</td>
</tr>
<tr>
<td>Maine</td>
<td>138</td>
<td>165</td>
<td>170</td>
<td>400</td>
<td>67</td>
</tr>
<tr>
<td>Maryland</td>
<td>568</td>
<td>566</td>
<td>552</td>
<td>4 900</td>
<td>1 112</td>
</tr>
<tr>
<td>Massachusetts</td>
<td>808</td>
<td>782</td>
<td>757</td>
<td>4 225</td>
<td>617</td>
</tr>
<tr>
<td>Michigan</td>
<td>1 233</td>
<td>1 259</td>
<td>1 160</td>
<td>9 604</td>
<td>859</td>
</tr>
<tr>
<td>Minnesota</td>
<td>498</td>
<td>428</td>
<td>492</td>
<td>3 364</td>
<td>356</td>
</tr>
<tr>
<td>Mississippi</td>
<td>538</td>
<td>563</td>
<td>548</td>
<td>1 089</td>
<td>277</td>
</tr>
<tr>
<td>Missouri</td>
<td>686</td>
<td>582</td>
<td>622</td>
<td>5 041</td>
<td>525</td>
</tr>
<tr>
<td>Montana</td>
<td>117</td>
<td>121</td>
<td>121</td>
<td>144</td>
<td>63</td>
</tr>
<tr>
<td>Nebraska</td>
<td>204</td>
<td>192</td>
<td>206</td>
<td>324</td>
<td>76</td>
</tr>
<tr>
<td>Nevada</td>
<td>382</td>
<td>371</td>
<td>398</td>
<td>1 296</td>
<td>441</td>
</tr>
<tr>
<td>New Hampshire</td>
<td>86</td>
<td>94</td>
<td>94</td>
<td>169</td>
<td>56</td>
</tr>
<tr>
<td>New Jersey</td>
<td>929</td>
<td>1 004</td>
<td>896</td>
<td>9 025</td>
<td>968</td>
</tr>
<tr>
<td>New Mexico</td>
<td>426</td>
<td>400</td>
<td>425</td>
<td>784</td>
<td>222</td>
</tr>
<tr>
<td>New York</td>
<td>2 609</td>
<td>2 791</td>
<td>2 747</td>
<td>24 336</td>
<td>14 334</td>
</tr>
<tr>
<td>North Carolina</td>
<td>1 374</td>
<td>1 509</td>
<td>1 452</td>
<td>9 025</td>
<td>2 087</td>
</tr>
<tr>
<td>North Dakota</td>
<td>83</td>
<td>82</td>
<td>82</td>
<td>144</td>
<td>49</td>
</tr>
<tr>
<td>Ohio</td>
<td>1 541</td>
<td>1 550</td>
<td>1 690</td>
<td>11 025</td>
<td>721</td>
</tr>
<tr>
<td>Oklahoma</td>
<td>556</td>
<td>551</td>
<td>526</td>
<td>2 025</td>
<td>309</td>
</tr>
<tr>
<td>Oregon</td>
<td>506</td>
<td>478</td>
<td>503</td>
<td>4 761</td>
<td>231</td>
</tr>
<tr>
<td>Pennsylvania</td>
<td>1 483</td>
<td>1 542</td>
<td>1 470</td>
<td>11 449</td>
<td>2 226</td>
</tr>
</tbody>
</table>
Table 6: Number of people below 100% of the federal poverty threshold in thousands

<table>
<thead>
<tr>
<th>State</th>
<th>Model</th>
<th>Former</th>
<th>FH</th>
<th>Former Var</th>
<th>Model MSE</th>
<th>FH MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rhode Island</td>
<td>122</td>
<td>123</td>
<td>122</td>
<td>256</td>
<td>70</td>
<td>110</td>
</tr>
<tr>
<td>South Carolina</td>
<td>834</td>
<td>683</td>
<td>800</td>
<td>3136</td>
<td>536</td>
<td>685</td>
</tr>
<tr>
<td>South Dakota</td>
<td>117</td>
<td>118</td>
<td>115</td>
<td>100</td>
<td>43</td>
<td>92</td>
</tr>
<tr>
<td>Tennessee</td>
<td>982</td>
<td>973</td>
<td>1008</td>
<td>4900</td>
<td>438</td>
<td>407</td>
</tr>
<tr>
<td>Texas</td>
<td>4075</td>
<td>4036</td>
<td>4261</td>
<td>42849</td>
<td>3842</td>
<td>5623</td>
</tr>
<tr>
<td>Utah</td>
<td>239</td>
<td>279</td>
<td>245</td>
<td>900</td>
<td>245</td>
<td>339</td>
</tr>
<tr>
<td>Vermont</td>
<td>60</td>
<td>65</td>
<td>62</td>
<td>64</td>
<td>37</td>
<td>73</td>
</tr>
<tr>
<td>Virginia</td>
<td>1040</td>
<td>894</td>
<td>923</td>
<td>6561</td>
<td>2284</td>
<td>1718</td>
</tr>
<tr>
<td>Washington</td>
<td>855</td>
<td>819</td>
<td>824</td>
<td>2916</td>
<td>1457</td>
<td>1402</td>
</tr>
<tr>
<td>West Virginia</td>
<td>258</td>
<td>261</td>
<td>234</td>
<td>2116</td>
<td>141</td>
<td>235</td>
</tr>
<tr>
<td>Wisconsin</td>
<td>714</td>
<td>654</td>
<td>682</td>
<td>4624</td>
<td>439</td>
<td>339</td>
</tr>
<tr>
<td>Wyoming</td>
<td>56</td>
<td>56</td>
<td>65</td>
<td>64</td>
<td>31</td>
<td>69</td>
</tr>
</tbody>
</table>

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 \textit{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 $\ell_2$-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


**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.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median household income</td>
<td>US Census Bureau</td>
</tr>
<tr>
<td>Percentage of households with income below 10.000$</td>
<td>US Census Bureau</td>
</tr>
<tr>
<td>Percentage of households on the Supplemental</td>
<td>US Census Bureau</td>
</tr>
<tr>
<td>Nutrition Assistance Program</td>
<td>US Census Bureau</td>
</tr>
<tr>
<td>Unemployment rate</td>
<td>US Census Bureau</td>
</tr>
<tr>
<td>Number of people below 50% of the federal poverty threshold</td>
<td>US Census Bureau</td>
</tr>
<tr>
<td>Number of people below 100% of the federal poverty threshold</td>
<td>US Census Bureau</td>
</tr>
<tr>
<td>Number of violent crimes per 100.000</td>
<td>UCR Program</td>
</tr>
<tr>
<td>Number of murders and nonnegligent manslaughters per 100.000</td>
<td>UCR Program</td>
</tr>
<tr>
<td>Number of rapes per 100.000, legacy definition</td>
<td>UCR Program</td>
</tr>
<tr>
<td>Number of rapes per 100.000, revised definition</td>
<td>UCR Program</td>
</tr>
<tr>
<td>Number of robberies per 100.000</td>
<td>UCR Program</td>
</tr>
<tr>
<td>Number of aggravated assaults per 100.000</td>
<td>UCR Program</td>
</tr>
<tr>
<td>Number of property crimes per 100.000</td>
<td>UCR Program</td>
</tr>
<tr>
<td>Number of burglaries per 100.000</td>
<td>UCR Program</td>
</tr>
<tr>
<td>Number of larceny-thefts per 100.000</td>
<td>UCR Program</td>
</tr>
<tr>
<td>Number of motor vehicle thefts per 100.000</td>
<td>UCR Program</td>
</tr>
</tbody>
</table>
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 $\Sigma_{\Delta_i}$. Note that these measurement error covariance matrices differ per scenario. They are generated according to the following procedure.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Generation</th>
<th>Areas</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$\Sigma(\Delta_i^r) = 0_{7,7}$</td>
<td>$i = 1, \ldots, m$</td>
</tr>
<tr>
<td>2</td>
<td>$\Sigma(\Delta_i^r) \in \mathbb{R}^{7 \times 7}<em>{+}$ positive-definite, $\sigma</em>{j,l}^r \sim \text{unif}(a = 40, b = 60)$</td>
<td>$i = 1, \ldots, m$</td>
</tr>
<tr>
<td>3</td>
<td>$\Sigma(\Delta_i^r) = \text{diag}(\sigma_1^r, \ldots, \sigma_7^r)$, $\sigma_j^r \sim \text{unif}(a = 40, b = 60)$</td>
<td>$i = 1, \ldots, (m/2)$</td>
</tr>
<tr>
<td>3</td>
<td>$\Sigma(\Delta_i^r) = 0_{7,7}$</td>
<td>$i = (m/2 + 1), \ldots, m$</td>
</tr>
<tr>
<td>4</td>
<td>$\Sigma(\Delta_i^r) \in \mathbb{R}^{7 \times 7}<em>{+}$ positive-definite, $\sigma</em>{j,l}^r \sim \text{unif}(a = 40, b = 60)$</td>
<td>$i = 1, \ldots, (m/2)$</td>
</tr>
<tr>
<td>4</td>
<td>$\Sigma(\Delta_i^r) = 0_{7,7}$</td>
<td>$i = (m/2 + 1), \ldots, m$</td>
</tr>
<tr>
<td>5</td>
<td>$\Sigma(\Delta_i^r) = 0_{15,15}$</td>
<td>$i = 1, \ldots, m$</td>
</tr>
<tr>
<td>6</td>
<td>$\Sigma(\Delta_i^r) \in \mathbb{R}^{15 \times 15}<em>{+}$ positive-definite, $\sigma</em>{j,l}^r \sim \text{unif}(a = 40, b = 60)$</td>
<td>$i = 1, \ldots, m$</td>
</tr>
<tr>
<td>7</td>
<td>$\Sigma(\Delta_i^r) = \text{diag}(\sigma_1^r, \ldots, \sigma_{15}^r)$, $\sigma_j^r \sim \text{unif}(a = 40, b = 60)$</td>
<td>$i = 1, \ldots, (m/2)$</td>
</tr>
<tr>
<td>7</td>
<td>$\Sigma(\Delta_i^r) = 0_{15,15}$</td>
<td>$i = (m/2 + 1), \ldots, m$</td>
</tr>
<tr>
<td>8</td>
<td>$\Sigma(\Delta_i^r) \in \mathbb{R}^{15 \times 15}<em>{+}$ positive-definite, $\sigma</em>{j,l}^r \sim \text{unif}(a = 40, b = 60)$</td>
<td>$i = 1, \ldots, (m/2)$</td>
</tr>
<tr>
<td>8</td>
<td>$\Sigma(\Delta_i^r) = 0_{15,15}$</td>
<td>$i = (m/2 + 1), \ldots, m$</td>
</tr>
</tbody>
</table>