Human Guided Linear Regression with Feature-Level Constraints

Aubrey Gress and Ian Davidson

Department of Computer Science, University of California, Davis

AAAI-2018

Outline

- Background
- Introduction
- Methods
- Experiments
- Conclusion

Background

Linear Regression

- simplicity
- interpretability
- reduced likelihood of overfitting
- perform poorly if little labeled training data is available

active learning、semi-supervised learning、Transfer learning

large amounts of unlabeled data，“cluster assumption” to hold or additional sources of sufficiently related data.

Contents

- Background
- Introduction
- Methods
- Experiments
- Conclusion

Introduction

non-data centric approach-using knowledge the user may have about the features

Users may have a wealth of understanding about the relationships between the features and the output

Example，life expectancy- smoking habits

Three forms of feature level knowledge

- **parameter signs**
- **relative parameter effect ordering**
- **Pairwise parameter signs**

contributions

- Parameter Sign Constrained Regression (PSCR)
- Parameter Relative Constrained Regression (PRCR)
- Pairwise Parameter Sign Constrained Regression (PPSCR)

Contents

- Background
- Introduction
- Methods
- Experiments
- Conclusion

Parameter Sign Constrained Regression

$$\begin{aligned}
&\min_{\beta,\xi} \frac{1}{n}||X\beta - y||^2 + C||\beta||^2 \\
&s.t. \ (1-\xi_i)e_i\beta_i \geq 0, \forall e_i \in e \\
&\quad \xi \in \{0,1\}^p \\
&\quad \sum_i \xi_i \leq \lambda
\end{aligned}$$

X is a n×p matrix of n instances and p features and Y is a n×1 vector of responses.

e, a p×1 vector where ei is 1 if the sign of the coefficient should be nonnegative,−1 if nonpositive and 0 if no guidance is provided for the coefficient.

ξ_i, a set of discrete variables which, when set to 1,deactivate the corresponding sign constraint.

C and λ are regularization parameters.

Parameter Sign Constrained Regression

difficult discrete optimization problem which can be hard to solve in practice. Thus, we relax it to the following continuous, convex optimization problem:

$$\begin{aligned}
&\min_{\beta,\xi} \frac{1}{n}||X\beta - y||^2 + C||w||^2 \\
&s.t.\ E\beta + \xi \geq 0 \\
&\qquad \xi \geq 0 \\
&\qquad \sum_i \xi_i \leq \lambda
\end{aligned}$$

E is a diagonal matrix encoding the sign constraints, with Eii ∈{−1,0,1}depending on the sign guidance.

The $\xi_i$are now, in a manner analogous to the Support Vector Machine (SVM), slack variables which control the extent to which the sign constraints can be violated .

Parameter Relative Constrained Regression

$$
\begin{aligned}
\min_{\beta,\xi} \ & \frac{1}{n}||X\beta - y||^2 + C||\beta||^2 \\
s.t. \ & \beta_j - \beta_k + \xi_i \geq 0, \forall p_i \in P \\
& \xi \geq 0 \\
& \sum_i \xi_i \leq \lambda
\end{aligned}
$$

P be a set of pairs where each pi ∈ P is a tuple (j,k) indicating that coefficient j is greater than coefficient k.

Pairwise Parameter Sign Constrained Regression

$$\begin{aligned}
\min_{\beta,\xi} \; & \frac{1}{n}||X\beta - y||^2 + C||\beta||^2 \\
& + \lambda_2 \sum_{p_i \in P} \max(-(E_{j,k}\beta_j\beta_k + \xi_i), 0) \\
s.t. \; & \xi \geq 0 \\
& \sum_i \xi_i \leq \lambda_1
\end{aligned}$$

P - a set of pairs where each pi ∈ P is a tuple (j,k) indicating that coefficients j and k should have the same (or opposite) sign.

$E_{j,k}$is positive if they should have the same sign, and negative otherwise.

λ1 and λ2 are regularization parameters.

different from previous formulation, this leads to nonconvex quadratic constraints which most standard continuous optimization libraries cannot model. We have moved the guidance constraint to the objective.

User Generated Guidance

Ideally, parameter constraint guidance could be estimated from the training data, but this guidance will be noisy if little training data is available.

The user is a potentially better source, but it is not clear users will be able to accurately provide it.

They can likely answer these same questions when considering a small subset of the features. For example， life expectancy- smoking habits

Parameter Constraint Transfer

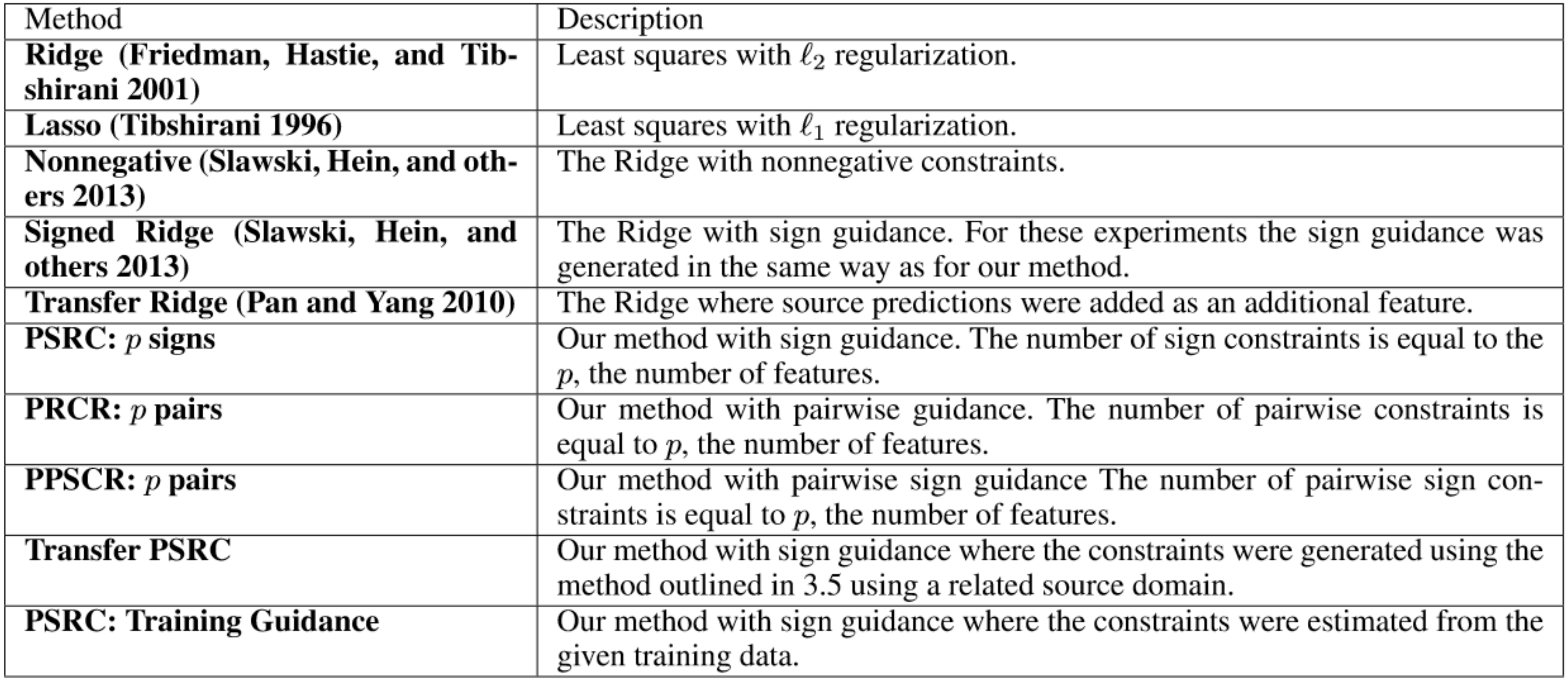
While our proposed guidance can be provided by domain experts, for domains where this is not possible, it can instead be learned through transfer learning.

the constraints are generated by estimating them from a related source domain which shares the same feature set.

Contents

- Background
- Introduction
- Methods
- Experiments
- Conclusion

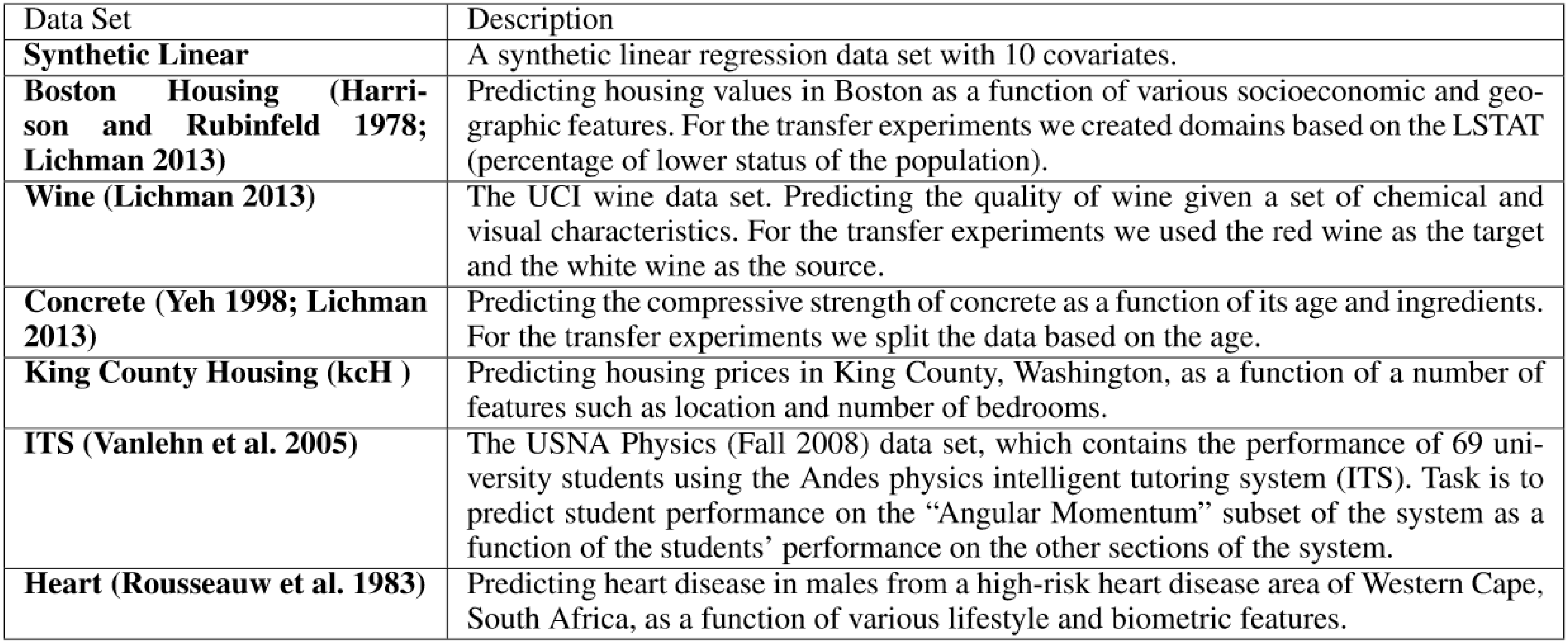
Experiments



Method	Description
Ridge (Friedman, Hastie, and Tibshirani 2001)	Least squares with ℓ_2 regularization.
Lasso (Tibshirani 1996)	Least squares with ℓ_1 regularization.
Nonnegative (Slawski, Hein, and others 2013)	The Ridge with nonnegative constraints.
Signed Ridge (Slawski, Hein, and others 2013)	The Ridge with sign guidance. For these experiments the sign guidance was generated in the same way as for our method.
Transfer Ridge (Pan and Yang 2010)	The Ridge where source predictions were added as an additional feature.
PSRC: p signs	Our method with sign guidance. The number of sign constraints is equal to the p, the number of features.
PRCR: p pairs	Our method with pairwise guidance. The number of pairwise constraints is equal to p, the number of features.
PPSCR: p pairs	Our method with pairwise sign guidance The number of pairwise sign constraints is equal to p, the number of features.
Transfer PSRC	Our method with sign guidance where the constraints were generated using the method outlined in 3.5 using a related source domain.
PSRC: Training Guidance	Our method with sign guidance where the constraints were estimated from the given training data.

Table 1: Methods we used in our experiments.

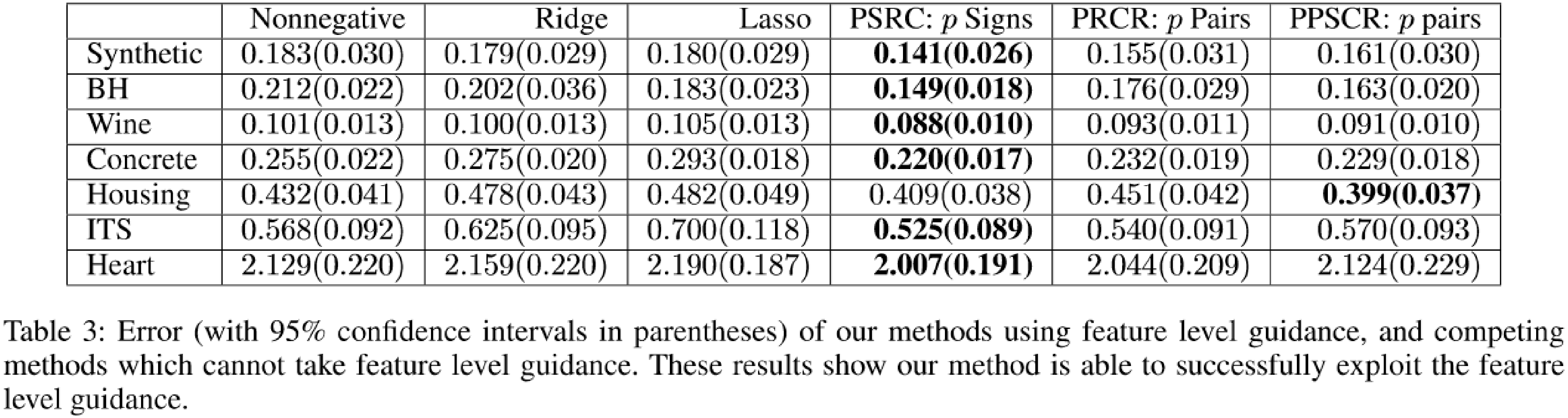
Experiments



Data Set	Description
Synthetic Linear	A synthetic linear regression data set with 10 covariates.
Boston Housing (Harrison and Rubinfeld 1978; Lichman 2013)	Predicting housing values in Boston as a function of various socioeconomic and geographic features. For the transfer experiments we created domains based on the LSTAT (percentage of lower status of the population).
Wine (Lichman 2013)	The UCI wine data set. Predicting the quality of wine given a set of chemical and visual characteristics. For the transfer experiments we used the red wine as the target and the white wine as the source.
Concrete (Yeh 1998; Lichman 2013)	Predicting the compressive strength of concrete as a function of its age and ingredients. For the transfer experiments we split the data based on the age.
King County Housing (kcH)	Predicting housing prices in King County, Washington, as a function of a number of features such as location and number of bedrooms.
ITS (Vanlehn et al. 2005)	The USNA Physics (Fall 2008) data set, which contains the performance of 69 university students using the Andes physics intelligent tutoring system (ITS). Task is to predict student performance on the "Angular Momentum" subset of the system as a function of the students' performance on the other sections of the system.
Heart (Rousseauw et al. 1983)	Predicting heart disease in males from a high-risk heart disease area of Western Cape, South Africa, as a function of various lifestyle and biometric features.

Table 2: Data sets we used in our experiments.

Experiments

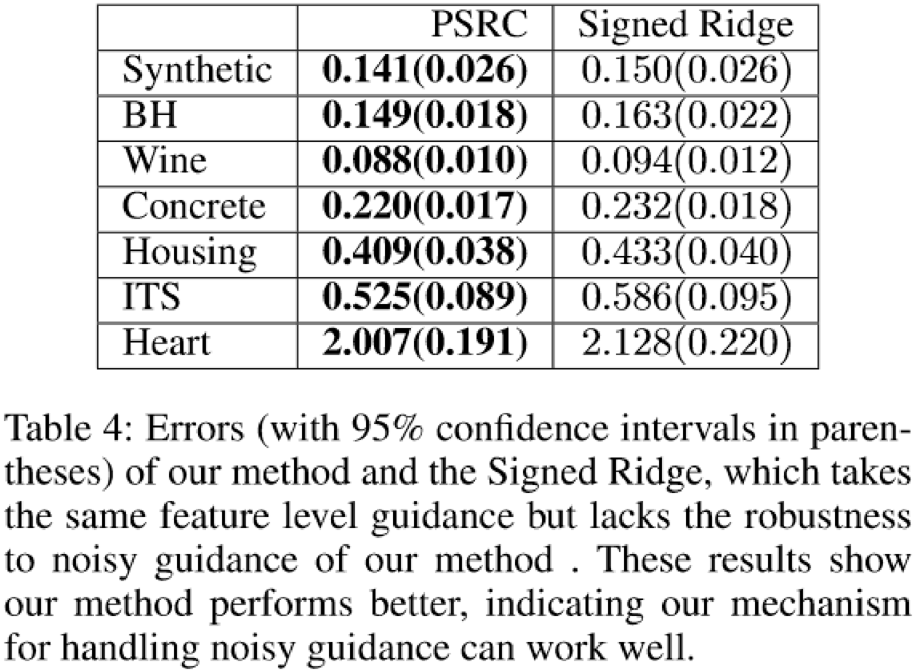


	Nonnegative	Ridge	Lasso	PSRC: p Signs	PRCR: p Pairs	PPSCR: p pairs
Synthetic	0.183(0.030)	0.179(0.029)	0.180(0.029)	**0.141(0.026)**	0.155(0.031)	0.161(0.030)
BH	0.212(0.022)	0.202(0.036)	0.183(0.023)	**0.149(0.018)**	0.176(0.029)	0.163(0.020)
Wine	0.101(0.013)	0.100(0.013)	0.105(0.013)	**0.088(0.010)**	0.093(0.011)	0.091(0.010)
Concrete	0.255(0.022)	0.275(0.020)	0.293(0.018)	**0.220(0.017)**	0.232(0.019)	0.229(0.018)
Housing	0.432(0.041)	0.478(0.043)	0.482(0.049)	0.409(0.038)	0.451(0.042)	**0.399(0.037)**
ITS	0.568(0.092)	0.625(0.095)	0.700(0.118)	**0.525(0.089)**	0.540(0.091)	0.570(0.093)
Heart	2.129(0.220)	2.159(0.220)	2.190(0.187)	**2.007(0.191)**	2.044(0.209)	2.124(0.229)

Table 3: Error (with 95% confidence intervals in parentheses) of our methods using feature level guidance, and competing methods which cannot take feature level guidance. These results show our method is able to successfully exploit the feature level guidance.

In particular, our methods performed much better than the Ridge with nonnegative constraints, showing that providing more accurate sign guidance is better than arbitrarily constraining the signs to be nonnegative.

Experiments

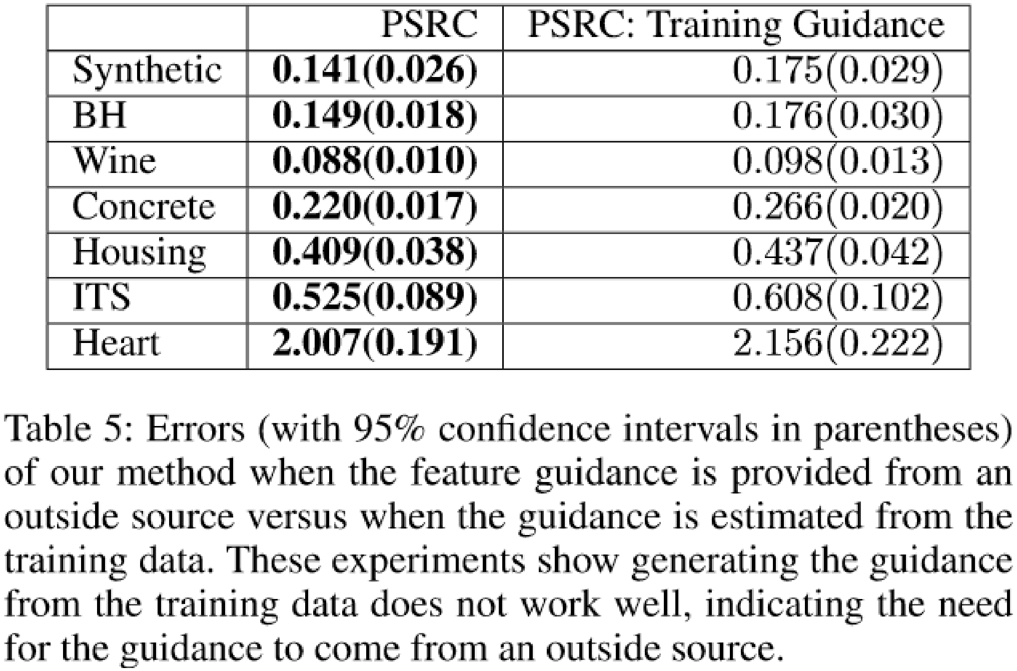


	PSRC	Signed Ridge
Synthetic	**0.141(0.026)**	0.150(0.026)
BH	**0.149(0.018)**	0.163(0.022)
Wine	**0.088(0.010)**	0.094(0.012)
Concrete	**0.220(0.017)**	0.232(0.018)
Housing	**0.409(0.038)**	0.433(0.040)
ITS	**0.525(0.089)**	0.586(0.095)
Heart	**2.007(0.191)**	2.128(0.220)

Table 4: Errors (with 95% confidence intervals in parentheses) of our method and the Signed Ridge, which takes the same feature level guidance but lacks the robustness to noisy guidance of our method . These results show our method performs better, indicating our mechanism for handling noisy guidance can work well.

Compared to “Signed Ridge,” which does not have a mechanism for relaxing the sign guidance, our method performs much better. This shows our relaxation can play an important role in preventing overfitting.

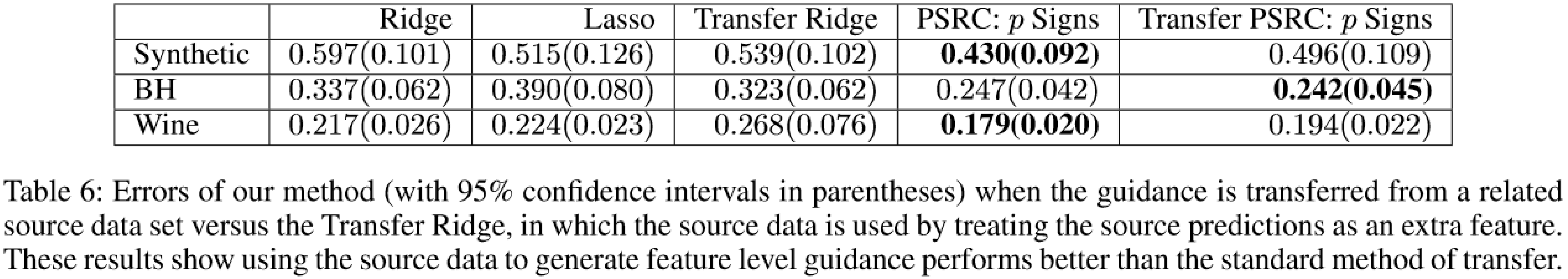
Experiments



	PSRC	PSRC: Training Guidance
Synthetic	**0.141(0.026)**	0.175(0.029)
BH	**0.149(0.018)**	0.176(0.030)
Wine	**0.088(0.010)**	0.098(0.013)
Concrete	**0.220(0.017)**	0.266(0.020)
Housing	**0.409(0.038)**	0.437(0.042)
ITS	**0.525(0.089)**	0.608(0.102)
Heart	**2.007(0.191)**	2.156(0.222)

Table 5: Errors (with 95% confidence intervals in parentheses) of our method when the feature guidance is provided from an outside source versus when the guidance is estimated from the training data. These experiments show generating the guidance from the training data does not work well, indicating the need for the guidance to come from an outside source.

Experiments



	Ridge	Lasso	Transfer Ridge	PSRC: p Signs	Transfer PSRC: p Signs
Synthetic	0.597(0.101)	0.515(0.126)	0.539(0.102)	**0.430(0.092)**	0.496(0.109)
BH	0.337(0.062)	0.390(0.080)	0.323(0.062)	0.247(0.042)	**0.242(0.045)**
Wine	0.217(0.026)	0.224(0.023)	0.268(0.076)	**0.179(0.020)**	0.194(0.022)

Table 6: Errors of our method (with 95% confidence intervals in parentheses) when the guidance is transferred from a related source data set versus the Transfer Ridge, in which the source data is used by treating the source predictions as an extra feature. These results show using the source data to generate feature level guidance performs better than the standard method of transfer.

Our results show our method with transfer performs better than the Transfer Ridge, but worse than our method where the sign guidance was generated through a simulated human.

Contents

- Background
- Introduction
- Approach
- Experiments
- Conclusion

Conclusion

- We proposed novel ways of constraining parameter along with formulations that are more robust to overfitting by allowing the guidance to be relaxed.
- We also presented two practical methods to provide this guidance: through simpler pointwise and pairwise queries and through transfer learning.