Dataset description

The eICU (eicu), MIMIC-III (mimic), MIMIC-IV (miiv), and HiRID (hirid) databases are open-access collections of de-identified patient data from multiple hospitals in the US (eICU), a single hospital in Boston, US (MIMIC-III, MIMIC-IV), and a single hospital in Bern, Switzerland (HiRID). They include 4 static and 48 dynamic variables. The static variables, sex, age, weight, and height, are measured once at entry to the ICU. The dynamic variables are measured in 1h intervals and include vital signs (e.g., heart rate, respiratory rate, and blood pressure) laboratory results (e.g., blood gases, electrolytes, and hematology), medications administered (e.g., antibiotics), and procedures performed in the ICU (e.g., mechanical ventilation).

We predict the patients' average heart rate 48-72h after entry to the ICU from the static variables and the average value of the dynamic variables 0-24h after entry to the ICU. As preprocessing, we log-transform certain variables, mean-impute missing values, and add a missingness indicator. The categorical feature sex is one-hot encoded. We remove patient visits if there was no measurement of heart rate in the time windows 0-24h or 48-72h. Then, from eicu, we remove observations corresponding to hospitals with fewer than 10 observations.

Afterward, the eICU dataset contains 74587 patient visits from 188 different hospitals, the MIMIC-III dataset contains 30335 patient visits, the MIMIC-IV dataset contains 35673 patient visits, and the HiRID dataset contains 8577 patient visits.

Task description

We fine-tune a model from the source distribution (eicu) to a target distribution (mimic, miiv, hirid). For this, we are given an 80% train split of the source dataset (eicu) and a small fine-tuning dataset with `n_test` (25, 50, 100, 200, 400, 800, and 1600) observations from the target distribution (mimic, miiv, hirid).

Models

Each model consists of a *prior*, that is fitted on the source distribution training dataset (eicu), and a *posterior*, that improves on the prior using the `n_test` samples from the fine-tuning dataset. Both the prior and the posterior can have hyperparameters. These are optimized by 5-fold cross-validation on the `n_test` samples from the fine-tuning dataset.

elastic net on target

This model has no (or a trivial) prior. The "posterior" model is an elastic net, which is fit to the fine-tuning dataset. In particular, this model does not the source distribution training dataset.

```
prior_params: {},
model_params: {
  "alpha": [0.001, 0.00316, 0.01, 0.0316, 0.1, 0.316, 1, 3.16, 10, 31.6, 100]
  "l1_ratio": [0, 0.2, 0.5, 0.8, 1]
}
```

anchor + emp bayes (id)

The prior of this model is (elastic net regularized) anchor regression. The posterior is empirical Bayes, minimizing

$$\hat{\beta}(\alpha_{\text{posterior}}, \gamma, \alpha_{\text{prior}}, \lambda) = \arg\min_{\beta} \frac{1}{n} \|y_{\text{target}} - X_{\text{target}}\beta\|^2 + \alpha \|\beta - \hat{\beta}_{\text{prior}}(\gamma, \alpha_{\text{prior}}, \lambda)\|^2,$$

where

$$\hat{\beta}_{\text{prior}}(\gamma, \alpha_{\text{prior}}, \lambda) = \arg\min_{\beta} \frac{1}{n} \|y_{\text{source}} - X_{\text{source}}\beta\|^2 + (\gamma - 1)\frac{1}{n} \|P_{A_{\text{source}}}(y_{\text{source}} - X_{\text{source}}\beta)\|^2 + \alpha_{\text{prior}}\lambda\|\beta\|_1 + \alpha_{\text{prior}}(1-\lambda)\|\beta\|_2^2.$$

Below, `l1_ratio` $= \lambda$.

```
prior_params: {
  "alpha": [0.00001, 0.0001, 0.001, 0.01, 0.1],
  "l1_ratio": [0, 0.2, 0.5, 0.8, 1],
  "gamma": [1, 3.16, 10, 31.6, 100, 316, 1000, 3162, 10000]
}
model_params: {
    {"alpha": [Infinity, 1.0, 2.61, 6.81, 17.7, 46.4, 121,
                  316, 825, 2154, 5623, 14677, 38311, 100000]}
}
```

anchor regression

The prior of this model is (regularized) anchor regression (as for anchor + emp. Bayes above). However, there is no (or, a trivial) posterior. The fine-tuning dataset is just used for model selection.

```
prior_params: {
  "alpha": [0.00001, 0.0001, 0.001, 0.01, 0.1],
  "l1_ratio": [0, 0.2, 0.5, 0.8, 1],
  "gamma": [1, 3.16, 10, 31.6, 100, 316, 1000, 3162, 10000]
}
model_params: {}
```

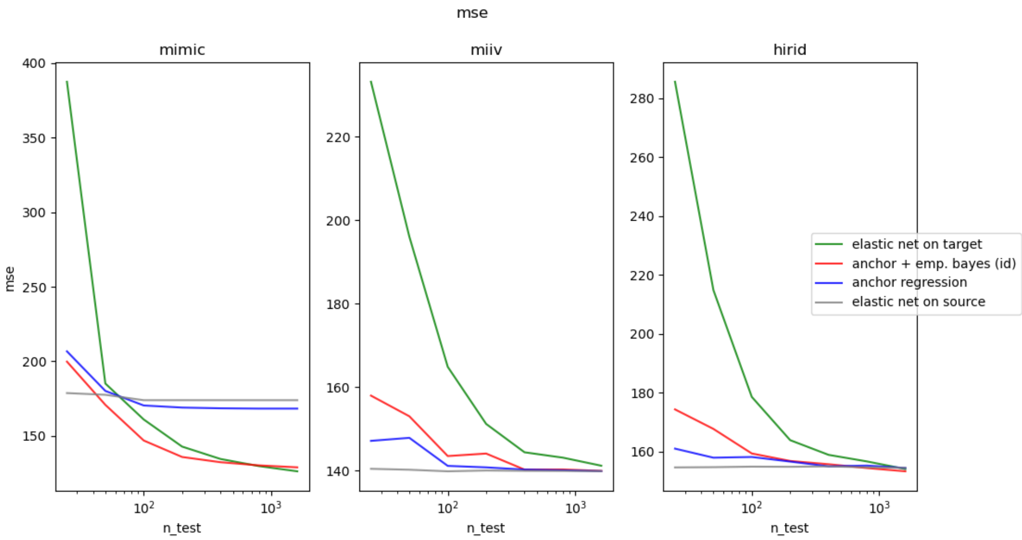
elastic net on source

The prior of this is elastic net (same as anchor regression, but with $\gamma = 1$ fixed). There is no (or, a trivial) posterior. The fine-tuning dataset is just used for model selection.

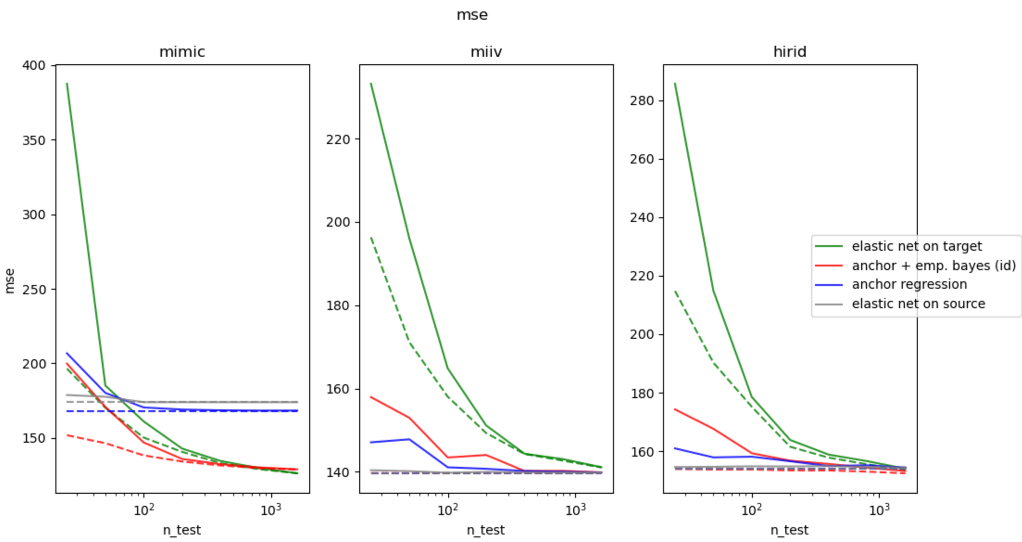
```
prior_params: {
  "alpha": [0.00001, 0.0001, 0.001, 0.01, 0.1],
  "l1_ratio": [0, 0.2, 0.5, 0.8, 1],
}
model_params: {}
```

Results

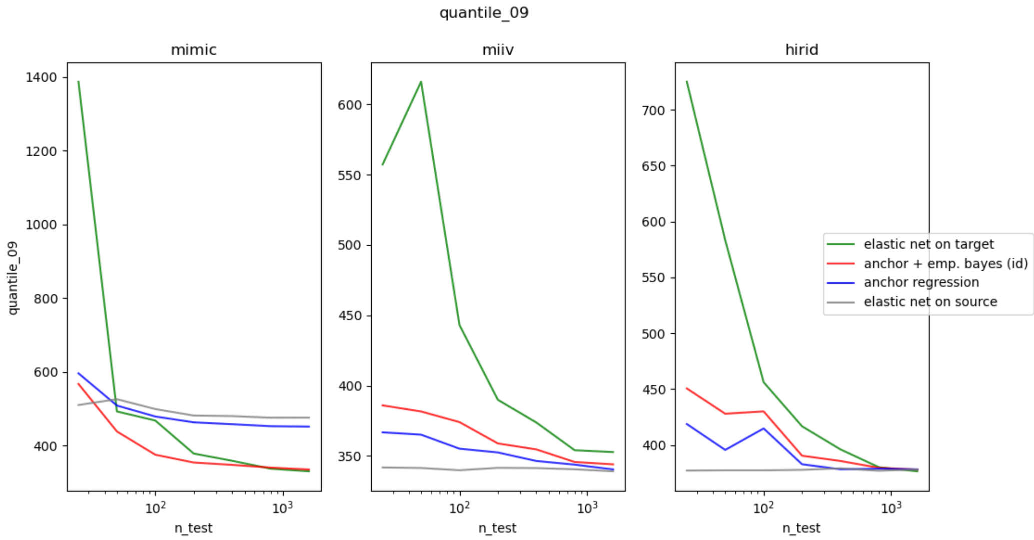
The following plot displays target distribution MSE (as measured on a large test set) by method and `n_test=25, 50, 100, 200, 400, 800, 1600`, averaged over 10 samples (seeds) of the `n_test` fine-tuning observations. The hyperparameters were chosen via 5-fold cross-validation on the fine-tuning dataset.



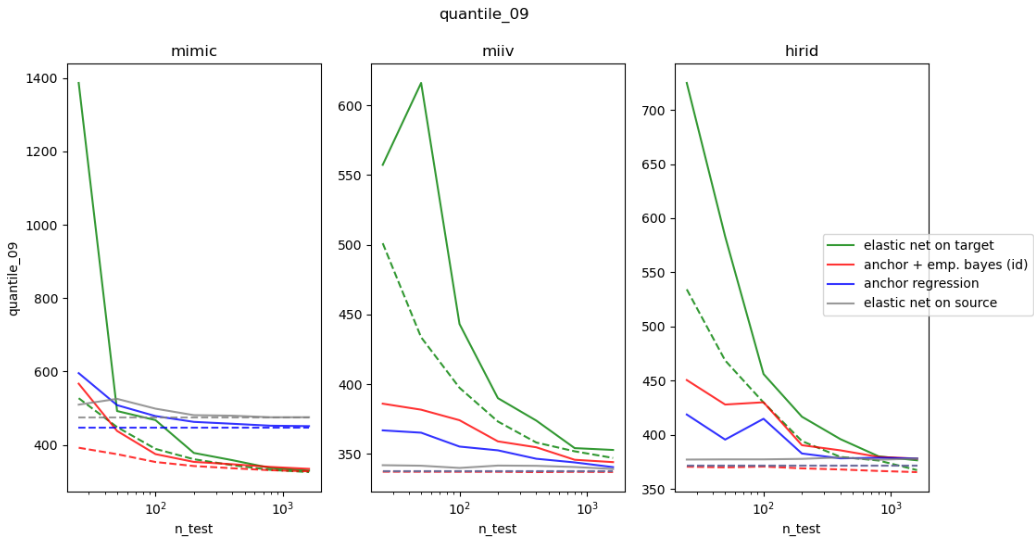
The following plots additionally show “oracle” performance, where the hyperparameters were chosen to minimize the test error, with a dashed line. For methods that use the fine-tuning dataset for hyperparameter optimization only (anchor regression, elastic net on source), this is independent of `n_test`.



The following plots display the same as above, but using the 90%-quantile of squared errors as a metric, both for evaluation and cross-validation. First without “oracle”:



Then with “oracle”:



Choice of γ

We are interested in which values for γ were chosen by the methods anchor regression and anchor + emp. Bayes.

For both methods, we display the median value of γ and mean test MSE over 10 samples of the fine-tuning dataset, of the method with hyperparameters (except for γ) chosen via 5-fold cross-validation, by the target and `n_test`.

anchor regression

Hyperparameters chosen via 5-fold cross-validation:

target	n_test	median_gamma	mean_test_mse
hirid	25	10	160.959
hirid	50	10	157.888
hirid	100	3.16228	158.125
hirid	200	3.16228	156.567
hirid	400	3.16228	154.945
hirid	800	6.58114	155.194
hirid	1600	10	154.404
miiv	25	51.5811	147.099
miiv	50	3.16228	147.822
miiv	100	2.08114	141.1
miiv	200	1	140.747
miiv	400	1	140.222
miiv	800	1	140.051
miiv	1600	1	139.844
mimic	25	100	206.687
mimic	50	31.6228	180.158
mimic	100	31.6228	170.394
mimic	200	31.6228	168.988
mimic	400	31.6228	168.495
mimic	800	31.6228	168.292
mimic	1600	31.6228	168.294

Hyperparameters chosen by looking at the test set (oracle):

target	median_gamma	mean_test_mse
hirid	1	154.068
miiv	1	139.711
mimic	31.6228	168.091

anchor + emp. Bayes

Hyperparameters chosen via 5-fold cross-validation:

target	n_test	median_gamma	mean_test_mse
hirid	25	10	174.283

target	n_test	median_gamma	mean_test_mse
hirid	50	65.8114	167.662
hirid	100	6.58114	159.316
hirid	200	6.58114	156.77
hirid	400	20.8114	155.658
hirid	800	17.3925	154.371
hirid	1600	2.08114	153.258
miiv	25	51.5811	157.941
miiv	50	10	152.988
miiv	100	3.16228	143.461
miiv	200	2.08114	144.056
miiv	400	1	140.213
miiv	800	1	140.247
miiv	1600	1	139.898
mimic	25	658.114	199.786
mimic	50	31.6228	170.821
mimic	100	31.6228	146.903
mimic	200	31.6228	135.856
mimic	400	20.8114	132.34
mimic	800	20.8114	130.214
mimic	1600	31.6228	128.905

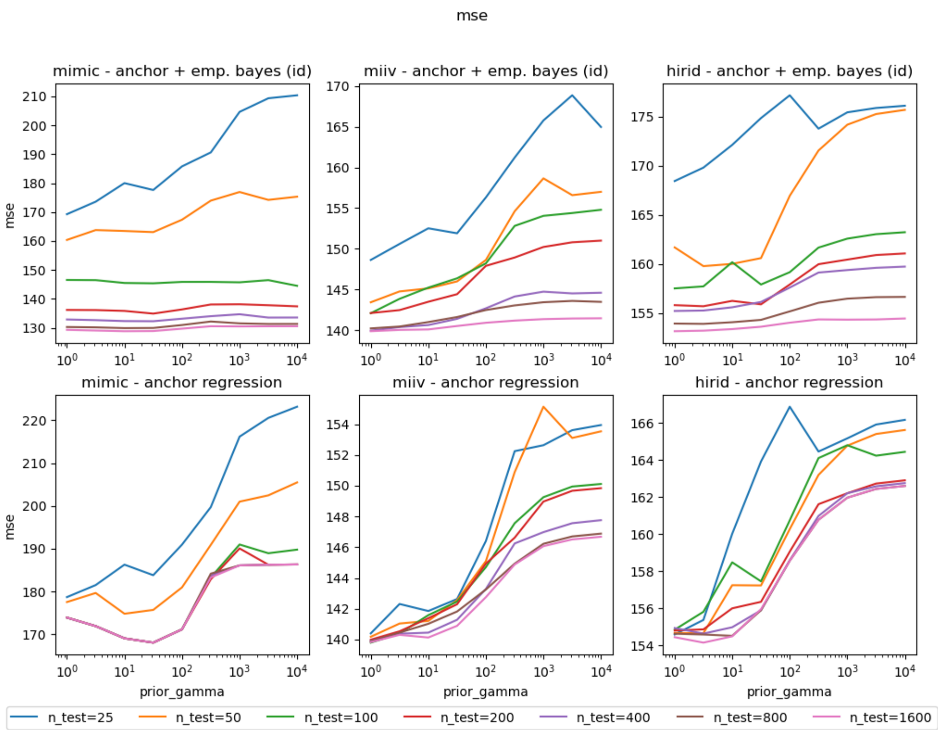
Hyperparameters chosen by looking at the test set (oracle):

target	n_test	median_gamma	mean_test_mse
hirid	25	1	153.924
hirid	50	1	153.726
hirid	100	1	153.745
hirid	200	3.16228	153.437
hirid	400	2.08114	153.429
hirid	800	3.16228	153.012
hirid	1600	1	152.446
miiv	25	1	139.708
miiv	50	1	139.706
miiv	100	1	139.705
miiv	200	1	139.686
miiv	400	1	139.7
miiv	800	1	139.676
miiv	1600	1	139.661
mimic	25	10	151.764
mimic	50	10	146.437
mimic	100	31.6228	138.249

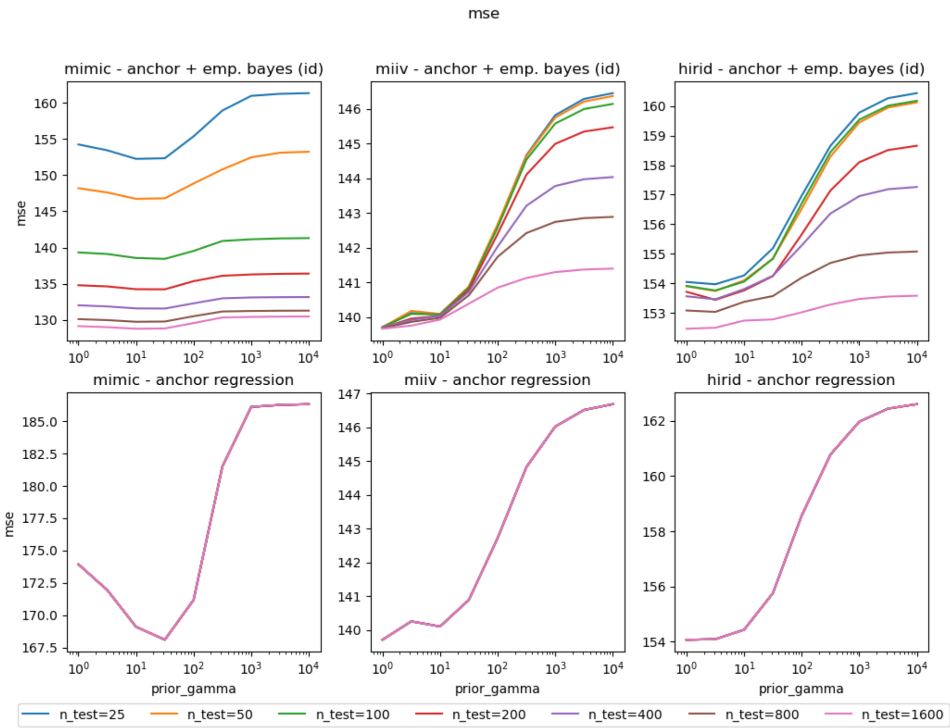
target	n_test	median_gamma	mean_test_mse
mimic	200	20.8114	134.141
mimic	400	31.6228	131.504
mimic	800	10	129.692
mimic	1600	10	128.754

Performance by `n_test` and γ

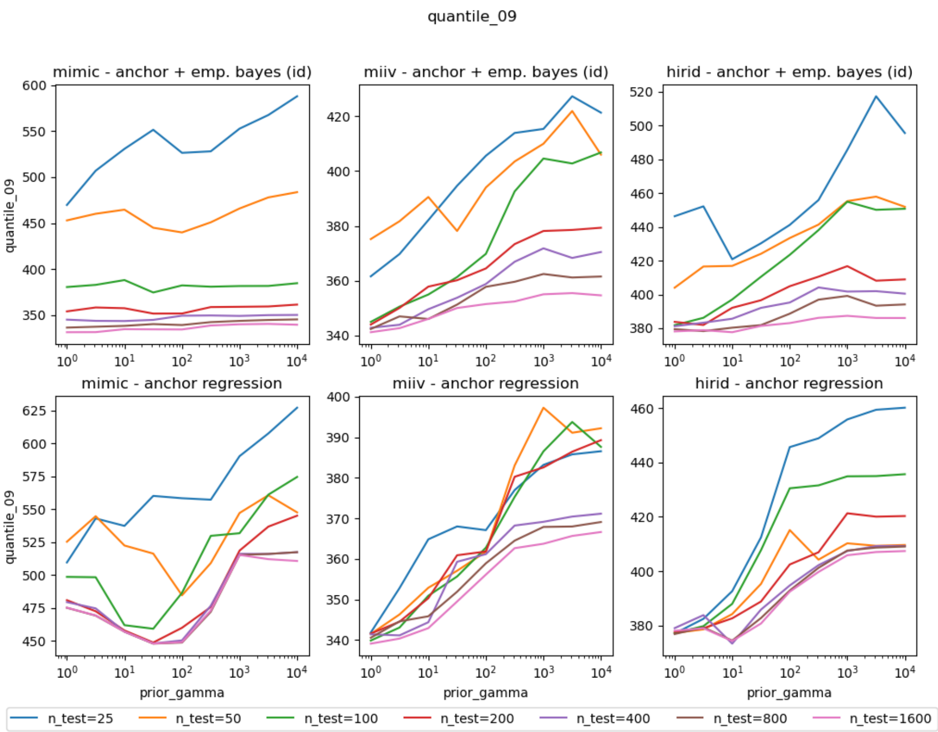
In the following, we plot the performance of the "best" hyperparameter configuration, by `n_test` and γ. That is, for both methods, for each target, each value of `n_test`, each value of γ, and each sample (seed) of the fine-tuning dataset, we choose the best hyperparameter combination (excluding γ) by 5-fold cross-validation on the fine-tuning dataset and average the test MSE over the samples (seeds).



The following chooses the hyperparameters by looking at the test set (oracle):



The following is the same as above, but using the 90% quantiles of squared errors as a metric. First using hyperparameters chosen via cross-validation on the fine-tuning dataset:



Next by looking at the test set (oracle):