Robust Prediction when Features are Missing
Xiuming Liu, Dave Zachariah, Petre Stoica, Fellow, IEEE

Abstract
Predictors are learned using past training data which may contain features that are unavailable at the time of prediction. We develop an approach that is robust against outlying missing features, based on the optimality properties of an oracle predictor which observes them. The robustness properties of the approach are demonstrated on both real and synthetic data.

I. INTRODUCTION
A common task in statistical machine learning and signal processing is to predict an outcome $y$ based on features $x$ and $z$, using past training data drawn from an unknown distribution

$$(x_i, z_i, y_i) \sim p(x, z, y), \quad i = 1, \ldots, n.$$ 

In certain problems, however, not all features in the training data are available at the time of prediction. For instance, in medical diagnosis certain features are more expensive or time-consuming to obtain than others, and therefore unavailable in an early stage of assessment. Other features are observable after the outcome has occurred. We let $z$ denote the features missing at the time of prediction and consider the task of predicting $y$ given only the observable features $x$.

A direct approach predicts only on the basis of the association between $x$ and $y$, and thus discard all past training data containing $z$. By contrast, missing data in statistics is commonly tackled by means of imputation [1], [2]. An indirect approach is then to predict $y$ using both $x$ and an imputed $\hat{z}(x)$. However, as we explain in Section [1] this turns out to be equivalent to the direct approach. For both approaches, learning a linearly parameterized predictor that minimizes the mean squared error (MSE), is shown to perform poorly when the missing features occur in the tails of the marginal distribution $p(z)$.

Robust statistics has typically focused on problems with contaminated training data [3] or heavy-tailed noise distributions [4], [5]. For the latter, Student-t distributions are often adopted in regression models...

Manuscript submitted 23 December, 2019. This research has been partly supported by the Swedish Research Council via the projects 2017-04543, 2017-04610, and 2018-05040.

The authors are with Department of Information Technology, Uppsala University, Sweden.
Fig. 1: Illustration of MSE conditioned on a missing scalar feature $z$ along with $p(z)$. The optimistic predictor $w_o$ may lead to high outlier MSE$_\alpha$ for the tails of $p(z)$. By contrast, the conservative predictor $w_c$ (see Section III for definition) can mitigate the outlier events. The MSE is lower bounded by that of an oracle predictor which observes $z$. The example used to generate this figure is specified in Section V-A.

in order to achieve robust estimation of model parameters such that outlying training data samples are downweighted. Our concern in this paper, however, is robust prediction in the case of outlying missing features.

Specifically, we achieve robustness using an adaptively weighted combination of optimistic and conservative predictors, which are derived in Section III. The approach of switching between modes during extreme events can be found in econometrics [6] and signal processing [7], [8], but has not been considered for prediction with missing features. We demonstrate the robustness properties of the proposed approach using both synthetic and real data sets.

**Notation:** We let $\|x\|_W = \sqrt{x^\top W x}$, where $W \succeq 0$, and define the sample mean of $x$ as $E_n[x] = n^{-1} \sum_{i=1}^{n} x_i$. The pseudoinverse of a matrix $A$ is denoted by $A^\dagger$.

II. PROBLEM FORMULATION

We consider scenarios in which

- $x$ and $z$ are correlated,
- the dimension of $x$ is greater than that of $z$,

and study the class of linearly parameterized predictors $\hat{y}(x; w) = w^\top x$, where $w \in \mathbb{R}^d$. Without loss of generality we consider $(x, z, y)$ to be centered. Note that the results in this paper can be readily extended
to arbitrary functions of the features by replacing $x$ in $w^\top x$ with a function $\phi(x)$.

The mean squared-error of a predictor is

$$\text{MSE}(w) \triangleq \mathbb{E} \left[ |y - \hat{y}(x; w)|^2 \right],$$

where the expectation is with respect to $(x, z, y)$. In the rest of this section we discuss briefly how a missing feature $z \in \mathbb{R}^q$ affects the prediction performance. The tails of the distribution of $z$ are contained in the region

$$Z_\alpha = \left\{ z : z^\top \left( \mathbb{E} \left[ zz^\top \right] \right)^{-1} z \geq q/\alpha \right\}$$

as $\alpha$ approaches 0. Indeed $\Pr\{z \in Z_\alpha\} \leq \alpha$ (see the appendix), and thus a small $\alpha$ corresponds to the probability of an outlier event. Making use of $Z_\alpha$, we can decompose the MSE into an outlier $\text{MSE}_\alpha = \mathbb{E} \left[ |y - \hat{y}|^2 \mid z \in Z_\alpha \right]$ and an inlier $\text{MSE}_{1-\alpha} = \mathbb{E} \left[ |y - \hat{y}|^2 \mid z \notin Z_{\alpha} \right]$:

$$\text{MSE} = \Pr\{z \in Z_{\alpha}\} \text{MSE}_\alpha + \Pr\{z \notin Z_{\alpha}\} \text{MSE}_{1-\alpha}.$$  

As Figure 1 illustrates, the prediction performance can degrade significantly for outlier events.

Using $n$ training samples, our goal is to formulate a robust predictor that will reduce the outlier $\text{MSE}_\alpha$ without incurring a significant increase of the inlier $\text{MSE}_{1-\alpha}$.

### III. Predictors: Optimistic, Conservative and Robust

If the feature $z$ were known, the optimal linearly parameterized predictor would be given by

$$\hat{y}_\star(x, z) = \alpha^\top x + \beta^\top z,$$

where

$$\begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} \mathbb{E}[xx^\top] & \mathbb{E}[xz^\top] \\ \mathbb{E}[zx^\top] & \mathbb{E}[zz^\top] \end{bmatrix}^{-1} \begin{bmatrix} \mathbb{E}[xy] \\ \mathbb{E}[zy] \end{bmatrix}$$

Its prediction errors are uncorrelated with both sets of features, that is,

$$\mathbb{E} [x(y - \hat{y})] = 0 \quad \text{and} \quad \mathbb{E} [z(y - \hat{y})] = 0,$$

which renders the predictor robust against outlier events for both $x$ and $z$, respectively. In the case of missing features $z$, we begin by considering predictors $\hat{y}(x; w) = w^\top x$ which satisfy either one of the orthogonality properties in \[5\].
A. Optimistic predictor

The predictors that satisfy the first equality in (5) are given by the parameter vectors in

$$\mathcal{W}_o = \left\{ w : \mathbb{E} [x(y - w^\top x)] = 0 \right\}$$

(6)

This set, however, consists of a single element $w_o$ which is also the minimizer of (1) [9]. That is,

$$w_o \equiv \arg \min_w \text{MSE}(w) = (\mathbb{E}[xx^\top])^{-1} \mathbb{E}[xy]$$

(7)

We denote the resulting predictor as ‘optimistic’ with respect to the missing $z$, because it does not attempt to satisfy the second equality in (5), see Fig. 1 for an illustration of its performance.

Remark: Given that $x$ and $z$ are correlated, we may consider using the MSE-optimal linear predictor

$$\hat{z} = \mathbb{E}[zz^\top](\mathbb{E}[xx^\top])^{-1}x$$

(8)

to impute the missing feature. An indirect predictor approach would then be to use $\hat{z}$ in (3) in lieu of $z$, but this is equivalent to (7). That is, $\hat{y}_\star(x, \hat{z}) \equiv w_o^\top x = \hat{y}(x; w_o)$, which can be shown by using the block matrix inversion lemma in (4).

B. Conservative predictor

The predictors that satisfy the second equality in (5) are given by all parameter vectors in

$$\mathcal{W}_c = \left\{ w : \mathbb{E} [z(y - w^\top x)] = 0 \right\}$$

(9)

This set is a $(d - q)$-dimensional subspace of $\mathbb{R}^d$ and therefore we can consider the parameter vector that minimizes (1), viz.

$$w_c = \arg \min_{w \in \mathcal{W}_c} \text{MSE}(w),$$

(10)

We denote the resulting predictor as ‘conservative’ with respect to the missing $z$, because it satisfies only the second equality in (5), see Fig. 1 for an illustration of its performance.

Remark: Comparing the error of $\hat{y}(x; w)$ with that of $\hat{y}_\star(x, z)$ in (3), the excess MSE can be expressed as

$$\text{MSE}(w) - \text{MSE}_\star = \|\Gamma(\alpha - w) + \beta\|_2^2 + \|\alpha - w\|_2^2 \geq 0,$$

(11)

where $\Gamma = (\mathbb{E}[zz^\top])^{-1} \mathbb{E}[zx^\top]$ and $\tilde{x} = x - \Gamma^\top z$ is a residual (see the appendix). Note that the first term in (11) is weighted by the dispersion of $z$. The constraint $w \in \mathcal{W}_c$ forces this term to zero. This leaves $d - q$ degrees of freedom that can be used to minimize the second term. By contrast, $w_o$ minimizes the sum of both terms.
C. Robust predictor

Satisfying only one of the equalities in (5) comes at a cost: The optimistic \( w_o \) yields robustness against outlying \( x \) but not \( z \) and, conversely, the conservative \( w_c \) yields robustness against outlying \( z \) but not \( x \). Since both equalities can be satisfied only by the infeasible predictor (3), we propose a predictor that interpolates between the optimistic and conservative modes using the side information that \( x \) provides about outliers in the missing features \( z \). That is, we propose to learn the adaptive parameter vector

\[
\hat{w}(x) = \Pr\{z \notin Z_\alpha | x\} w_o + \Pr\{z \in Z_\alpha | x\} w_c,
\]

such that the predictor \( \hat{y}(x) = w^\top(x)x \) becomes robust against outliers in both \( x \) and \( z \).

IV. LEARNING THE ROBUST PREDICTOR

Learning the robust predictor requires finding finite-sample approximations of \( w_o, w_c \) and \( \Pr\{z \in Z_\alpha | x\} \) in (12) using \( n \) training samples \( \{(x_i, z_i, y_i)\} \).

We begin by defining \( \text{MSE}_n(w) = \mathbb{E}_n[y - w^\top x]^2 \), which yields the empirical counterpart of (7):

\[
\hat{w}_o = \arg \min_w \text{MSE}_n(w) = (\mathbb{E}_n[xx^\top])^{\dagger} \mathbb{E}_n[xy] \quad (13)
\]

Note that the pseudoinverse is used to include cases in which the sample covariance matrices are degenerated. Similarly, for (10) we note that the empirical counterpart of the constraint in (9) is

\[
\mathbb{E}_n[z(y - w^\top x)] = 0 \iff \mathbb{E}_n[zx^\top]w = \mathbb{E}_n[z y] \quad (14)
\]

All vectors that satisfy (14) can therefore be parameterized as

\[
\hat{w}_c = w(\hat{\theta}),
\]

where \( \hat{\theta} \) is the minimizer of the quadratic function \( \text{MSE}_n(w(\theta)) \).

Next, we consider learning a model of the probability of an outlier event, \( \Pr\{z \in Z_\alpha | x\} \), conditioned on \( x \). Noting the definition (2), we predict an outlier event using the scalar

\[
\delta(x) = \sqrt{\hat{z}^\top(x) \mathbb{E}_n[zz^\top] \hat{z}(x)} \geq 0,
\]

where \( \hat{z}(x) = \mathbb{E}_n[zx^\top](\mathbb{E}_n[xx^\top])^{\dagger}x \) is the empirical version of (8). The conditional outlier probability is modeled using a standard logistic function,

\[
\hat{\Pr}\{z \in Z_\alpha | x\} = \frac{1}{1 + \exp \kappa(\delta(x) - \delta_0)} \quad (16)
\]
Fig. 2: Outlier and inlier $z$ in the training data (for $\alpha = 10\%$) versus statistic $\delta(x)$, along with fitted logistic model (16) of outlier probability conditioned on $x$. In this example, the learned parameters were $\delta_0 = 4.21$ and $\kappa = -1.78$, respectively. The example used to generate this figure is specified in Section V-A.

The model parameters $\kappa$ and $\delta_0$ are learned from the training data $\{(x_i, z_i)\}$ by minimizing the standard cross-entropy criterion

$$\min_{\kappa, \delta_0} - \mathbb{E}_n \left[ I(z \in Z_\alpha) \ln \hat{\Pr}\{z \in Z_\alpha|x\} + I(z \notin Z_\alpha) \ln (1 - \hat{\Pr}\{z \in Z_\alpha|x\}) \right].$$

(17)

This approach takes into account the inherent uncertainty of predicting an outlying $z$ from $x$. An example of a fitted model as in (17) is presented in Figure 2.

In sum, we learn a robust predictor with an adaptive parameter vector

$$\hat{w}(x) = \hat{\Pr}\{z \in Z_\alpha|x\} \hat{w}_o + \hat{\Pr}\{z \notin Z_\alpha|x\} \hat{w}_c$$

(18)

using $n$ samples, as described in Algorithm 1.

**Algorithm 1** Learning the robust predictor

1: **Input:** Training data $\{(x_i, y_i, z_i)\}$ and $\alpha \in (0, 1]$

2: Compute $\hat{w}_o$ via (13)

3: Compute $\hat{w}_c$ via (15)

4: For each $(x_i, z_i)$, form $(\delta(x_i), I(z_i \in Z_\alpha))$

5: Learn $\hat{\Pr}\{z \in Z_\alpha|x\}$ via (17)

6: **Output:** $\hat{w}(x)$ in (18)

Remark: In the case of high-dimensional features $x$ and $z$ one may use regularized methods, such as ridge regression, LASSO, or the tuning-free SPICE method [10], to learn $\hat{w}_o$, $\hat{w}_c$ and $\hat{z}(x)$. 
Fig. 3: Conditional MSE versus the missing feature $z$ for different predictors. The bands correspond to the 75% percentiles and the lines represent the median over 50 Monte Carlo runs. In comparison with the optimistic $w_o$, the robust predictor $\hat{w}(x)$ significantly reduces the errors in the tails of $z$, while incurring only a small increased error for inlier $z$.

V. EXPERIMENTAL RESULTS

We evaluate the robustness of the proposed predictor using both synthetic and real data.

A. Synthetic data

Consider the following data-generating process of $z \in \mathbb{R}$, $x \in \mathbb{R}^d$, and $y \in \mathbb{R}$:

$$z \sim \text{St}(0, 1, \nu_z),$$

$$x = 1z + u + \epsilon_x,$$

$$y = z + 1^\top x + \epsilon_y,$$

where $u \sim \text{St}(0, \Sigma_u, \nu_u)$ is a $d = 3$-dimensional t-distributed latent variable with $\nu_u$ degrees of freedom and $(\epsilon_x, \epsilon_y)$ are white Gaussian processes of corresponding dimensions.

We evaluate the predictors $\hat{y}(x; w)$ of a new outcome $y$ given only $x$, where the vector $w$ is learned from $n$ samples of training data. Specifically, we evaluate the optimistic $\hat{w}_o$, conservative $\hat{w}_c$ and proposed $\hat{w}(x)$ predictors in a case of missing features with heavy tails ($\nu_z = 3$), where $n$ ranges from 100 to 1000 and $\alpha = 0.1$. A comparison of the conditional MSE functions of the learned predictors ($n = 1000$) is given in Fig. 3, where it is seen that the robust predictor smoothly interpolates between the two modes.

When averaging over $z$, the distributions of MSEs for 50 training datasets are illustrated in Fig. 4. We see that the robust predictor drastically reduces the outlier MSE$_{\alpha}$, while yielding only a small increase in
the inlier MSE$_{1-\alpha}$. The differences in MSEs, when averaged over all training datasets, are summarized in Tables I which demonstrates the robustness of the proposed approach.

| n    | \( \hat{w}_c \) | \( \Delta \text{MSE}_{1-\alpha} \) | \( \Delta \text{MSE}_\alpha \) | \( \hat{w}(x) \) | \( \Delta \text{MSE}_{1-\alpha} \) | \( \Delta \text{MSE}_\alpha \) |
|------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| 100  | 58.9            | -81.4           | 17.3            | -72.4           |                  |                  |
| 500  | 63.6            | -81.0           | 16.9            | -66.9           |                  |                  |
| 1000 | 56.5            | -87.7           | 14.4            | -75.3           |                  |                  |

TABLE I: Changes in averaged MSE in comparison to the MSE of \( \hat{w}_c \) in [%], for \( \alpha = 10\% \). Average of 50 simulations using \( 10^6 \) test samples.

B. Air quality data

Next, we demonstrate the proposed method using real-world air quality data. Nitrogen-oxides (NO\(_x\)) emitted by the fossil fuel vehicles are major air pollutants in urban environments, with negative impacts on the health of inhabitants.

The aim here is to predict the daily average of NO\(_x\) concentration, denoted \( y \), based on NO\(_x\) and ozone (O\(_3\)) measurements from \( L \) previous days. That is, \( x \) is of dimension \( d = 2L \) and contains the daily average NO\(_x\) and O\(_3\) levels from the \( L \) past days. In the training data we have also access to \( z \), the O\(_3\) concentration, at the same time as the outcome \( y \). This feature \( z \) is correlated with \( y \) and \( x \). For the prediction of a new outcome, however, \( z \) is a missing feature.

The dataset contains 10 years of daily average NO\(_x\) and O\(_3\) measurements from 2006-01-01 to 2015-12-31. Data is split into 7 years of training data (2006-2012), and 3 years of test data (2012-2015). Using
\( \alpha = 30\% \) in the definition of \( Z_\alpha \) the results in Table II show that we are able to reduce the outlier MSE\(_\alpha\) by about 10% while incurring a minimal increase of the inlier MSE\(_{1-\alpha}\).

\[
\begin{array}{cccccc}
L & \hat{w}_c & \hat{w}(x) & \Delta \text{MSE}_{1-\alpha} & \Delta \text{MSE}_\alpha & \Delta \text{MSE}_{1-\alpha} \\
7 & 4.1 & -18.0 & 0.7 & -6.9 & \\
28 & 3.5 & -24.7 & 0.4 & -11.4 & \\
56 & 3.1 & -23.1 & 0.4 & -11.2 & \\
\end{array}
\]

TABLE II: Real-world NO\(_x\) and O\(_3\) data. Changes in the MSE in comparison to the MSE of \( \hat{w}_o \) in [%], for \( \alpha = 30\% \).

VI. CONCLUSION

Based on the orthogonality properties of an optimal oracle predictor, we developed a predictor that is robust against outliers of the missing features. The proposed predictor is formulated as a convex combination of optimistic and conservative predictors, and requires only specifying the intended outlier level against which it must be robust. The ability of the robust predictor to suppress outlier errors, while incurring only a minor increase in the inlier errors, was demonstrated using both simulated and real-world datasets.

APPENDIX A

PROOFS

A. Probability bound for (2)

The probability bound for an event \( z \in Z_\alpha \) follows readily from a Chebychev-type inequality:

\[
\Pr \{ z \in Z_\alpha \} = \int_{Z_\alpha} p(z) \, dz 
\leq \int_{Z_\alpha} \left[ (\alpha/q) z^\top (\mathbb{E}[zz^\top])^{-1} z \right] p(z) \, dz 
\leq \alpha \int \left[ z^\top (\mathbb{E}[zz^\top])^{-1} z/q \right] p(z) \, dz 
= \alpha
\]

\[1\]In this example \( \alpha \) is set to a larger value such that there are enough outlier events in the training data for the model fitting in (17).
B. MSE decomposition \[(11)\]

The outcome $y$ can always be decomposed as

$$y = \alpha^\top x + \beta^\top z + v,$$

where $\text{MSE}_\star = \mathbb{E}[v^2]$. The random variable $v$ is orthogonal to any linear function of $x$ and $z$; which includes the residual $\tilde{x} = x - \Gamma^\top z$. Using (20) we can express the prediction error as

$$y - w^\top x = [\Gamma(\alpha - w) + \beta]^\top z + (\alpha - w)^\top \tilde{x} + v$$

Squaring this expression and taking the expectation yields (11). Similarly, inserting it into the constraint in (9) yields $\Gamma(\alpha - w) + \beta = 0$.

REFERENCES

[1] R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data.*, 3rd Edition. Wiley, 2019.

[2] O. Chapelle, B. Schölkopf, and A. Zien, *Semi-supervised learning*. Cambridge, Mass: MIT Press, 2006.

[3] P. J. Huber, “Robust estimation of a location parameter,” *The Annals of Mathematical Statistics*, pp. 73–101, 1964.

[4] K. L. Lange, R. J. Little, and J. M. Taylor, “Robust statistical modeling using the t distribution,” *Journal of the American Statistical Association*, vol. 84, no. 408, pp. 881–896, 1989.

[5] J. Christmas and R. Everson, “Robust autoregression: Student-t innovations using variational bayes,” *IEEE Transactions on Signal Processing*, vol. 59, no. 1, pp. 48–57, 2010.

[6] J. D. Hamilton and R. Susmel, “Autoregressive conditional heteroskedasticity and changes in regime,” *Journal of econometrics*, vol. 64, no. 1-2, pp. 307–333, 1994.

[7] A. Poritz, “Linear predictive hidden markov models and the speech signal,” in *ICASSP’82. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 7, pp. 1291–1294, IEEE, 1982.

[8] Y. Ephraim and W. J. Roberts, “Revisiting autoregressive hidden markov modeling of speech signals,” *IEEE Signal processing letters*, vol. 12, no. 2, pp. 166–169, 2005.

[9] T. Kailath, A. H. Sayed, and B. Hassibi, *Linear estimation*. Upper Saddle River, NJ: Prentice Hall, 2000.

[10] D. Zachariah and P. Stoica, “Online hyperparameter-free sparse estimation method,” *IEEE Transactions on Signal Processing*, vol. 63, no. 13, pp. 3348–3359, 2015.