Making Energy Forecasting Resilient to Missing Features: a Robust Optimization Approach

Akylas Stratigakos, Panagiotis Andrianesis, Andrea Michiorri, and Georges Kariniotakis

Mines Paris, PSL University, Center PERSEE
Sophia Antipolis, France

akylas.stratigakos@minesparis.psl.eu

42nd International Symposium on Forecasting

PSL | PERSEE

Outline

Outline

Forecasting applications in power systems (*energy forecasting*) are mostly data-driven:

- Performance depends on data quality and availability.
- Data-management issues [1] that appear in industrial applications: missing data, outliers, distribution shift.
- Some issues emerge only after the model is deployed.

Missing features (or *feature deletion*) in an operational setting:

- Subset of features used for training is unavailable at test time.
- Reasons: network latency, APIs, cyber-attacks, equipment failures...
- Assessment on ENTSO-E's Transparency platform: "for every data domain, fewer than 40% of users reported that data were always there when needed" [2].

Forecasting applications in power systems (*energy forecasting*) are mostly data-driven:

- Performance depends on data quality and availability.
- Data-management issues [1] that appear in industrial applications: missing data, outliers, distribution shift.
- Some issues emerge only after the model is deployed.

Missing features (or *feature deletion*) in an operational setting:

- Subset of features used for training is unavailable at test time.
- Reasons: network latency, APIs, cyber-attacks, equipment failures...
- Assessment on ENTSO-E's Transparency platform: "for every data domain, fewer than 40% of users reported that data were always there when needed" [2].

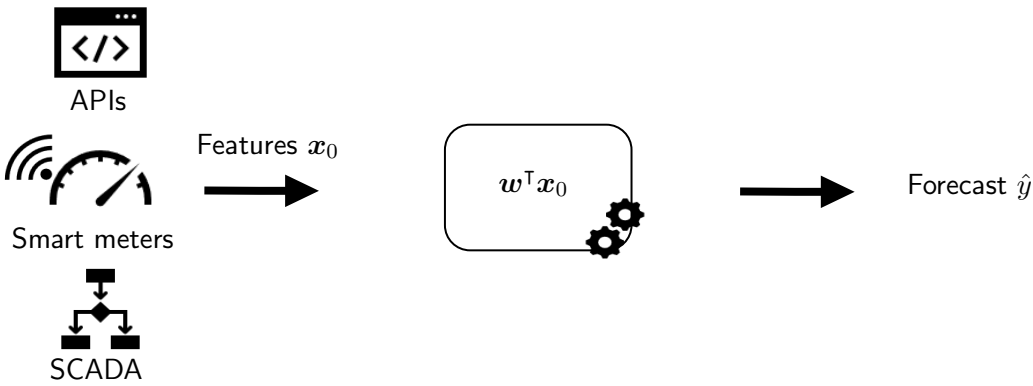
Model training

Data set $\{\boldsymbol{x}_i, y_i\}_{i=1}^n$

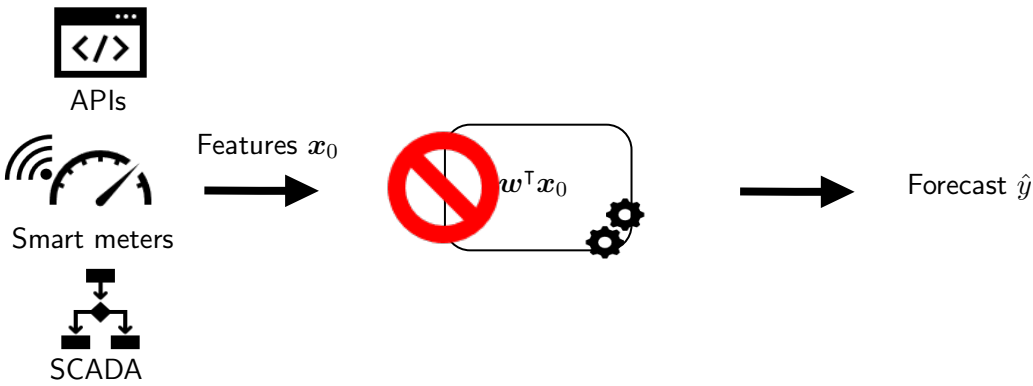
$$\min_{\boldsymbol{w}} \frac{1}{n} \sum_{i=1}^{n} l\left(y_i - \boldsymbol{w}^\intercal \boldsymbol{x}_i\right)$$

Parameters $\boldsymbol{w}$

Model deployment



Model deployment



Aim and contribution

Ad-hoc solutions:

- Involve manual tuning and heuristics, increase modeling complexity.
- Retrain without missing features outperforms "impute, then forecast" [3], but is impractical.

Ideally, deployed models should be resilient and maintain consistent performance without increasing complexity.

Design regression models that optimally resilient to feature deletion at test time

- Principled approach to improve model resilience, only requires solving an LP.
- Benchmarking energy forecasting under feature deletion.

Aim and contribution

Ad-hoc solutions:

- Involve manual tuning and heuristics, increase modeling complexity.
- Retrain without missing features outperforms "impute, then forecast" [3], but is impractical.

Ideally, deployed models should be resilient and maintain consistent performance without increasing complexity.

> Design regression models that optimally resilient to feature deletion at test time

- Principled approach to improve model resilience, only requires solving an LP.
- Benchmarking energy forecasting under feature deletion.

Outline

Proposed Robust Formulation

Standard linear regression problem:

- Given n observations of target $y \in \mathbb{R}$ and features $\boldsymbol{x} \in \mathbb{R}^p$, estimate parameters $\boldsymbol{w} \in \mathbb{R}^p$ by minimizing loss function l:

$$\min_{\boldsymbol{w}} \frac{1}{n} \sum_{i=1}^{n} l\left(y_i - \boldsymbol{w}^\intercal \boldsymbol{x}_i\right)$$

Modeling feature uncertainty:

- Introduce $\boldsymbol{\alpha} \in \{\mathbf{0}, \mathbf{1}\}^p$ and model features as $\boldsymbol{x}_i \odot (\mathbf{1} - \boldsymbol{\alpha})$, where $\alpha_j = 1$ if the j-th feature is missing (*same* features are missing in all samples).
- Some features cannot be deleted (e.g. calendar variables) and others are grouped (e.g. polynomial, interactions) $\rightarrow$use $M \in \mathbb{R}^{m \times p}$ to model additional constraints.

Proposed Robust Formulation

Standard linear regression problem:

- Given n observations of target $y \in \mathbb{R}$ and features $\boldsymbol{x} \in \mathbb{R}^p$, estimate parameters $\boldsymbol{w} \in \mathbb{R}^p$ by minimizing loss function l:

$$\min_{\boldsymbol{w}} \frac{1}{n} \sum_{i=1}^{n} l\left(y_i - \boldsymbol{w}^\intercal \boldsymbol{x}_i\right)$$

Modeling feature uncertainty:

- Introduce $\boldsymbol{\alpha} \in \{\mathbf{0}, \mathbf{1}\}^p$ and model features as $\boldsymbol{x}_i \odot (\mathbf{1} - \boldsymbol{\alpha})$, where $\alpha_j = 1$ if the j-th feature is missing (*same* features are missing in all samples).
- Some features cannot be deleted (e.g. calendar variables) and others are grouped (e.g. polynomial, interactions) $\rightarrow$use $\boldsymbol{M} \in \mathbb{R}^{m \times p}$ to model additional constraints.

Proposed Robust Formulation

- Discrete uncertainty set: $\mathcal{U} = \{\boldsymbol{\alpha} \mid \boldsymbol{\alpha} \in \{\mathbf{0}, \mathbf{1}\}^p, \mathbf{1}^\intercal \boldsymbol{\alpha} = \Gamma, \boldsymbol{M}\boldsymbol{\alpha} = \mathbf{0}\}$, where Γ (integer) is the budget of robustness.
- *Feature-deletion robust regression* (FDRR) minimizes the worst-case loss when Γ features are missing:

$$\min_{\boldsymbol{w}} \max_{\boldsymbol{\alpha} \in \mathcal{U}} \sum_{i=1}^{n} l(y_i - \boldsymbol{w}^\intercal(\boldsymbol{x}_i \odot (\mathbf{1} - \boldsymbol{\alpha})))$$

Choosing a loss l:

- Quantile (pinball) loss and ℓ_1-norm $l(\cdot) = |\cdot|$
- For example, FDRR with ℓ_1 loss: $\min_{\boldsymbol{w}} \max_{\boldsymbol{\alpha} \in \mathcal{U}} \sum_{i=1}^{n} \left| y_i - \boldsymbol{w}^\intercal(\boldsymbol{x}_i \odot (\mathbf{1} - \boldsymbol{\alpha})) \right|$

Proposed Robust Formulation

- Discrete uncertainty set: $\mathcal{U} = \{\boldsymbol{\alpha} \mid \boldsymbol{\alpha} \in \{\mathbf{0}, \mathbf{1}\}^p, \mathbf{1}^\intercal \boldsymbol{\alpha} = \Gamma, \boldsymbol{M}\boldsymbol{\alpha} = \mathbf{0}\}$, where Γ (integer) is the budget of robustness.
- *Feature-deletion robust regression* (FDRR) minimizes the worst-case loss when Γ features are missing:

$$\min_{\boldsymbol{w}} \max_{\boldsymbol{\alpha} \in \mathcal{U}} \sum_{i=1}^{n} l\big(y_i - \boldsymbol{w}^\intercal(\boldsymbol{x}_i \odot (\mathbf{1} - \boldsymbol{\alpha}))\big)$$

Choosing a loss l:

- Quantile (pinball) loss and ℓ_1-norm $l(\cdot) = |\cdot|$
- For example, FDRR with ℓ_1 loss: $\min_{\boldsymbol{w}} \max_{\boldsymbol{\alpha} \in \mathcal{U}} \sum_{i=1}^{n} \big|y_i - \boldsymbol{w}^\intercal(\boldsymbol{x}_i \odot (\mathbf{1} - \boldsymbol{\alpha}))\big|$

Solution methodology

Exact solution:

- $\mathcal{U}$ is finite $\rightarrow$robust problem can be solved with vertex enumeration, but this leads to an LP which grows combinatorially.

Conservative approximation:

- Define polyhedral uncertainty set:
  $\mathcal{A} = \text{conv}(\mathcal{U}) = \{\boldsymbol{\alpha} \mid \mathbf{0} \leq \boldsymbol{\alpha} \leq \mathbf{1}, \mathbf{1}^\intercal \boldsymbol{\alpha} = \Gamma, M\boldsymbol{\alpha} = \mathbf{0}\}.$
  - LP relaxation of inner $\max$
  - Affinely Adjustable Reformulation of $|\cdot|$
  - Duality reformulation
  - Tractable LP

Key takeaway: the problem is solvable.

Solution methodology

Exact solution:

- $\mathcal{U}$ is finite $\rightarrow$robust problem can be solved with vertex enumeration, but this leads to an LP which grows combinatorially.

Conservative approximation:

- Define polyhedral uncertainty set:
  $\mathcal{A} = \text{conv}(\mathcal{U}) = \{\boldsymbol{\alpha} \mid \mathbf{0} \leq \boldsymbol{\alpha} \leq \mathbf{1}, \mathbf{1}^\intercal \boldsymbol{\alpha} = \Gamma, \boldsymbol{M}\boldsymbol{\alpha} = \mathbf{0}\}$.
  - LP relaxation of inner $\max$
  - Affinely Adjustable Reformulation of $|\cdot|$
  - Duality reformulation
  - Tractable LP

Key takeaway: the problem is solvable.

Outline

Setting: Day-ahead horizon (12h-36h ahead), data arriving in batches, point forecasts

Data set	Source	Features
Prices	FR, ENTSO-E	Lags, calendar, net load, system margin
Load* (21 series)	GEFCom 2012	Vanilla model [4] for multiple weather stations
Wind* (10 series)	GEFCom 2014	Wind speed/dir. (10m, 100m), Fourier terms for diurnal patterns
Solar† (3 series)	GEFCom 2014	Numerical weather predictions

*: features deleted in groups, †: one model per hour.

Benchmarks

Construct a feature vector $\boldsymbol{x}$, train the following models:

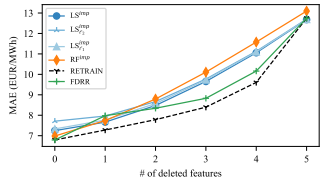
- $\texttt{LS}^*$: a least squares regression with adequate performance.
- $\texttt{LS}^*_{\ell_1 \backslash \ell_2}$: the same model as above with ℓ_1 (lasso) and ℓ_2 (ridge) regularization penalty.
- $\texttt{RF}^*$: a Random Forest model trained on the same set of features.
- RETRAIN [3]: an ℓ_1 regression model retrained for each combination of missing features. A total of $\sum_{k=1}^{p} \binom{p}{k}$ additional models is required.
- $\texttt{FDRR}^{\Gamma}$: a robust ℓ_1 regression with Γ indicating the robustness budget (a different model is trained for each Γ).

* missing data is filled with mean imputation.

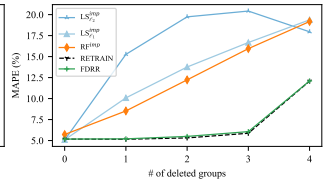
Forecast accuracy versus # of missing features

Randomly deleting a number of features from the test set:

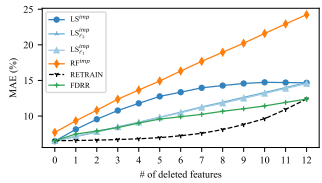
- FDRR improves resiliency compared to LS, outperforms regularized models with imputation.
- Performance of FDRR is comparable to RETRAIN.
- RETRAIN's complexity: for solar production, > 4000 models are trained *per* hour.



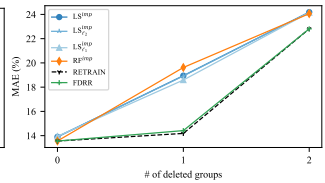
(a) Electricity prices.



(b) Load.

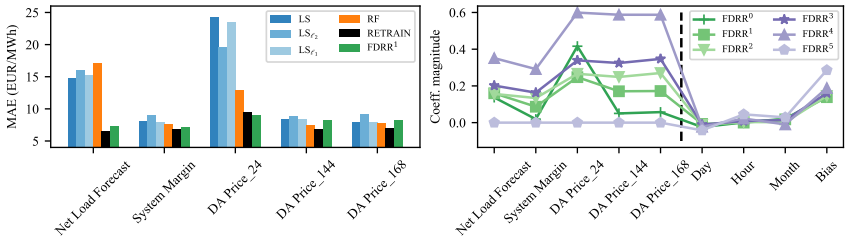


(c) Solar production.



(d) Wind production.

Worst-case scenario and interpretation



(a) Uniform feature deletion.

(b) Estimated coefficients.

- FDRR hedges against the worst-case scenario of uniformly deleting the most important feature from test set (left).
- As Γ increases, the weights are more evenly distributed across features (right).
- Intuitively, FDRR finds the most important features and mitigates their relative impact on accuracy.

Realistic setting

Varying the percentage of test observations with missing features:

- RETRAIN is the best overall, but the difference with FDRR is negligible up to 10%.
- Standard regularization is also beneficial.
- Relative decrease of FDRR from 0% to 50%: 21% for prices, 19% for wind production, 20% for load, and 30% for solar.
- Most resilient LS-type: 18% for prices (but worse in absolute terms), 27% for wind, 96% for load, and 33% for solar.

	% of obs.	0	1	5	10	25	50
Prices	LS	7.25	7.27	7.39	7.52	7.91	8.57
	LS_{ℓ_2}	7.71	7.73	7.83	7.95	8.29	8.87
	LS_{ℓ_1}	7.33	7.36	7.47	7.6	7.99	8.64
	RF	6.95	6.98	7.12	7.28	7.78	8.61
	FDRR	**6.79**	**6.82**	**6.93**	7.07	7.5	8.21
	RETRAIN	**6.79**	**6.82**	**6.91**	**7.03**	**7.39**	**7.98**
Load	LS	5.22	6.94	13.98	22.03	47.1	88.78
	LS_{ℓ_2}	**5.07**	5.21	5.73	6.41	8.41	11.61
	LS_{ℓ_1}	5.09	5.2	5.58	6.08	7.54	9.98
	RF	5.72	5.8	6.13	6.55	7.75	9.86
	FDRR	5.18	**5.19**	**5.28**	5.38	5.68	6.21
	RETRAIN	5.17	**5.19**	**5.27**	**5.37**	**5.64**	**6.16**
Wind	LS	13.9	13.98	14.28	14.66	15.83	17.79
	LS_{ℓ_2}	13.9	13.98	14.28	14.65	15.83	17.78
	LS_{ℓ_1}	13.95	14.02	14.32	14.68	15.83	17.71
	RF	13.57	13.65	13.98	14.38	15.62	17.77
	FDRR	**13.55**	**13.6**	**13.79**	14.04	14.82	16.14
	RETRAIN	**13.55**	**13.59**	**13.78**	**14.02**	**14.79**	**16.07**
Solar	LS	**6.47**	6.54	6.8	7.11	8.06	9.58
	LS_{ℓ_2}	**6.47**	**6.53**	6.7	6.92	7.58	8.67
	LS_{ℓ_1}	6.51	6.57	6.73	6.94	7.59	8.66
	RF	7.73	7.84	8.2	8.63	10.01	12.27
	FDRR	6.54	6.58	6.74	6.95	7.56	8.52
	RETRAIN	6.54	6.55	**6.62**	**6.71**	**6.96**	**7.37**

Outline

Conclusions and Next Steps

Resilient energy forecasting to handle missing data at test time:

- Consistent performance, lower degradation, hedging against the worst-case scenario.
- Requires only an LP, instead of training a large number of additional models

Broader perspective:

- It is important to also consider model resiliency, besides accuracy.
- Robust optimization offers tools to deal with feature uncertainty.

Next steps:

- Accuracy-resilience trade-off in standard regularization methods (ridge, lasso)
- Case studies: smart meter data, intra-day wind forecasting, etc.

Outline

References I

[1] N. Polyzotis, S. Roy, S. E. Whang, and M. Zinkevich, "Data management challenges in production machine learning," in *Proceedings of the 2017 ACM International Conference on Management of Data*, pp. 1723–1726, 2017.

[2] European Commission, "A review of the entso-e transparency platform," 2017.

[3] R. Tawn, J. Browell, and I. Dinwoodie, "Missing data in wind farm time series: Properties and effect on forecasts," *Electric Power Systems Research*, vol. 189, p. 106640, 2020.

[4] T. Hong, *Short term electric load forecasting*. North Carolina State University, 2010.

Thanks!