Every Sample a Task: Pushing the Limits of Heterogeneous Models with Personalized Regression

Ben Lengerich[1,2], Bryon Aragam[2], Eric P. Xing[1,2,3]
{blengeri, naragam, epxing}@cs.cmu.edu
1) Machine Learning Department, Carnegie Mellon University
2) Computer Science Department, Carnegie Mellon University, 3) Petuum, Inc.

Guiding Question

Can **collections** of **simple** models outperform large models if each simple model is used for only a single sample?

Motivation

Typical tug-of-war:

Accuracy ⟷ **Interpretability**

What if this tradeoff is a byproduct of using population-level models to learn effects which actually differ between samples?

Can we instead learn *sample-specific models*? Would give us:

- A simple, interpretable model for each sample
- Representational capacity from the entire collection of models.

Let's use a multi-task framework, defining each training sample as a task, to share power and learn these *sample-specific* models.

Problem Formulation

Given samples with predictors $X^1, \dots, X^n$, covariates $U^1, \dots, U^n$ and labels $Y^1, \dots, Y^n$ with $X^i \in \mathbb{R}^p$, $U^i \in \mathbb{R}^k$, $Y^i \in \mathbb{R}$, we seek a collection of models $\boldsymbol{B} = \{\beta^1, \dots, \beta^n\}$ and a distance function $d(\cdot,\cdot)$ which jointly minimize expectation of the test loss $l(Y^{test}, X^{test}, \beta^{\eta(U^{test})})$ where the model to use is chosen from $\boldsymbol{B}$ by the distance metric.

Related Work

- **Mixture Models** estimate a small number of components, typically independently.
- **Varying Coefficients (VC) [1]** specify a function to generate regression parameters from covariates.
- **Contextual Parameters [2]** generalize VC models to use deep networks as context encoders.
- **Sample-specific PGMs [4,5]** use model structure to test individual observations for deviations away from a population mean.

We are testing a general framework for sample-specific model inference which does not require specifying a parameter-generating function.

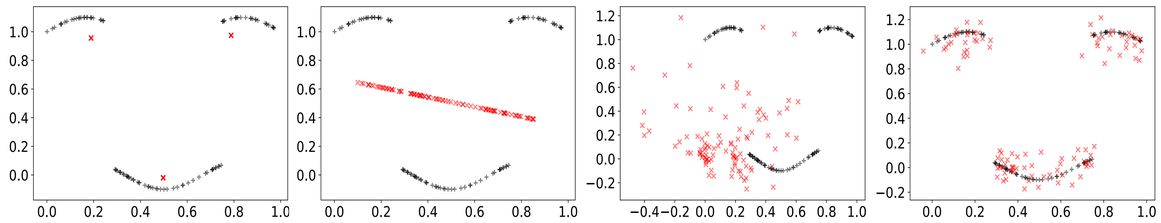


Figure 1 Illustration of the benefits of personalized models. Each point represents the regression parameters for a sample. Black points indicate true effect sizes, while the red points are estimates. Mixture models (a) estimate a limited number of models. The varying-coefficients model (b) estimates sample-specific models but the non-linear structure of the true parameters violates the model assumptions, leading to a poor fit. The locally-linear models induced by a deep learning model (c) do not accurately recover the underlying effect sizes. In contrast, personalized regression (d) accurately recovers effect sizes.

Personalized Regression

Individual samples become different tasks in a multi-task framework:

$$\mathcal{L}(X, Y, \boldsymbol{B}, d) = \sum_{i=1}^{n} l(X^i, Y^i, \beta^i) + \lambda\rho(\beta^i) + \gamma\varrho(\boldsymbol{B}, d)$$

Prediction Loss — Regularizer — Distance-Matching Regularizer

Distance-Matching Regularization

To share power, match structure in *covariates* to structure in *regression parameters*:

$$\varrho(\boldsymbol{B}, d) = \sum_{i=1}^{n}\sum_{j\neq i}\left(d(U^i, U^j) - \|\beta^i - \beta^j\|\right)^2$$

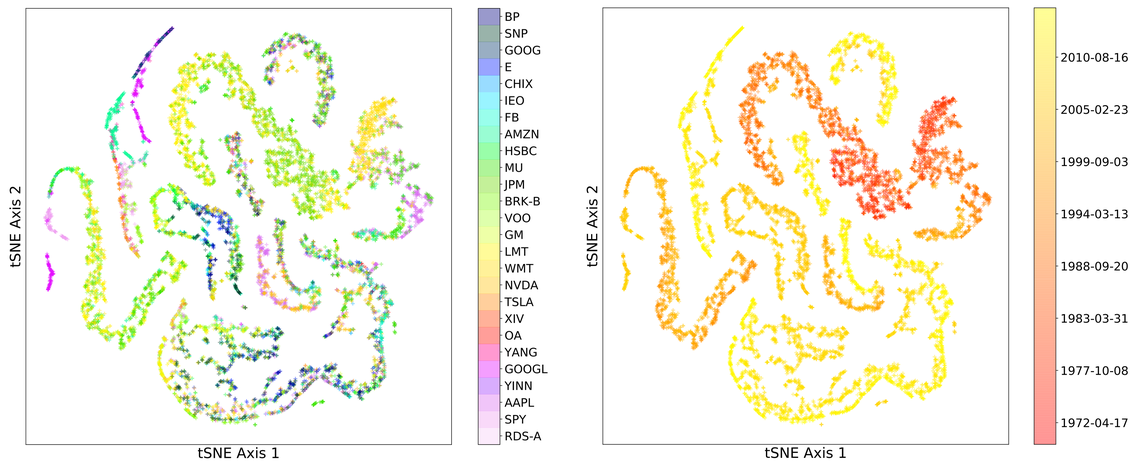


Figure 2 Visualization of personalized models trained on a financial dataset. Personalized financial models (t-SNE (Van Der Maaten, 2014) projection). Each point represents a regression model for one security at a single date. There is strong clustering in models according to both industry (a) and time (b), but neither covariate would be sufficient to completely characterize each sample.

Conclusions

- We have presented *Personalized Regression* to estimate collections of regression models by matching the structure of regression parameters to the structure of covariates.
- Personalized models often outperform the predictive accuracy of larger models because they can model effect sizes which vary between samples.
- Underscores the importance of treating sample heterogeneity directly rather than building increasingly-complicated cohort-level models.

Open Questions

- How many datasets have hidden sample-specific effects which could use sample-specific models?
- Is Personalized Regression the best framework to learn sample-specific models? Should models vary according to predictors or according to covariates?
- Can these ideas be used to regularize estimation of standard models (e.g. mixture models with DMR)?

Model	Simulation			Financial		Cancer	
	$\|\hat{\beta} - \beta\|_2$	R^2	MSE	R^2	MSE	AUROC	Acc
Population	17.40	0.632	0.092	0.006	64144	0.794	0.962
Mixture	15.50	0.857	0.036	0.738	16146	0.876	0.939
Varying-Coefficients	14.25	0.710	0.073	0.057	60694	0.430	0.863
Neural Network	22.49[3]	**0.940**	**0.015**	-0.024	63028	0.901	0.955
Personalized	**3.23**	0.911	0.022	**0.857**	**4822**	**0.923**	**0.975**

Predictive performance. For continuous response variables, we report correlation coefficient (R^2) and mean squared error (MSE) of the predictions. For classification tasks, we report area under the receiver operating characteristic curve (AUROC) and the accuracy (ACC). For the simulation, we also report recovery error of the true regression parameters.

Figure 3 Personalized models for patients in the training set of the cancer dataset. Each point represents a model for a single sample, colored by the patient ID. There is strong clustering according to patient label, but also intra-patient heterogeneity (notably Patients 1,3,4, and 6).

References

1. Hastie, T. and Tibshirani, R. Varying-coefficient models. Journal of the Royal Statistical Society. Series B (Methodological), pp. 757–796, 1993.
2. M. Al-Shedivat, A. Dubey, and E. P. Xing. Contextual explanation networks. arXiv preprint arXiv:1705.10301, 2017.
3. Kuijjer, M. L., Tung, M., Yuan, G., Quackenbush, J., and Glass, K. Estimating sample-specific regulatory networks. arXiv preprint arXiv:1505.06440, 2015.
4. Jabbari, F., Visweswaran, S., and Cooper, G. F. Instance-specific bayesian network structure learning. In Kratochvıl, V. and Studeny, M. (eds.), Proceedings of the Ninth International Conference on Probabilistic Graphical Models, volume 72 of Proceedings of Machine Learning Research, pp. 169–180, Prague, Czech Republic, 11–14 Sep 2018. PMLR.
5. Lengerich, Benjamin J., Bryon Aragam, and Eric P. Xing. "Personalized regression enables sample-specific pan-cancer analysis." Bioinformatics 34.13 (2018): i178-i186.

Every Sample a Task:
Pushing the Limits of Heterogeneous Models with Personalized Regression

Ben Lengerich[1,2], Bryon Aragam[2], Eric P. Xing[1,2,3]
{blengeri, naragam, epxing}@cs.cmu.edu
1) Machine Learning Department, Carnegie Mellon University
2) Computer Science Department, Carnegie Mellon University, 3) Petuum, Inc.

Guiding Question

Can **collections** of **simple** models outperform large models if each simple model is used for only a single sample?

Motivation

Typical tug-of-war:

Accuracy

Interpretability

What if this tradeoff is a byproduct of using population-level models to learn effects which actually differ between samples?

Can we instead learn *sample-specific models*? Would give us:

- A simple, interpretable model for each sample
- Representational capacity from the entire collection of models.

Let's use a multi-task framework, defining each training sample as a task, to share power and learn these *sample-specific* models.

Personalized Regression

Individual samples as tasks in a multi-task framework:

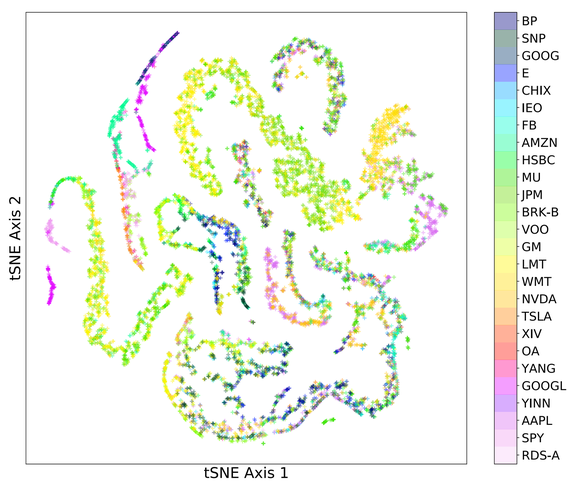
$$\mathcal{L}(X, Y, \boldsymbol{B}, d) = \sum_{i=1}^{n} l(X^i, Y^i, \beta^i) + \lambda\rho\left(\beta^i\right) + \gamma\varrho(\boldsymbol{B}, d)$$

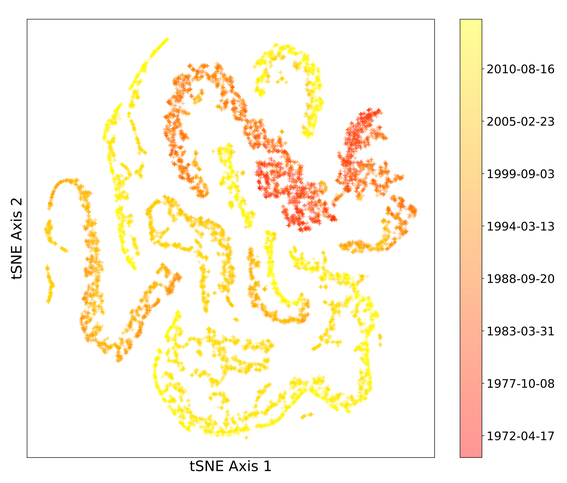
Prediction Loss | Regularizer | Distance-Matching Regularizer

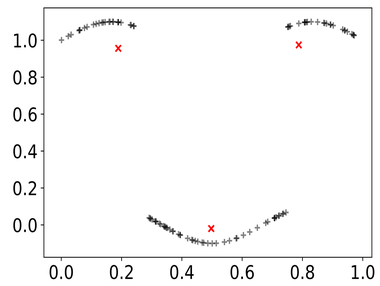
Distance-Matching Regularization

Match structure in *covariates* to structure in *parameters*:

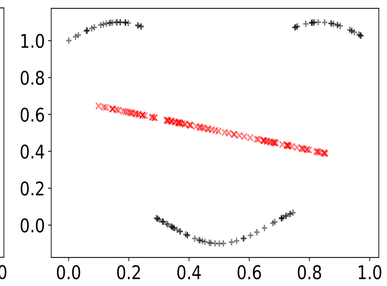
$$\varrho(\boldsymbol{B}, d) = \sum_{i=1}^{n}\sum_{j\neq i}\left(d\left(U^i, U^j\right) - \left\|\beta^i - \beta^j\right\|\right)^2$$



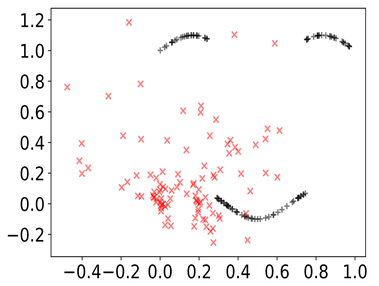




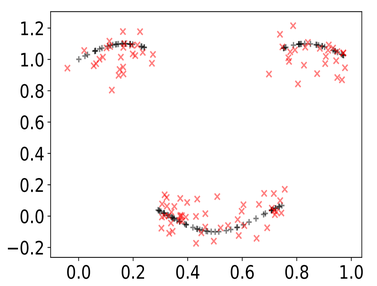
a) Mixture Model



b) Varying-Coefficients



c) DL+LIME



c) Personalized

The benefits of personalized models. Each point represents the regression parameters for a sample. **Black points** indicate true effect sizes, while the **red points** are estimates.

Mixture models (a) estimate a limited number of models. The varying-coefficients model (b) estimates sample-specific models but the non-linear structure of the true parameters violates the model assumptions, leading to a poor fit. The locally-linear models induced by a deep learning model (c) do not accurately recover the underlying effect sizes. In contrast, personalized regression (d) accurately recovers effect sizes.

Visualization of personalized models trained on a financial dataset (t-SNE projection). Each point represents a regression model for one security at a single date. There is strong clustering in models according to both industry (a) and time (b), but neither covariate would be sufficient to completely characterize each sample.