Learning Sample-Specific Models with Low-Rank Personalized Regression

Personalized regression enables sample-specific pan-cancer analysis

Benjamin Lengerich, Bryon Aragam, Eric P. Xing

Carnegie Mellon University, University of Chicago

Motivation

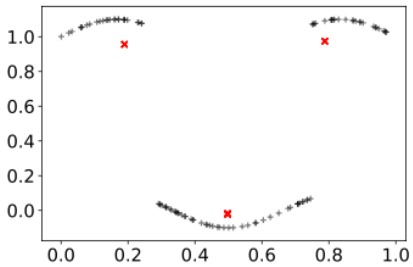
Adapting to heterogeneity in complex data to infer **individual-level effects**

No need for the model be complex: simple linear and logistic regression models will suffice

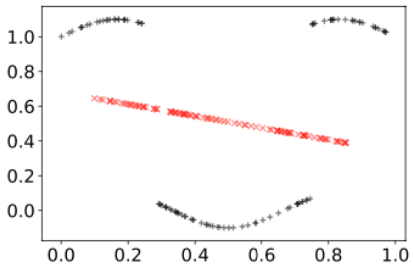
A tradeoff between effect complexity and effect personalization

Universal effect $\leftrightarrow$ Personalized effect

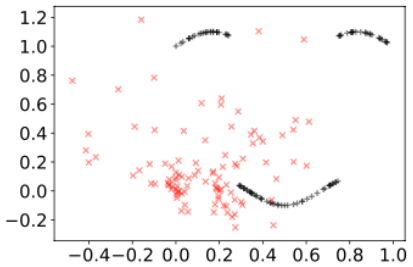
Complex model $\leftrightarrow$ Simple model



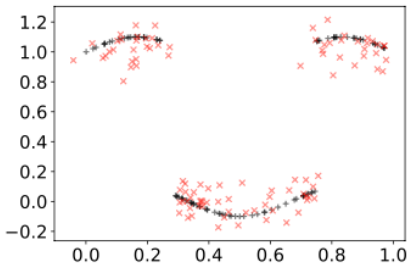
(a) Mixture model



(b) Varying-Coefficient



(c) Deep Neural Net



(d) Personalized

Traditional model v.s. Personalized model

$$\begin{aligned}
\text{n samples: } & (X^{(i)}, Y^{(i)}) \\
\text{Predictors: } & X^{(i)} \in \mathbb{R}^p \\
\text{Response: } & Y^{(i)} \\
\text{Traditional model: } & Y^{(i)} = f(X^{(i)}; \theta) \\
\text{Personalized model: } & Y^{(i)} = f(X^{(i)}; \theta^{(i)})
\end{aligned}$$

- Multiple simple models for each individual are estimated jointly with a single objective function
- Without additional constraints, this model is overparametrized
- Definitely over-fitted, but still fixable

Learning sample-specific models

$$Y^{(i)} = f(X^{(i)}, \theta^{(i)}) + w^{(i)} \qquad (1)$$

The key is to choose a solution $\theta^{(i)}$ that simultaneously leads to good generalization and accurate inferences about the i-th sample.

- a low-rank latent representation of $\theta^{(i)}$
- a novel regularization scheme.

Low-rank representation

Ideas of dictionary learning: good-behaved personalized modeling has a sparse representation

$$\text{Low rank: } \Omega = [\theta^{(1)}|...|\theta^{(n)}] \in \mathbb{R}^{p \times n}$$
$$i.e.\ \theta^{(i)} = Q^T Z^{(i)}$$
$$\text{Loadings: } Z^{(i)} \in \mathbb{R}^q; Z \in \mathbb{R}^{q \times n}$$
$$\text{Dictionary: } Q \in \mathbb{R}^{q \times p}$$
$$\text{Low rank: } \Omega = Q^T Z$$

Special form of sparse dictionary learning aka sparse coding

- Sparse coding is a representation learning method which aims at finding a sparse representation of the input data
- In the form of a linear combination of basic elements as well as those basic elements themselves
- These elements are called atoms and they compose a dictionary.

The choice of q is determined by the user's desired latent dimensionality;

for $q \ll p$, using only $\Theta(q(n+p))$ instead of $\Theta(np)$ of a full-rank solution can improve computational and statistical efficiency.

With $\theta^{(i)} = Q^T Z^{(i)}$, lower rank formulation enables to use L_2 distance in Z to restrict Euclidean distances between the $\theta^{(i)}$

$$\|\theta^{(i)} - \theta^{(j)}\| \leq \sqrt{p}\|Z^{(i)} - Z^{(j)}\| \qquad (2)$$

- Sparsity in θ can be realized by sparsity in Z, Q
- The low-rank formulation constrains the number of personalized sparsity patterns, by changing the latent dimensionality q.

Distance-matching

Regularize the parameters $\theta^{(i)}$ by requiring that similarity in θ corresponds to similarity in U,

- i.e. $\|\theta^{(i)} - \theta^{(j)}\| \approx \rho(U^{(i)}, U^{(i)})$
- The $U(i)$ represent exogenous variables that we do not wish to directly model.

Distance-matching regularization (DMR)

Adapt a distance-matching regularization (DMR) scheme to penalize the squared difference in implied distances.

The covariate distances are modeled as a weighted sum.

$$\rho_\phi(u, v) = \sum_{l=1}^{k} \phi_l d_l(u_l, v_l), \ \phi_l \geq 0 \qquad (3)$$

Each $d_l (l = 1, \ldots, k)$ is a metric for a covariate;

ϕ is a positive (non-negative) vector represents a linear transformation into latent distance function;

Distance Matching Regularized Loss

Main idea: Distance between sample parameters should be similar to distance between sample covariates, as well as the distance between latent variable U

In order for $\|\theta^{(i)} - \theta^{(j)}\| \approx \rho_\phi(U^{(i)}, U^{(j)})$, it is suffices to require $\|Z^{(i)} - Z^{(j)}\| \approx \rho_\phi(U^{(i)}, U^{(j)})$

Then we can define the **distance-matching regularizer**:

$$D_\gamma^{(i)}(d_\beta, d_U) = \frac{\gamma}{2} \sum_{j \in B_r(i)} \left(\rho_\phi(U^{(i)}, U^{(j)}) - \|Z^{(i)}, Z^{(j)}\|_2 \right)^2 \tag{4}$$

$B_r(i) = \{j : \|Z^{(i)} - Z^{(j)}\|^2 < r\}$

The hyperparameter γ trades off sensitivity to prediction of the response variable against sensitivity to covariate structure

For example for simple linear relationship both under L_2 norm

$$D_\gamma^{(i)}(d_\beta, d_U) = \frac{\gamma}{2} \sum_{j \neq i} \left(d_\theta(\theta^{(i)}, \theta^{(j)}) - d_Z(Z^{(i)}, Z^{(j)}) \right)^2 = \frac{\gamma}{2} \sum_{j \neq i} \left(\|\theta^{(i)}, \theta^{(j)}\|_2^2 - \|Z^{(i)}, Z^{(j)}\|_2^2 \right)^2$$

Personalized Regression

Seeking a model for inference, not necessarily for accurate predictive models

Seeking relatively simple personalized effects not universal effects

covariate data as informative of each sample

Here is the $\ell(x, y, \theta)$ is the loss function we want to minimize

$$\mathcal{L}^{(i)}(Z, Q, \phi) = \ell\left(X^{(i)}, Y^{(i)}, Q^T Z^{(i)}\right) + \psi_\lambda(Q^T Z^{(i)}) + D_\gamma^{(i)}(Z, \phi) \quad (5)$$

- where ψ_λ is a regularization such as L_1 penalty
- $D_\gamma^{(i)}$ is the distance-matching regularization defined in *Eq.*(4)

Where we learn Ω and ϕ by minimizing the following objective:

$$\mathcal{L}(Z, Q, \phi) = \sum_{i=1}^{n} \mathcal{L}^{(i)}(Z, Q, \phi) + v\|\phi - 1\|_2^2 \qquad (6)$$

- where $v\|\phi - 1\|_2^2$ regularize the distance function ρ_ϕ with strength set v
- again, $\Omega = Q^T Z$

Algorithm

- The objective function is optimized with sub-gradient descent.
- Initialize Σ by setting $\theta^{(i)}$ $N(\hat{\theta}, \varepsilon I)$ for population model such as lasso and elastic net.
- Initialize Z and Q by PCA factorization.
- Each personalized estimator is endowed with a personalized learning rate $\alpha_t^{(i)} = \alpha_t / ||\hat{\theta}_t^{(i)} - \hat{\theta}^{(pop)}||_\infty$.
- This learning rate scales the global learning rate α_t according to how far the estimator has traveled.
- This scheme ensures that the personalized coefficients' center of mass stays close to the initial $\hat{\theta}^{(pop)}$ despite unconstrained $\theta^{(i)}$.

Algorithm

Algorithm 1 Personalized Estimation

Require: $\widehat{\theta}^{pop}, \lambda, \gamma, \upsilon, \alpha, c$

$\theta^{(1)}, \ldots, \theta^{(n)} \leftarrow \widehat{\theta}^{pop}$

$\Omega \leftarrow [\theta^{(1)} | \ldots | \theta^{(n)}]$

$Z, Q \leftarrow \text{PCA}(\Omega)$

$\phi \leftarrow \mathbf{1}$

$\alpha \leftarrow \alpha_0$

do

$\widetilde{Z}, \widetilde{Q}, \widetilde{\phi} \leftarrow Z, Q, \phi$

$\phi \leftarrow \phi - \alpha \frac{\partial}{\partial \phi} \mathcal{L}(\widetilde{Z}, \widetilde{Q}, \widetilde{\phi}; \lambda, \gamma, \upsilon)$

$Z^{(i)} \leftarrow Z^{(i)} - \frac{\alpha}{\|\theta^{(i)} - \widehat{\theta}^{\text{pop}}\|_\infty} \Big[\frac{\partial}{\partial Z^{(i)}} \sum_{i=1}^n D_\gamma^{(i)}(\widetilde{Z}, \widetilde{\phi}) +$
$\widetilde{Q}\big(\partial \ell(X^{(i)}, Y^{(i)}, \theta^{(i)}) + \partial \psi_\lambda(\theta^{(i)})\big)\Big] \quad \forall\, i \in [1, \ldots, n]$

$Q \leftarrow Q - \alpha \Big[\frac{\partial}{\partial Q} \sum_{i=1}^n D_\gamma^{(i)}(\widetilde{Z}, \widetilde{\phi}) + \sum_{i=1}^n \widetilde{Z}^{(i)} \big(\partial \ell(X^{(i)}, Y^{(i)}, \theta^{(i)})^T + \partial \psi_\lambda(\theta^{(i)})^T\big)\Big]$

$\alpha \leftarrow \alpha c$

$\theta^{(i)} \leftarrow Q^T Z^{(i)} \quad \forall\, i \in [1, \ldots, n]$

$\Omega \leftarrow [\theta^{(1)} | \ldots | \theta^{(n)}]$

while not converged

return Ω, Z, Q, ϕ

Prediction

- Given a test point (X, U), averaging the model parameters of the k_n nearest training points based on distance ρ_ϕ
- Increasing k_n drives the test models toward the population model to control overfitting
- Intentionally avoided using X to select θ so that interpretation of θ is not confounded by X

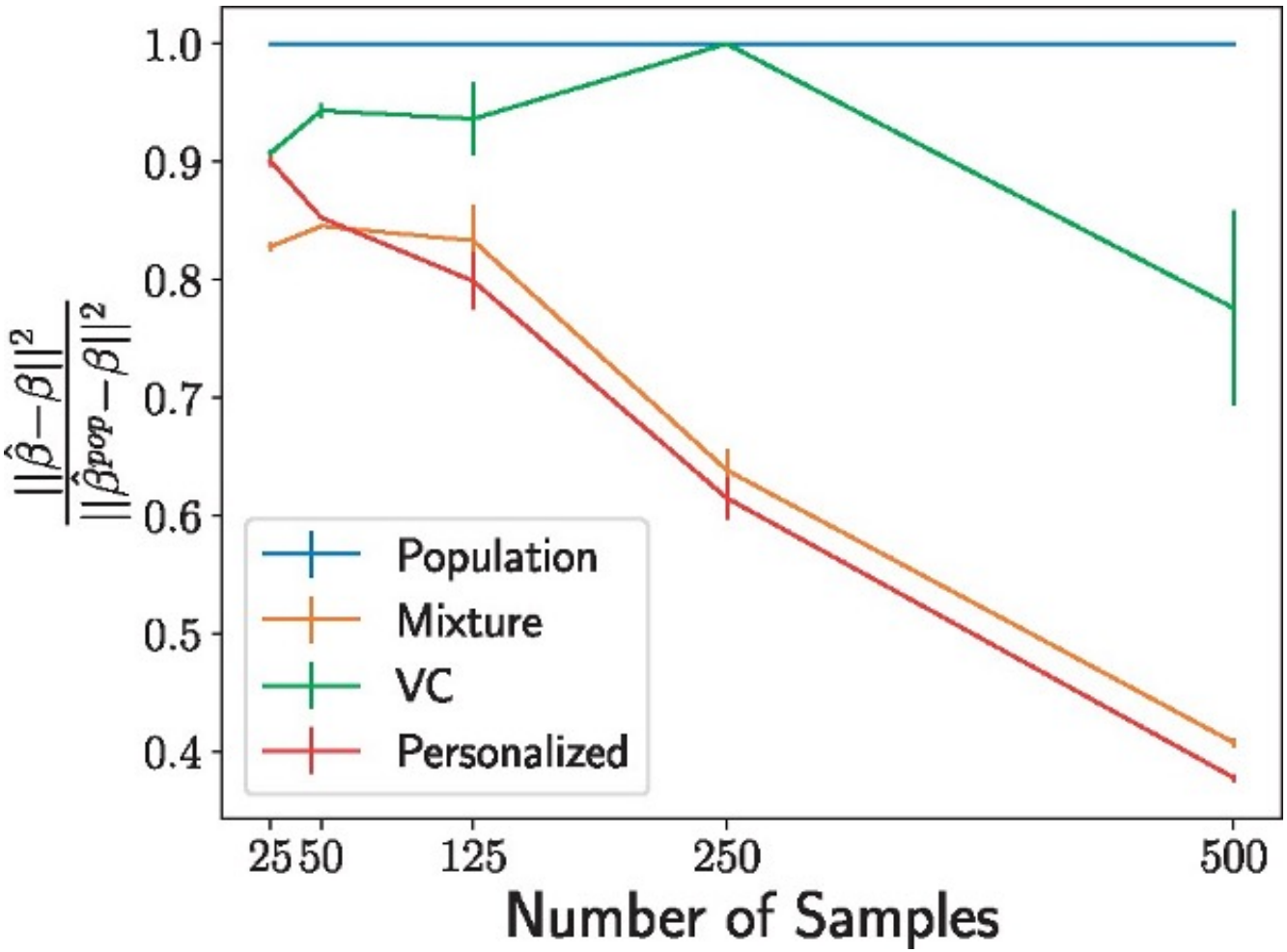
$$\theta = \frac{1}{k_n} \sum_{j=1}^{k_n} \theta^{(\eta(\rho_\phi, U))[j]}; \;\; \eta(\rho_\phi, U) = \mathit{argsort}_{a \leq i \leq n} \rho_\phi(U, U^{(i)}) \qquad (7.a)$$

Results

- Performance of personalized regression will be tested through simulation study.
- Fix $X \in \mathbb{R}^{N \times P}$, generate sample-specific $\beta^{(i)}$ Unif (0,1), and $Y^{(i)}) \in (0, 1)$.
- Covariates $U^{(i)}$ are generated by projecting $\beta^{(j)}$ into $K < P$ dimension with mult-dimensional scaling.
- covariates that are related to the personalized regression coefficients in a highly nonlinear, nonparametric manner.
- Set $K = 10, K = 3$.

Model	Train error (%)	Test error (%)
Population	6.9	6.8
Tissue-population	6.5	6.8
Mixture	6.7	6.8
VC	7.5	8.7
LMM	7.0	7.1
Personalized	**6.3**	**6.7**

Simulation

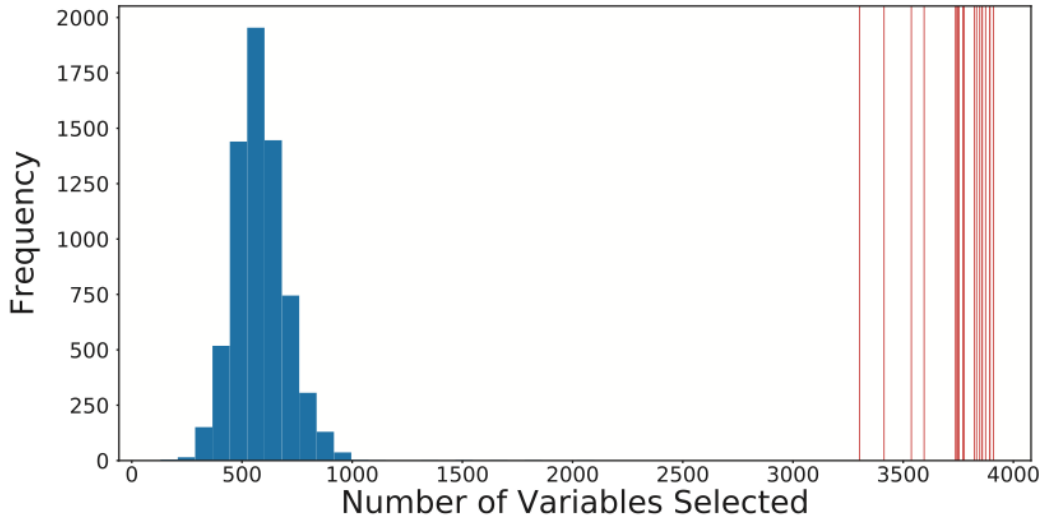

1.0
0.9
0.8
0.7
0.6
0.5
0.4
$\frac{||\hat{\beta}-\beta||^2}{||\hat{\beta}^{pop}-\beta||^2}$
Population
Mixture
VC
Personalized
25 50
125
250
500
Number of Samples

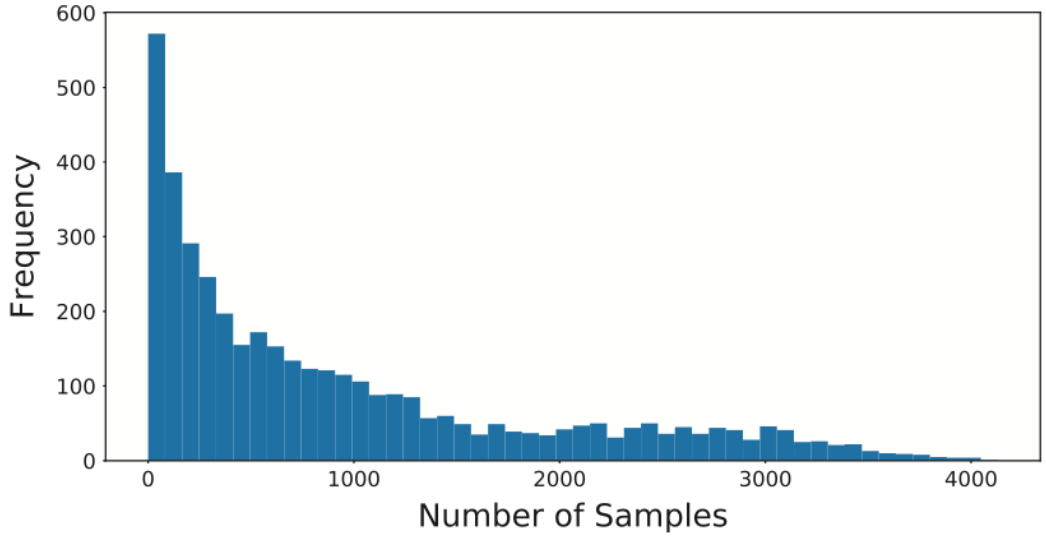
Sample-specific pan-cancer analysis

They also use gene expression (RNA-Seq) quantification data from The Cancer Genome Atlas (TCGA).

- This dataset compiles data from 37 projects spanning 36 disease types in 28 primary sites.
- After pruning for missing values, this dataset contains 9663 profiles for 8944 case and 719 matched control samples
- This resulting in $P = 4123$ features when an intercept term is added.
- They divide this set into 75% training data and 25% testing.

PR has good variable selections



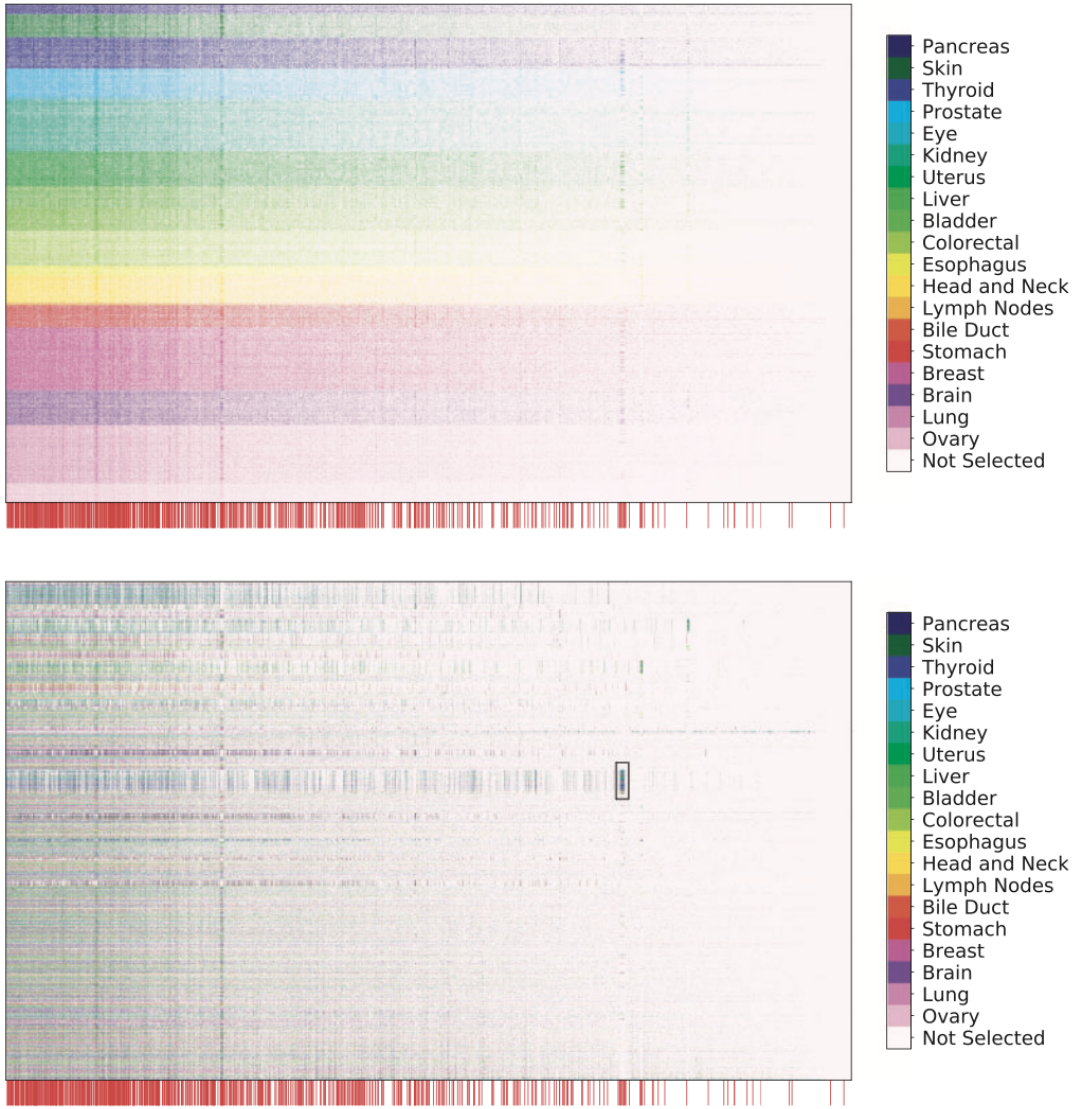


PR has good variable selections

Table 3. Enrichment analysis of complete variable rankings

Model	Biological process	*P*-value
Population	mRNA processing	2.06e-8
	DNA metabolic process	3.18e-6
	Organelle organization	3.86e-2
Tissue-Population	mRNA processing	3.09e-9
	Metabolic process	3.26e-5
	Transcription, DNA-dependent	9.61e-5
	DNA metabolic process	5.9e-3
Mixture	mRNA processing	1.45e-8
	DNA Metabolic process	1.96e-5
	Transcription, DNA-dependent	2.62e-4
	Organelle organization	7.32e-3
VC	None	NA
LMM	DNA metabolic process	2.02e-2
Personalized	mRNA processing	5.83e-6
	Metabolic process	1.1e-3
	DNA metabolic process	3.15e-2

Here is the result of RNA-seq data



Cancer is a highly personalized problem

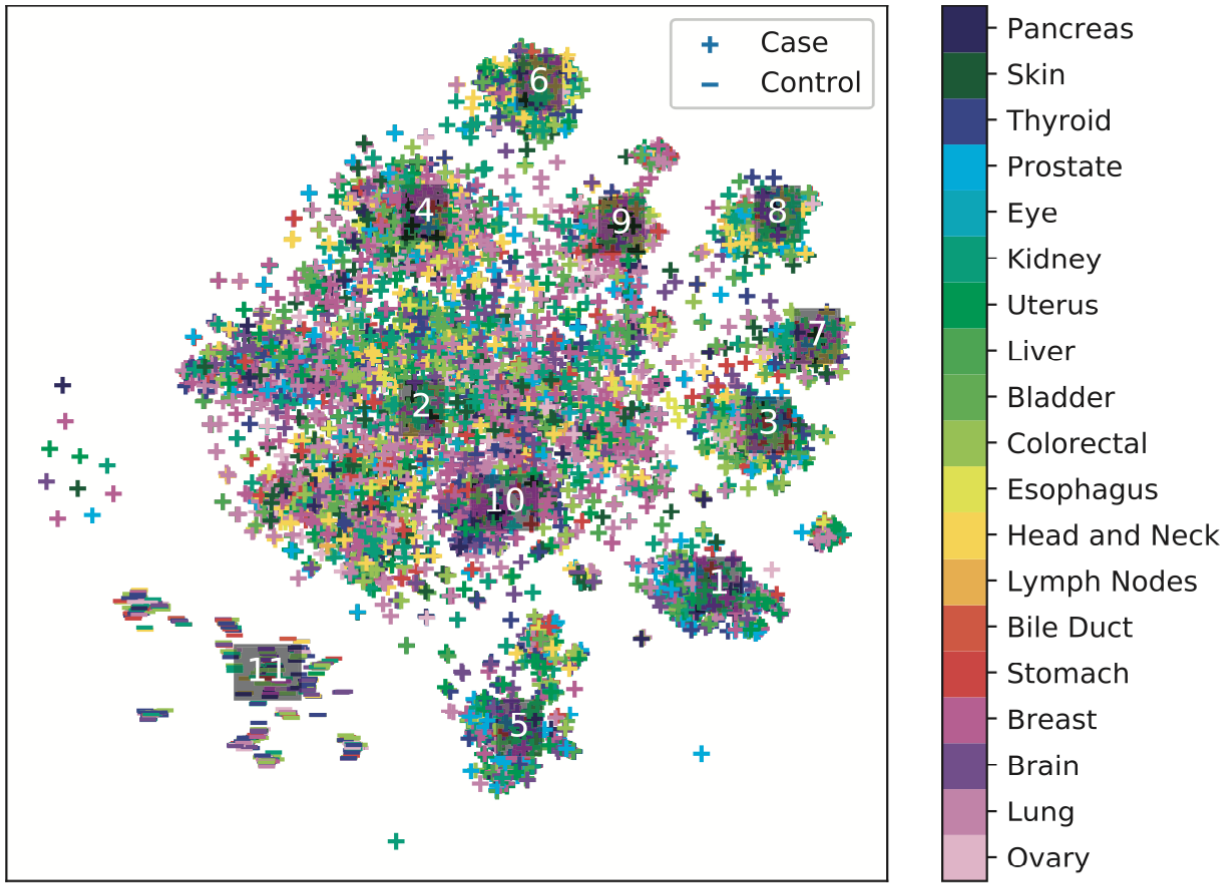


Fig. 6. tSNE projection of personalized regression parameters learned from a pan-cancer dataset. Each point represents a single sample with color indicating primary tumor site and marker type indicating case/control status of the patient. Labelled points indicate the centroids of clusters analyzed in Table 4

The clusters have biological and clinical meanings

Table 4. Enrichment analysis of tumor clusters

Cluster	Biological process	*P*-value
1	Symbiont process	2.62e-3
	Regulation of cellular catabolic process	1.96e-2
	Protein modification process	3.43e-2
2	DNA repair	3.21e-12
	RNA splicing, via transesterification Reactions with bulged adenosine as nucleophile	3.64e-7
	DNA replication	1.00e-6
3	Symbiont process	1.4e-3
	Antigen processing and presentation of peptide antigen	1.06e-2
	Antigen processing and presentation of exogenous antigen	1.08e-2
4	DNA metabolic process	3.83e-8
	DNA repair	1.68e-6
	Regulation of cellular macromolecule biosynthetic process	5.06e-6
5	Plasma membrane bounded cell projection morphogenesis	1.45e-2
	Neuron projection development	3.02e-2
6	mRNA catabolic process	8.78e-4
	Gene expression	6.02e-4
	Macromolecule biosynthetic process	3.32e-2
7	None	N/A
8	Generation of precursor metabolites and energy	4.75e-5
	Oxidation-reduction process	4.52e-5
	Citrate metabolic process	9.84e-3
9	DNA metabolic process	3.96e-10
	Cellular response to DNA damage stimulus	5.57e-9
	Protein complex subunit organization	1.41e-4
10	DNA metabolic process	7.15e-8
	ncRNA metabolic process	1.33e-4
	Chromatin organization	8.27e-4

Discussion

- The Generalization Problem? Sensitive to outliers?
- Is it really interpreting as its claimed to be? Confounding and Collinearity?
- Tuning the parameters $\lambda, \gamma . v, c, \alpha$?
- Why PCA? How Kernel PCA? ?