SUQUAN: Supervised quantile normalisation

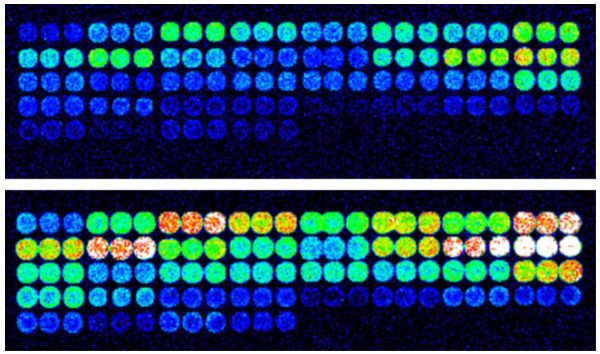
Marine Le Morvan
Joint work with **Jean-Philippe Vert**

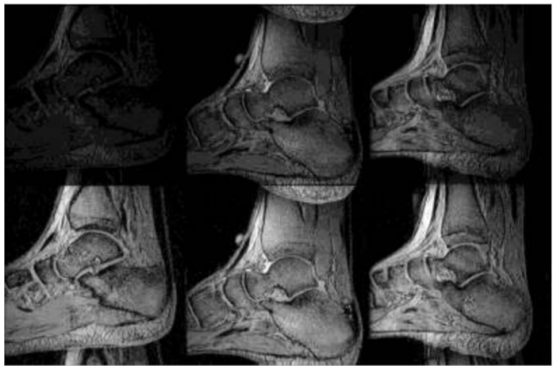
CBIO - Mines Paristech, INSERM U900 - Curie institute, Paris, France

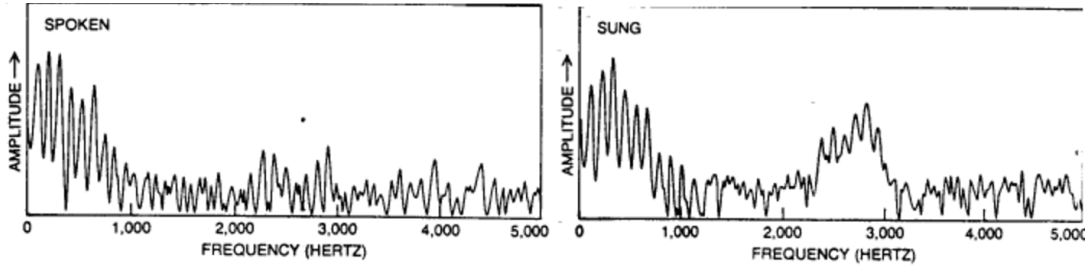
May 2nd, 2017

How do we deal with technical variability?

- Data acquisition is often plagued with various sources of perturbations which induce unwanted variations.







  - ✓ Gene expression microarrays, RNAseq, Genotyping arrays, DNA methylation, ChIP-Sequencing, Brain imaging, Photos, Speech...

- Need to remove technical variability from noisy data.

Quantile normalisation

Quantile normalization monotonically modifies the entries of a given sample so that after normalization, all samples have the same distribution of entries.

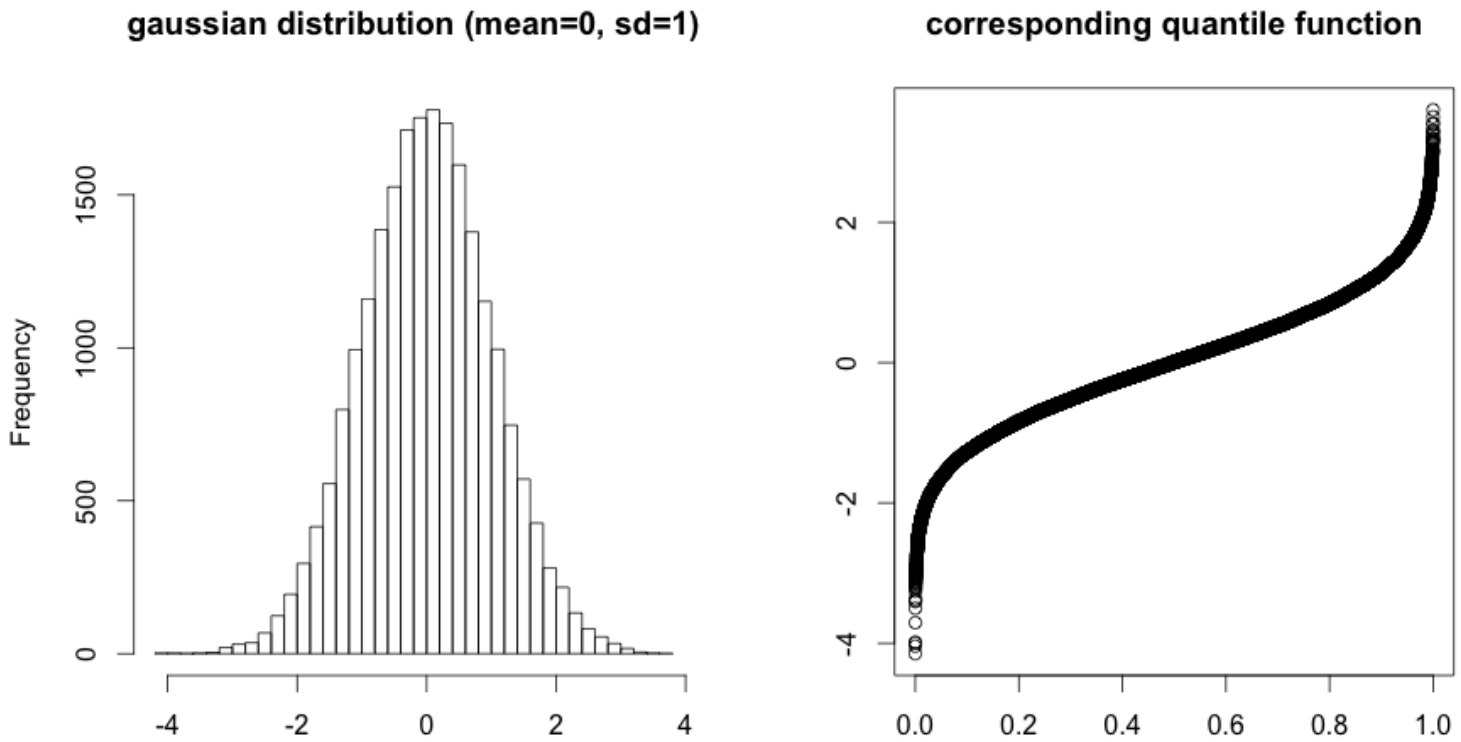
Standard full quantile normalization

Quantile normalisation in practice:

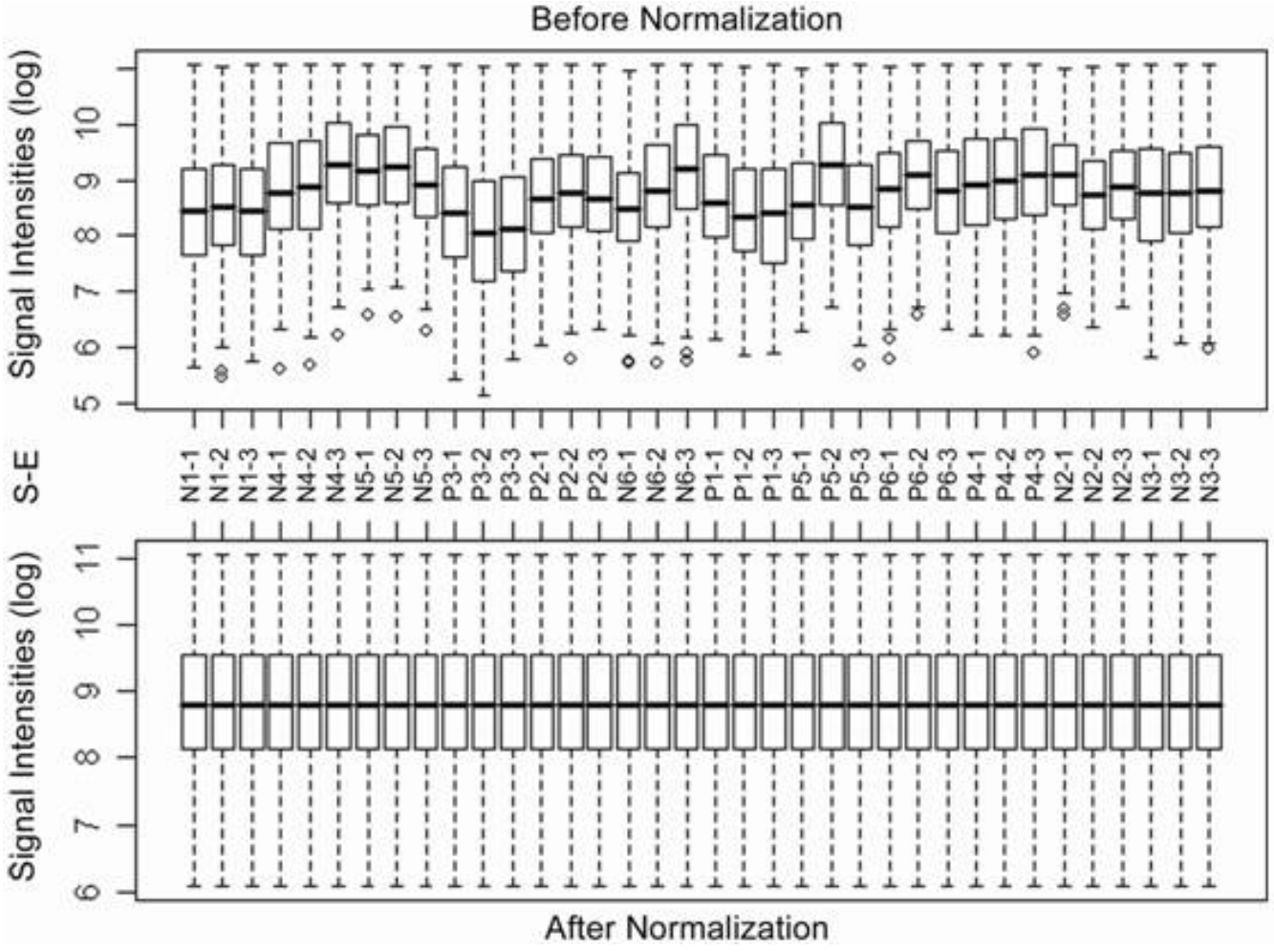
- ✓ Define a target quantile function (equivalently a target distribution)

$$f = (f_1, f_2, ..., f_p) \quad \text{such that} \quad f_1 \leq f_2 \leq ... \leq f_p$$

- ✓ Set the smallest entry of each sample to f_1
- ✓ Set the second smallest entry of each sample to f_2
- ✓ ...
- ✓ Set the largest entry of each sample to f_p



Standard full quantile normalization

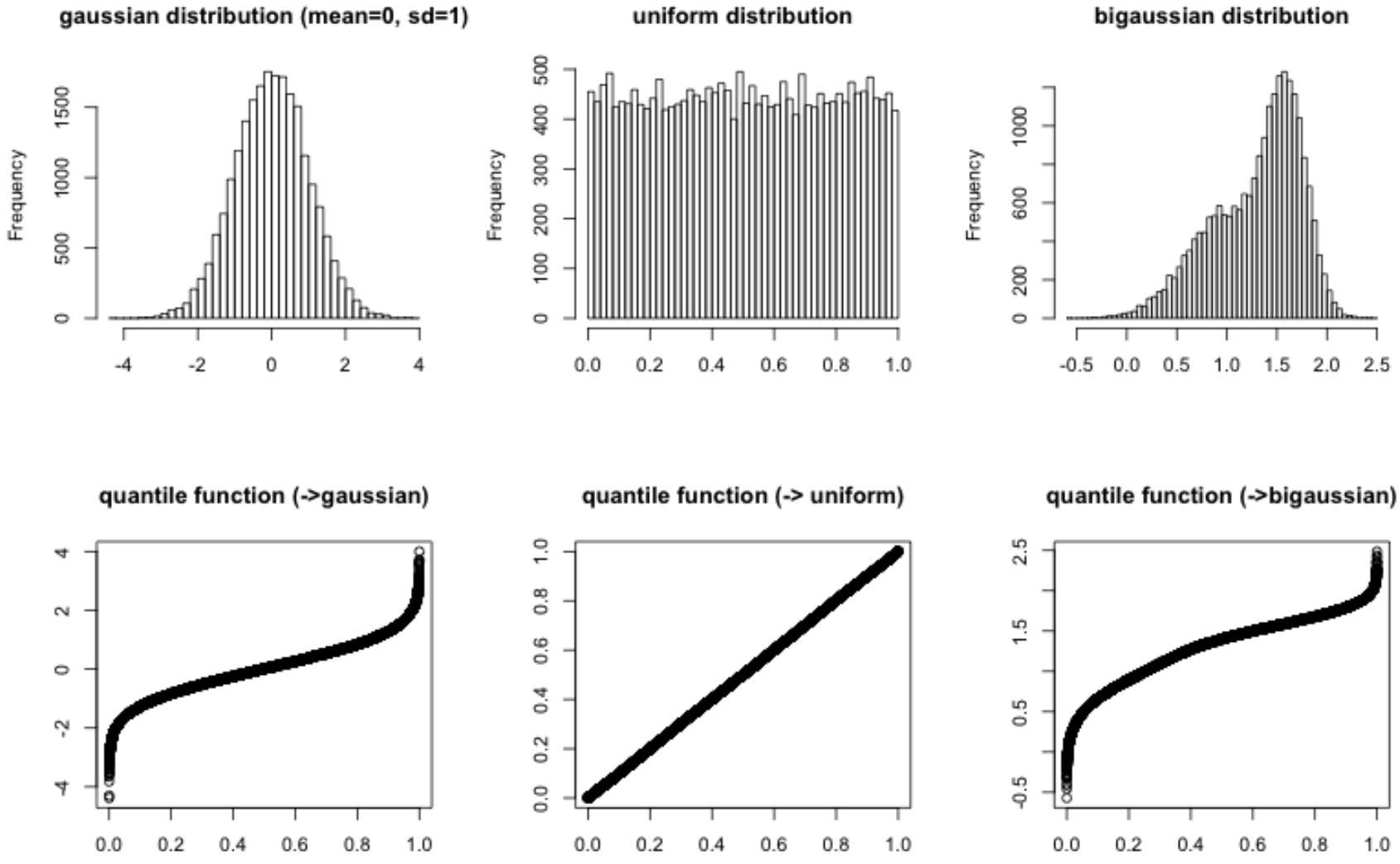


We believe the ”true” signal should have the same distribution but is perturbed by ”unwanted variations”.

Standard full quantile normalization

$\Rightarrow$ QN suffers from a practical question:

How to choose the target distribution?



Standard full quantile normalization

$\Rightarrow$ QN suffers from a practical question:

How to choose the target distribution?

- In biology, the target distribution was empirically chosen as the median of the empirical distribution of the samples (obtained by taking the median of each k^{th} order statistic across samples).

- Quantile normalization was originally developed for gene expression microarrays (Bolstad et al., 2003):
  - ✓ *While there might be some advantages to using a common, non-data driven, distribution with the quantile method, it seems unlikely an agreed standard could be reached. [...]. For this reason we prefer the minimalist approach of a data based normalization.*

Learning the target distribution

- ✓ $x_1, \ldots, x_n$ a set of p-dimensional samples
- ✓ $\mathcal{F} \subset \mathbb{R}^p$ the set of target functions
- ✓ $f \in \mathcal{F}$ a target function
- ✓ For $x \in \mathbb{R}^p$, let $\Phi_f(x) \in \mathbb{R}^p$ be the data after QN with target distribution f

- Standard approaches (NSQN, NetNorM, ...)
  1. Fix f arbitrarily
  2. QN all samples to get $\Phi_f(x_1), \ldots, \Phi_f(x_n)$
  3. Learn a generalized linear model (w, b) on normalized data:

$$\min_{w,b} \frac{1}{n} \sum_{i=1}^{n} \ell_i \left(w^\top \Phi_f(x_i) + b \right) + \lambda \Omega(w)$$

- SUQUAN: jointly learn f and (w, b):

$$\min_{w,b,f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \ell_i \left(w^\top \Phi_f(x_i) + b \right) + \lambda \Omega(w) + \gamma \Omega(f)$$

Learning the target distribution

- For $x \in \mathbb{R}^p$, let $\Pi_x \in \mathbb{R}^{p \times p}$ the permutation matrix of x's entries

$$x = \begin{pmatrix} 4.5 \\ 1.2 \\ 10.1 \\ 8.9 \end{pmatrix} \quad \Pi_x = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix} \quad f = \begin{pmatrix} 0 \\ 1 \\ 3 \\ 4 \end{pmatrix}$$

- Quantile normalized x with target distribution f is:

$$\Phi_f(x) = \Pi_x f$$

- SUQUAN solves

$$\min_{w,b,f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \ell\left(w^\top \Pi_{x_i} f + b\right) + \lambda\Omega(w) + \gamma\Omega(f)$$

$$= \min_{w,b,f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \ell\left(< wf^\top, \Pi_{x_i} > + b\right) + \lambda\Omega(w) + \gamma\Omega(f)$$

- A particular rank-1 matrix optimization, x is replaced by Π_x

Three variants of SUQUAN

SUQUAN solves:

$$\min_{w,b,f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \ell\left(< wf^\top, \Pi_{x_i} > +b\right) + \lambda\Omega(w) + \gamma\Omega(f)$$

We consider three sets of candidate target functions $\mathcal{F}$:

- the set of bounded target functions:

$$\mathcal{F}_0 = \left\{ f \in \mathbb{R}^p \ : \ \frac{1}{p}\sum_{i=1}^{p} f_i^2 \leq 1 \right\} .$$

  A caveat with $\mathcal{F}_0$ is that the target function may not be non-decreasing.

- the set of bounded non-decreasing target functions

$$\mathcal{F}_{\text{BND}} = \mathcal{F}_0 \cap \mathcal{I} \quad \text{where} \quad \mathcal{I} = \left\{ f \in \mathbb{R}^p \ : \ f_1 \leq f_2 \leq \ldots \leq f_p \right\} .$$

- the set of non-decreasing and smooth target functions

$$\mathcal{F}_{\text{SPAV}} = \left\{ f \in \mathcal{I} \ : \ \sum_{j=1}^{p-1} (f_{j+1} - f_j)^2 \leq 1 \right\} .$$

Algorithms

We propose SUQUAN-SVD as an efficient method to approximately solve SUQUAN when $\mathcal{F} = \mathcal{F}_0$ and $\Omega(w) = \|w\|_2^2$.

Algorithm 1 SUQAN-SVD

Input: $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^p \times \{-1, 1\}$

Output: $f \in \mathcal{F}_0$ target quantile

1: $M \leftarrow -\sum_{i=1}^n y_i \Pi_{x_i}$

2: $(\sigma, w, f) \leftarrow SVD(M, 1)$

Let $\mathcal{S}(f, w, b) = \frac{1}{n}\sum_{i=1}^n \ell\left(< wf^\top, \Pi_{x_i} > + b\right)$. The first-order Taylor expansion of $\mathcal{S}(f, w, 0)$ at the origin is:

If ℓ is the logistic loss:

$$\mathcal{S}(f, w, 0) \approx \frac{1}{2} - \frac{1}{n}\sum_{i=1}^n y_i w^T \Pi_{x_i} f$$

If ℓ is the square loss:

$$\mathcal{S}(f, w, 0) \approx 1 - \frac{2}{n}\sum_{i=1}^n y_i w^T \Pi_{x_i} f$$

Under the constraints $\|f\|_2 = 1$ and $\|w\|_2 = 1$, the first left and right singular vectors of M minimize the first-order Taylor expansion of $\mathcal{S}(f, w, 0)$.

Algorithms

SUQUAN-BND and SUQUAN-SPAV approximately solve SUQUAN when $\mathcal{F} = \mathcal{F}_{BND}$ and $\mathcal{F} = \mathcal{F}_{SPAV}$ respectively using an alternate optimisation scheme in w and f.

Algorithm 2 SUQAN-BND and SUQUAN-SPAV

Input: $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^p \times \{-1, 1\}, f_{init} \in \mathcal{I}$, $\lambda \in \mathbb{R}$

Output: $f \in \mathcal{I}$ target quantile

1: **for** $i = 1$ to n **do**

2: $rank_i, order_i \leftarrow \text{sort}(x_i)$

3: **end for**

4: $w, b \leftarrow \underset{w,b}{\text{argmin}} \frac{1}{n} \sum_{i=1}^{n} \ell_i \left(w^\top f[rank_i] + b\right) + \lambda ||w||^2$

(standard linear model optimisation)

5: $f \leftarrow \underset{f \in \mathcal{F}_{BND}}{\text{argmin}} \frac{1}{n} \sum_{i=1}^{n} \ell_i \left(f^\top w[order_i] + b\right)$

(isotonic optimisation problem using PAVA as prox)

OR

$f \leftarrow \underset{f \in \mathcal{F}_{SPAV}}{\text{argmin}} \frac{1}{n} \sum_{i=1}^{n} \ell_i \left(f^\top w[order_i] + b\right)$

(smoothed isotonic optimisation problem using SPAV as prox)

Simulations

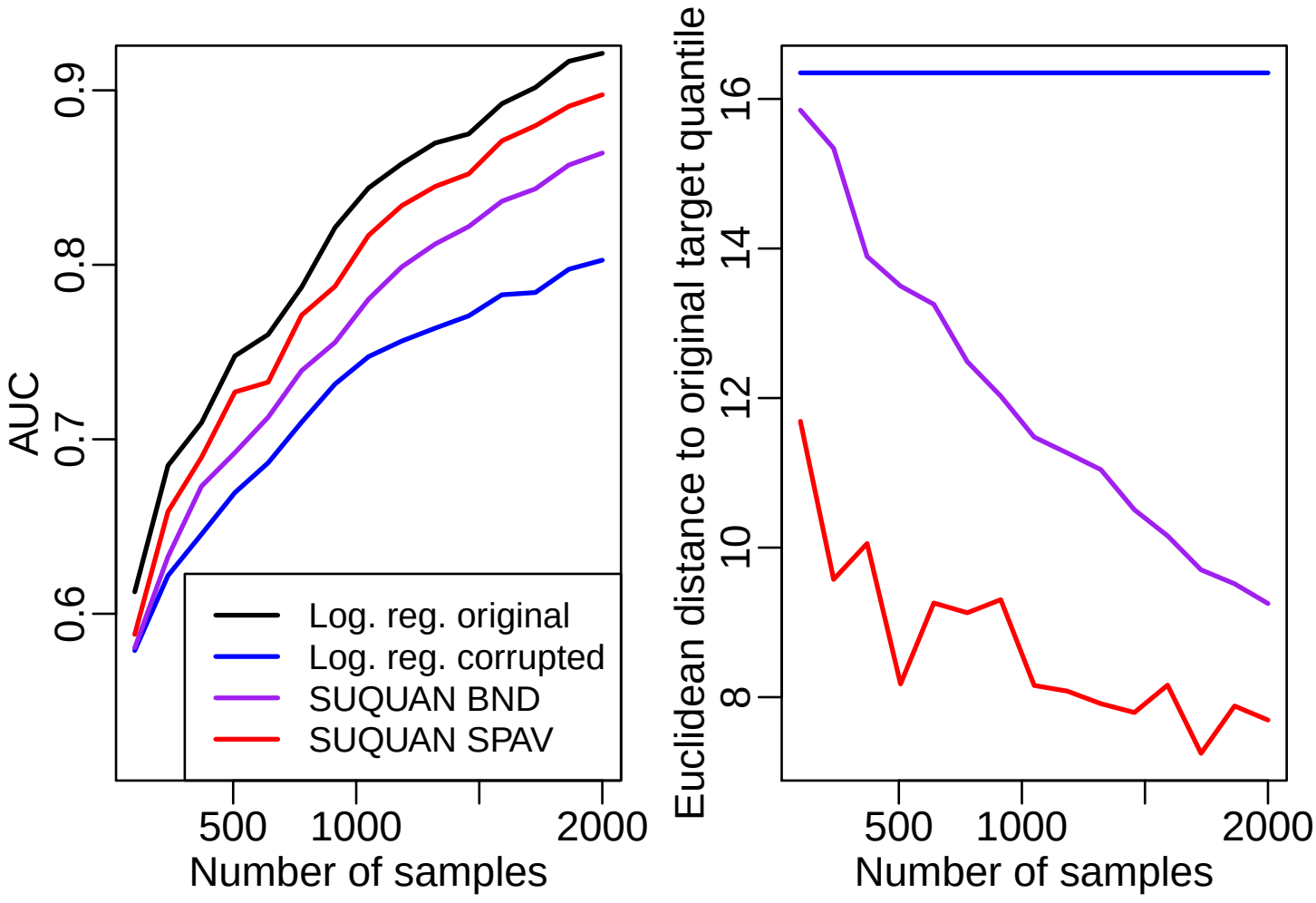
1. Fix $f \in \mathcal{F}$ to be the quantile function of the normal distribution.
2. Randomly sample $w \in \mathbb{R}^p$ from a multivariate normal distribution.
3. Simulate $(\Phi_f(X), Y) \in \mathbb{R}^p \times \{-1, 1\}$ pairs according to the model

$$P(Y = 1 \mid X = \Phi_f(X)) = \frac{1}{1 + \exp(-w^\top \Phi_f(X))},$$

   where $\Phi_f(X)$ is a random shuffling of the entries of f.
4. Estimate w from n observations:
   - Ridge logistic regression on the correct data $(\Phi_f(X_i), Y_i)_{i=1,\ldots,n}$.
   - Ridge logistic regression on the corrupted data $(\Phi_g(X_i), Y_i)_{i=1,\ldots,n}$, where g is a corrupted quantile function.
   - SUQUAN-BND and SUQUAN-SPAV on the corrupted data $(\Phi_g(X_i), Y_i)_{i=1,\ldots,n}$.
5. Assess the model on an independently generated test set of 1000 samples.

Simulations

$$p = 1000,\ 100 \leq n \leq 2000$$



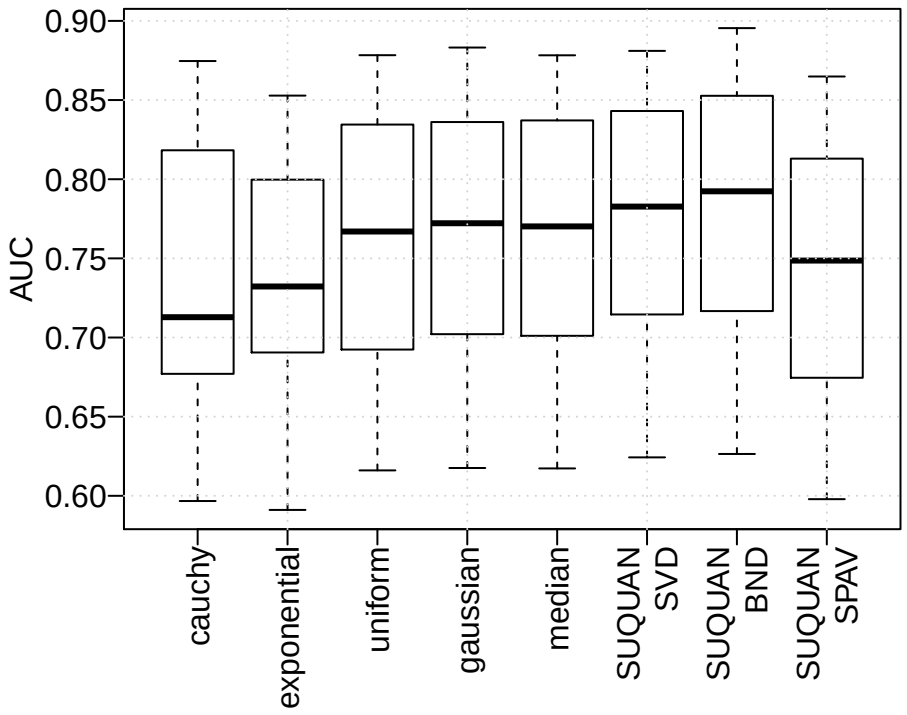
Real data experiments - CIFAR10

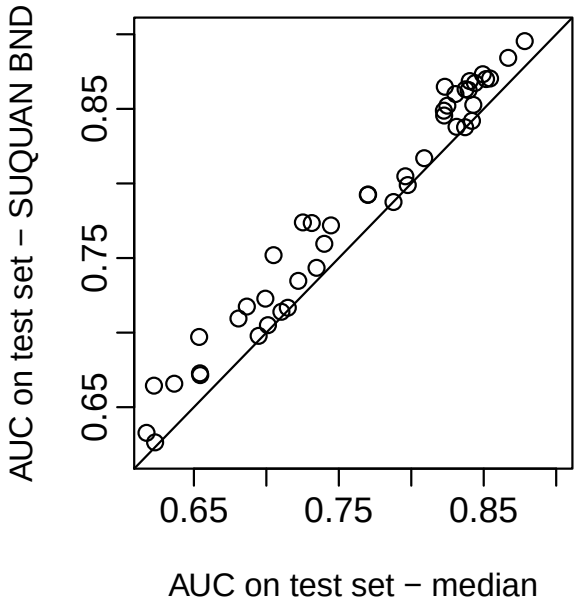
- Proof of concept on an image classification task
- 32 $\times$ 32 tiny color images from 10 different classes.

  $\Rightarrow$ 45 binary classification tasks.

  - ✓ 10 000 training images + 2000 test images per task
  - ✓ Images were converted to grayscale and transformed into a feature vector of length 1024.

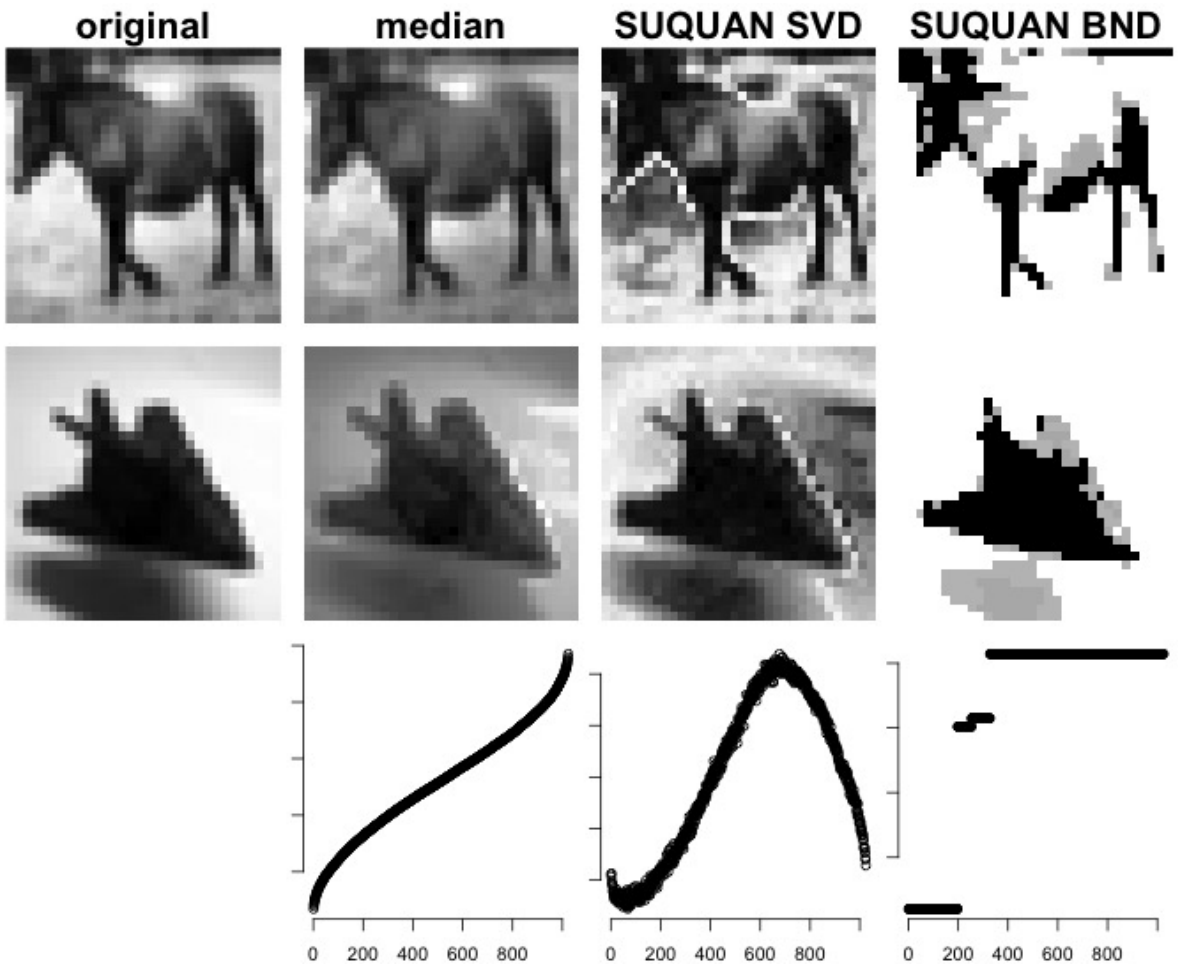
Real data experiments - CIAFR10





Real data experiments - CIAFR10

- Target quantiles for the ’airplane’ versus ’horse’ binary classification task.



Real data experiments - Breast cancer gene expression

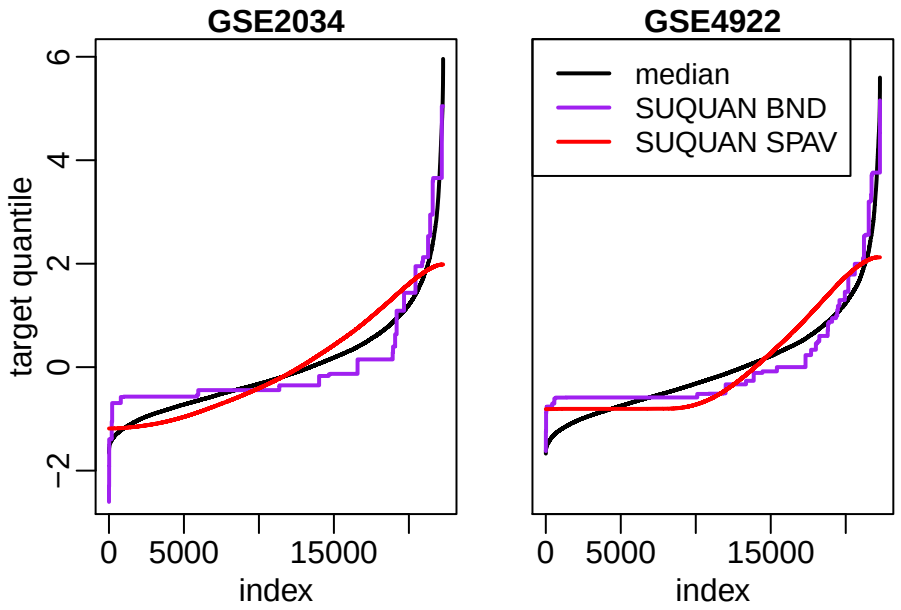
- Breast cancer prognosis from gene expression data.
- Two classes of patients: those who relapsed within 6 years of diagnosis and those who did not.

Dataset name	# genes	# patients	# positives	% positives
GSE4922	22283	225	73	0.32
GSE2990	22283	106	32	0.30
GSE2034	22283	271	104	0.38
GSE1456	22283	141	37	0.26

$\Rightarrow$ Binary classification task.

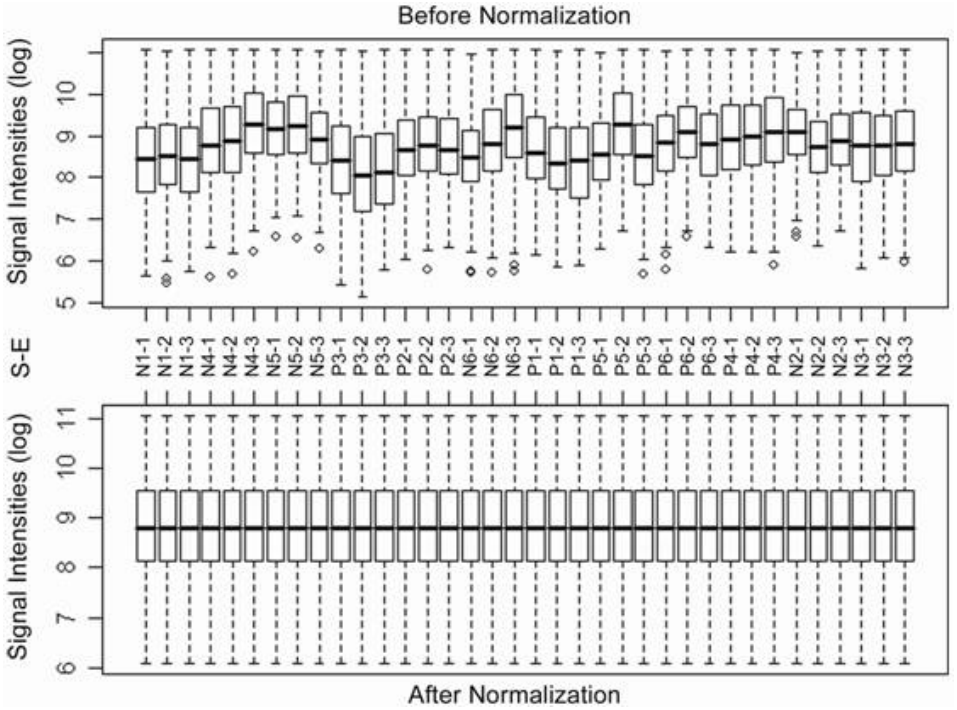
Real data experiments - Breast cancer gene expression

	LOGISTIC REGRESSION							SUQUAN		
	RAW	RMA	CAUCHY	EXP.	UNIFORM	GAUSSIAN	MEDIAN	SVD	BND	SPAV
GSE1456	65.94	68.73	59.56	68.86	68.72	69.00	69.06	57.60	71.44	69.60
GSE2034	74.52	75.42	61.91	74.53	75.22	76.45	74.92	52.61	70.50	76.11
GSE2990	57.01	60.43	54.72	61.25	56.25	58.66	59.72	52.51	59.22	59.94
GSE4922	58.52	58.86	55.24	58.81	55.66	60.01	59.18	52.39	61.82	61.41
AVERAGE	64.00	65.86	57.86	65.86	63.96	66.03	65.72	53.78	65.75	**66.77**



Example of target quantiles learned for two gene expression datasets and an arbitrary split in train/test sets.

SUQUAN summary



- The target distribution of QN can be seen as a parameter to optimize.
- SUQUAN boils down to
  - Represent each sample x by the permutation matrix Π_x that represents the ranking of its features
  - Learn a linear model over these matrices, with a rank-1 matrix of weights

Thank you for your attention!