QUT Centre for Data Science

ACEMS
AUSTRALIAN RESEARCH COUNCIL CENTRE OF EXCELLENCE FOR MATHEMATICAL AND STATISTICAL FRONTIERS

Bayesian Modelling and Analysis of Challenging Data

Kerrie Mengersen
School of Mathematical Sciences
QUT

PC Mahalanobis Lecture Series
January 2021

ACEMS
AUSTRALIAN RESEARCH COUNCIL CENTRE OF EXCELLENCE FOR
MATHEMATICAL AND STATISTICAL FRONTIERS

Australian Research Council Centre of Excellence

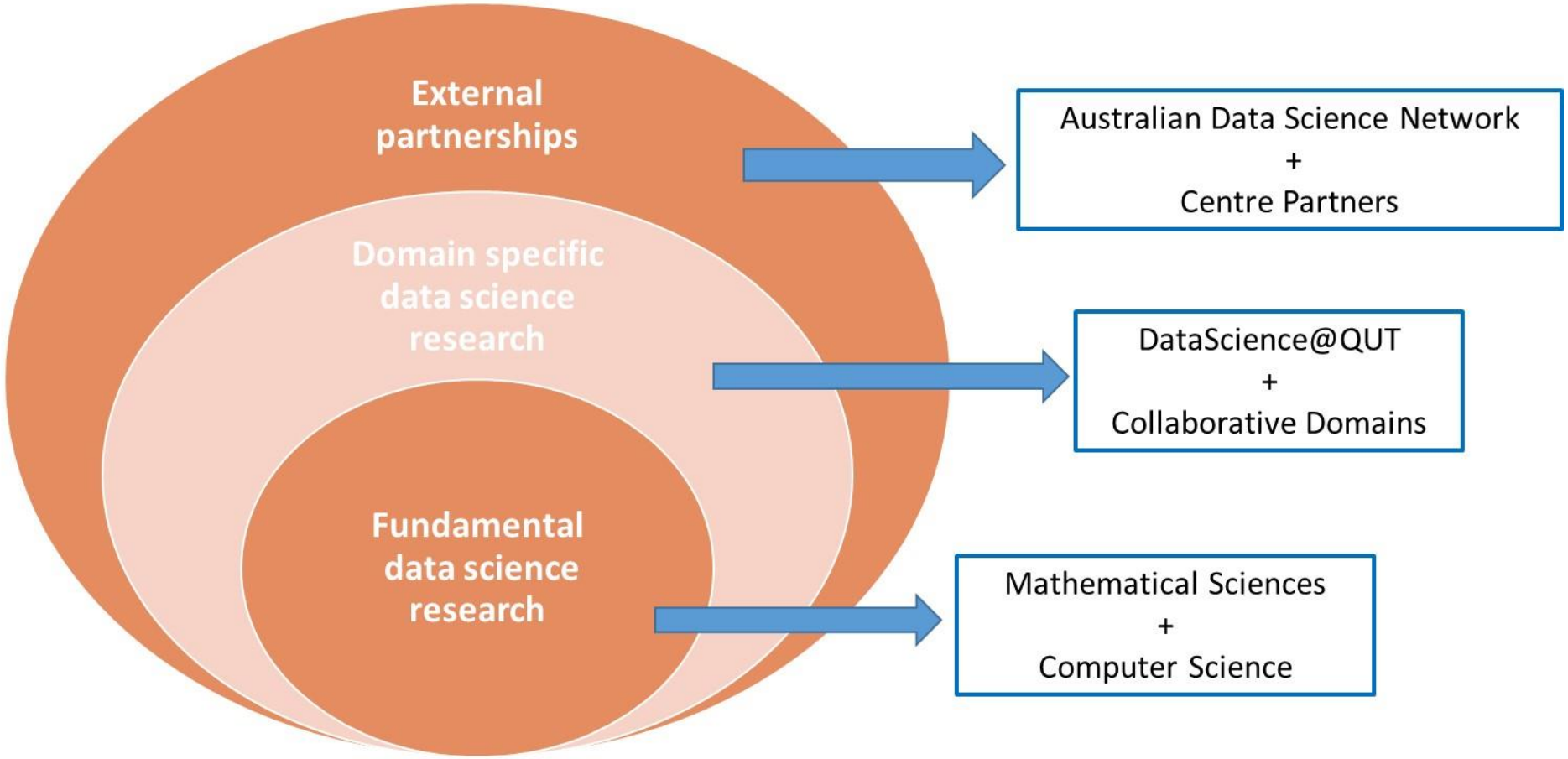
Mathematical & Statistical Frontiers:

Big Data, Big Models, New Insights

QUT Centre for Data Science

…to be a national and global leader in the development of frontier methods for the purposeful use of data to benefit our world

Centre for Data Science



Leadership Team

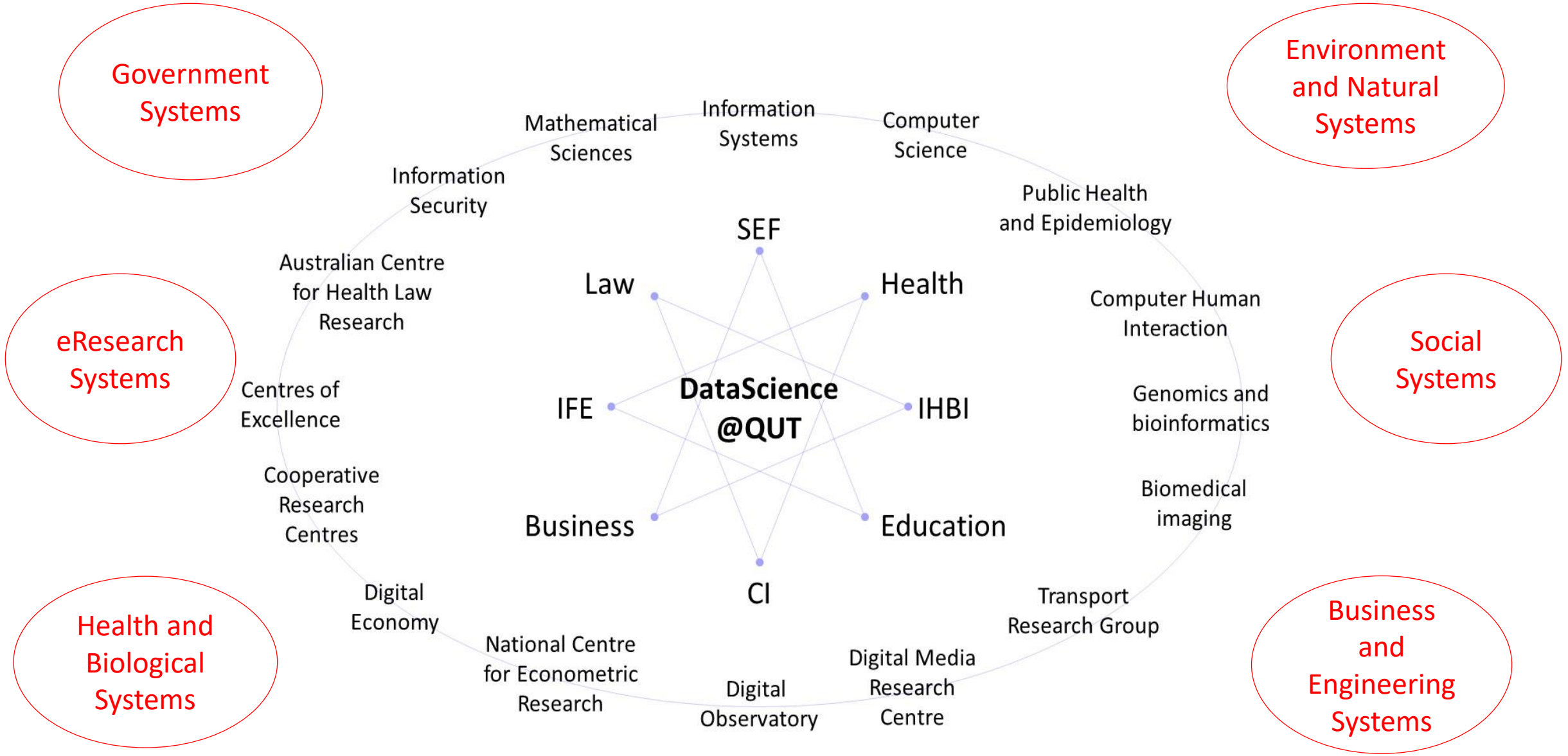
Kerrie Mengersen
Director

Chris Drovandi
Models and Algorithms

Richi Nayak
Applied Data Science

James McGree
Data for Discovery

David Lovell
Data Focused Decision Making


Government Systems
Environment and Natural Systems
eResearch Systems
Social Systems
Health and Biological Systems
Business and Engineering Systems
DataScience @QUT
SEF
Law
Health
IFE
IHBI
Business
Education
CI
Information Systems
Computer Science
Mathematical Sciences
Information Security
Public Health and Epidemiology
Australian Centre for Health Law Research
Computer Human Interaction
Centres of Excellence
Genomics and bioinformatics
Cooperative Research Centres
Biomedical imaging
Digital Economy
Transport Research Group
National Centre for Econometric Research
Digital Media Research Centre
Digital Observatory

AMSI
ACEMS
Swinburne
Monash
RMIT
Queensland
Data61
Corporate Research
NSW
UTS
Australian Data Science Network
SA Govt
Professional Societies
Victoria
UQ
Wollongong
NSW Treasury
Government
ACT
South Australia
Uni SA
Melbourne
Deakin
Adelaide
La Trobe

Programme of Lectures

January 27th:

- Lecture 1: 10-1045am IST (230pm-3:15pm AEST)
  Identifying the Intrinsic Dimension of High-Dimensional Data

- Lecture 2: 11-11:45am IST (3:30pm-4:15pm AEST)
  Finding Patterns in Highly Structured Spatio-Temporal Data

January 29th:

- Lecture 3: 10-1045am IST (230pm-3:15pm AEST)
  Describing Systems of Data

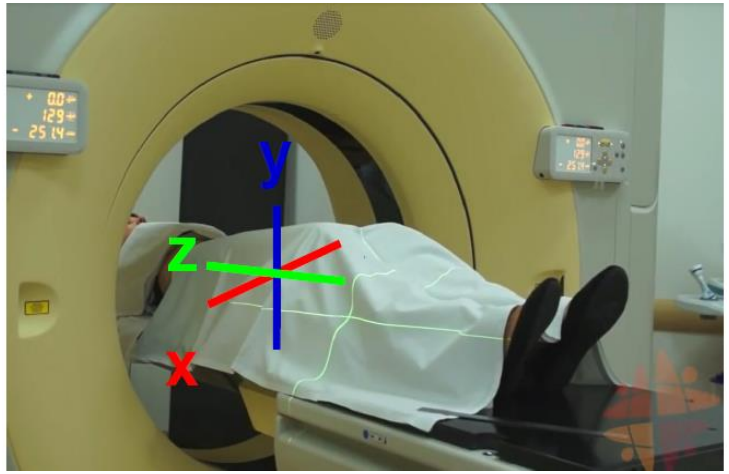
- Lecture 4: 11-11:45am IST (3:30pm-4:15pm AEST)
  Making New Sources of Data Trustworthy

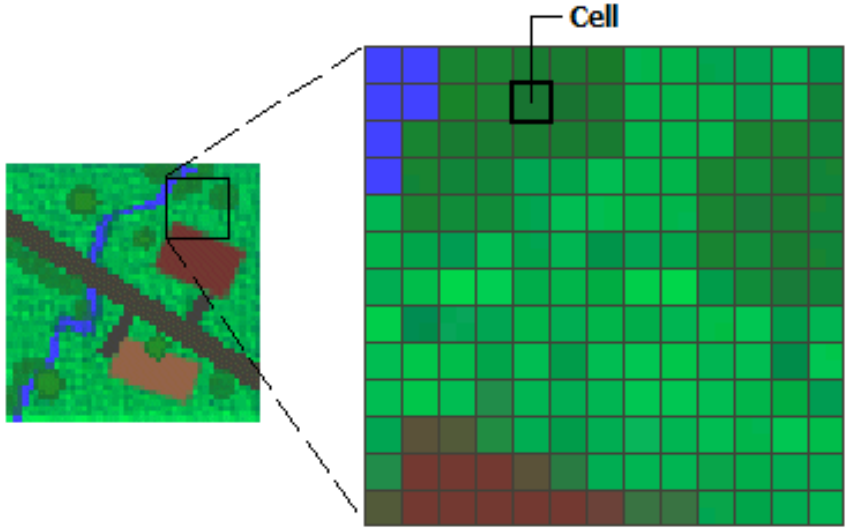
Bayesian Modelling and Analysis of Challenging Data

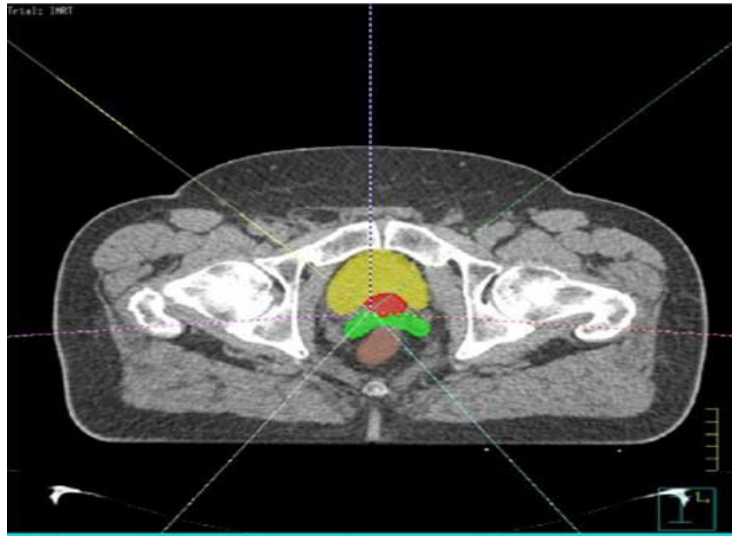
Lecture 1:
Identifying the intrinsic dimension of high-dimensional data

Hongbo Xie, Insha Ullah,
Edgar Santos-Fernandez, Antonietta Mira,
Benoit Liquet, Matthew Sutton

Case Study 1: Analysis of images



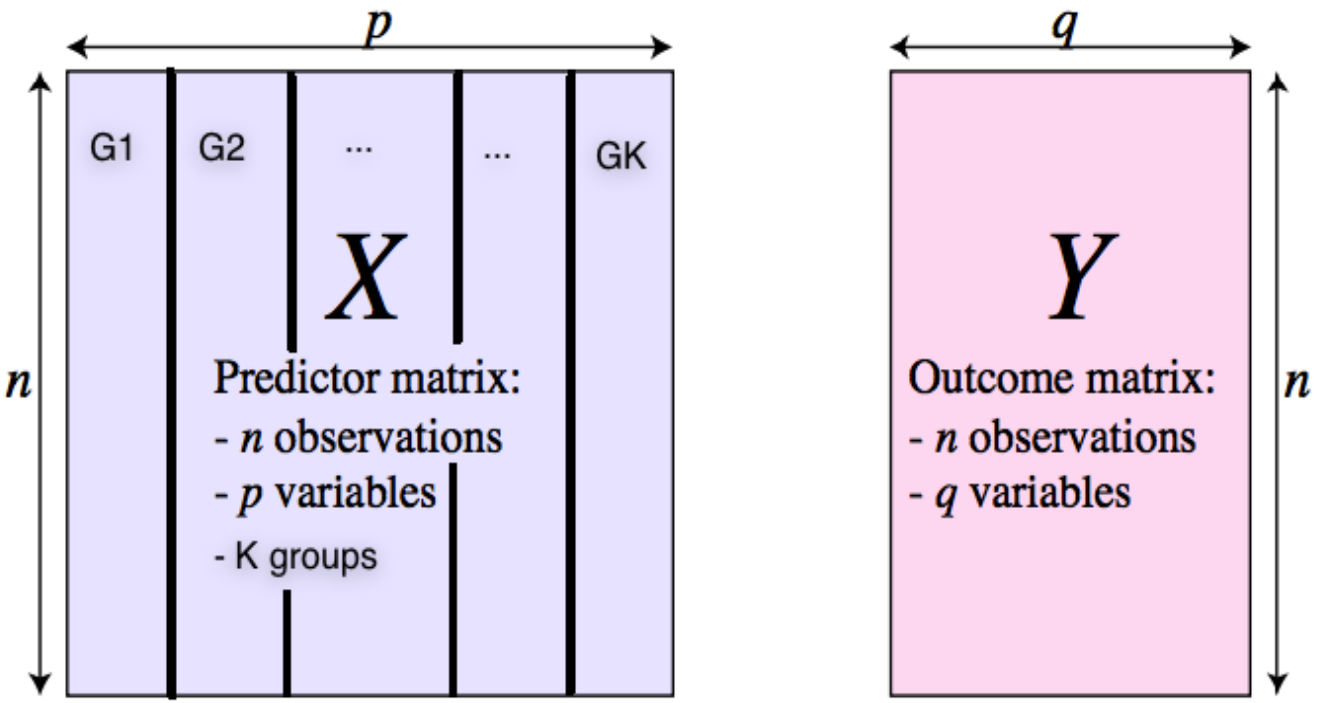




Case study 2: player tracking

- Individual player tracking devices
- What to analyse and compare plays
- Home vs away? Winning vs losing teams?

Case study 3: genomic data



- Multiple disease outcomes
- Many potential predictors (genes)
- $p >> n$

Intrinsic dimension

Lower bound of the dimension of a dataset == measure of complexity of dataset or signal

- Pattern recognition: the number of variables needed in a minimal representation of the data.
- Signal processing: how many variables are needed to generate a good approximation of the signal.
- Estimation: a representation in the intrinsic dimension does only need to exist locally, i.e. different intrinsic dimensions in different parts of the data set.

- History:
  - Term coined by Bennet (1965) in information theory
  - 1960's: Estimation of ID in multidimensional scaling
  - 1970's: ID estimation methods based on local eigenvalues, distance functions, etc
  - 1980's: ID of sets and probability measures in dynamical systems (fractal dimension)
  - 2000's: 'curse of dimensionality'

Wikipedia

Case Study 1: Matrix Factorization

Aim: to extract low-rank and/or sparse structures (e.g., classes)

Approach: Use matrix factorization techniques

Data: Y ($M \times N$) matrix

Solve: Recover the actual low rank matrix X, i.e.

$$\mathrm{Y} = \mathrm{X} + \mathrm{E} = \mathrm{U}\mathrm{V}^T + \mathrm{E}$$

$$\mathrm{Y} \epsilon R^{M \times N}, \mathrm{U} \epsilon R^{M \times r}, \mathrm{V} \epsilon R^{N \times r}, \mathrm{E} \epsilon R^{M \times N}$$

$r \ll \min(M, N)$ for sparsity

Bayesian model: use priors to induce sparsity

Gaussian priors for columns of U and V:

$$p(\mathrm{U}|\gamma) = \prod_{j=1}^{r} N(u_{.j}|0, \gamma_j^{-1}\mathrm{I}_M)$$

$$p(\mathrm{V}|\gamma) = \prod_{j=1}^{r} N(v_{.j}|0, \gamma_j^{-1}\mathrm{I}_M)$$

Same γ_j

$$p(\gamma_j) = \mathrm{Gamma}(a, \frac{1}{b})$$

Small a,b

Couple U with a kernel matrix K_U to give a latent matrix G with prior

$$p(\mathrm{G}|\mathrm{U}, \mathrm{K_U}, \sigma_g) = \prod_{j=1}^{r} N(g_{.j}|\mathrm{K_U^T} \cdot u_{.j}, \sigma_g^{-1}\mathrm{I}_M)$$

Jeffreys prior:

$$p(\sigma_g) = \sigma_g^{-1}$$

Similarly,

$$p(\mathrm{H}|\mathrm{V}, \mathrm{K_V}, \sigma_h) = \prod_{j=1}^{r} N(h_{.j}|\mathrm{K_V^T} \cdot v_{.j}, \sigma_h^{-1}\mathrm{I}_N)$$

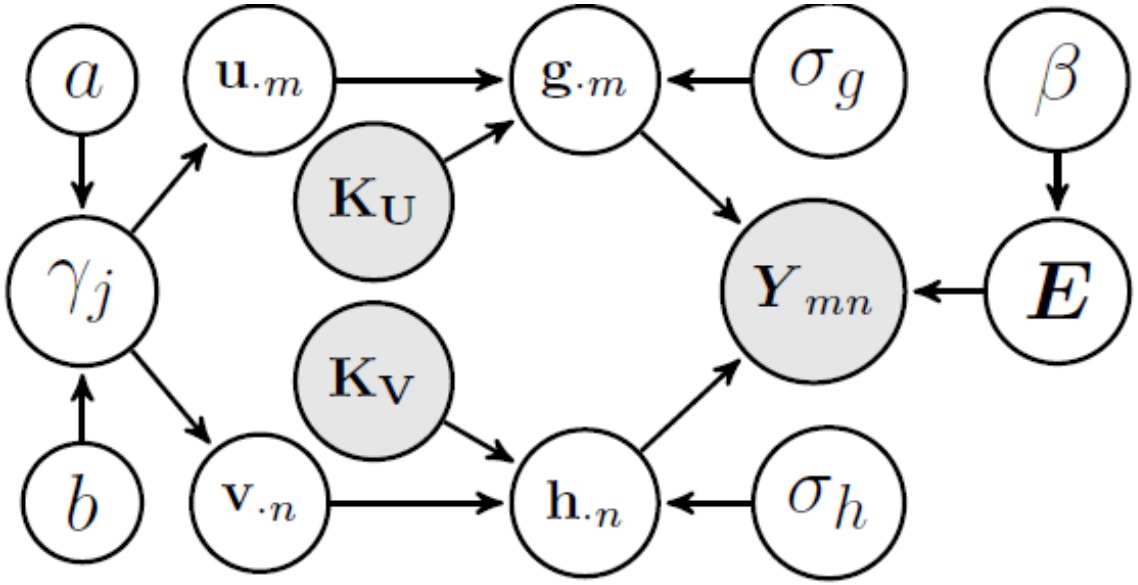
$$p(\sigma_h) = \sigma_h^{-1}$$

Residual term

$$p(\mathrm{E}) = \prod_{i=1}^{M} \prod_{j=1}^{N} N(\varepsilon_{mn}|0, \beta^{-1})$$

$$p(\beta) = \beta^{-1}$$

Graphical model



Conditional and joint distributions

Conditional distribution for observation model:

$$p(\mathrm{Y}|\mathrm{G},\mathrm{H},\beta) = N\left(\mathrm{Y}\middle|\mathrm{GH}^{\mathrm{T}},\beta^{-1}\mathrm{I}_{MN}\right)$$

Joint distribution:

$$p\left(\mathrm{Y},\mathrm{U},\mathrm{V},\mathrm{G},\mathrm{H},\sigma_g,\sigma_h,\gamma,\beta\right) =$$

$$p(\mathrm{Y}|\mathrm{G},\mathrm{H},\beta)p\left(\mathrm{G}\middle|\mathrm{U},\sigma_g\right)p(\mathrm{H}|\mathrm{V},\sigma_V)p(\mathrm{U}|\gamma)p(\mathrm{V}|\gamma)p\left(\sigma_g\right)p(\sigma_h)p(\gamma)p(\beta)$$

Computation: Variational Bayes

Choice of kernel

Many different types of kernels. For images, we want one that incorporates similarity information between patches into patch group matrix factorisation.

Let $d_E^{i,j}$ be the Euclidean distance between a pair of patches (i,j).

Define the similarity between them, i.e., an entry of K_u or K_v as

$$k_{ij} = \sqrt[4]{\frac{1}{1 + d_E^{i,j}/M}}$$

M is the total number of pixels in the patch

- A pixel and its nearest neighbours in a $\sqrt{M} \times \sqrt{M}$ window are modelled as a column vector.
- Construct the $M \times N$ patch group matrix Y by grouping other $N - 1$ patches with similar local spatial structures to the underlying one in the local window.
- Since each column shares similar underlying image structures, the noise-free patch group matrix Y has the low-rank property.

Overall algorithm

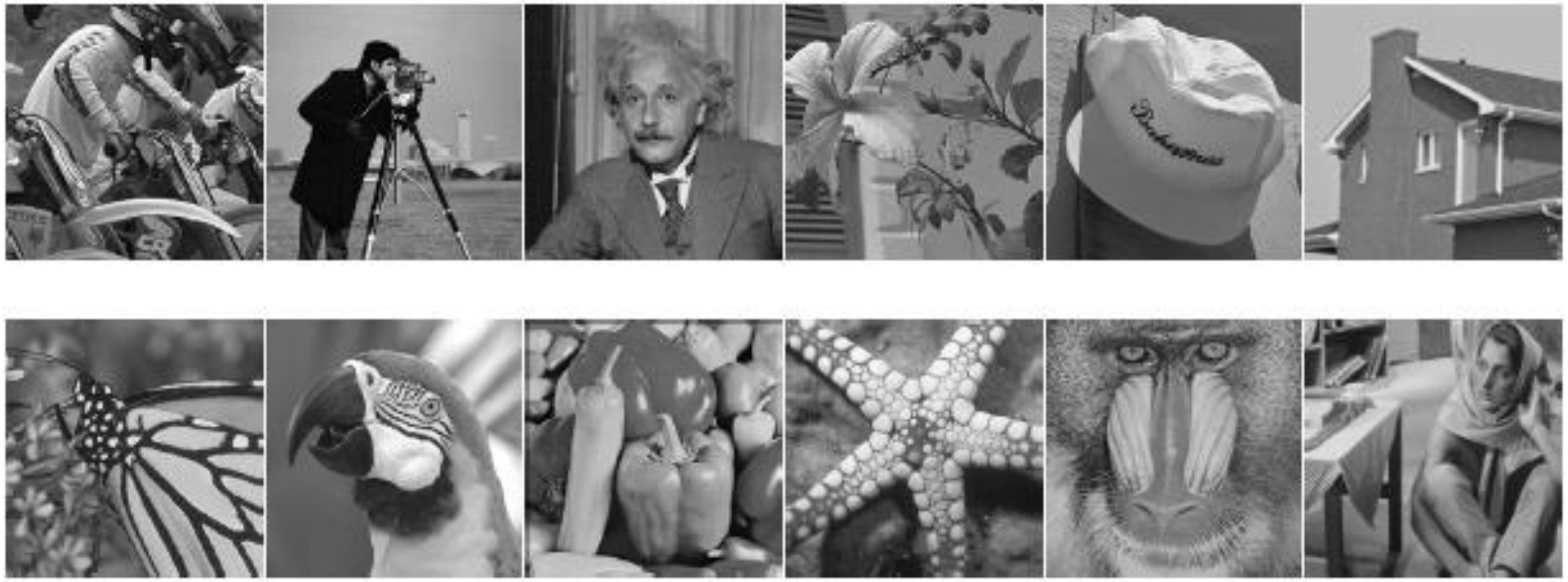
1. Cluster patches with similar spatial structure to form a patch matrix.

2. Apply KSBMF in succession on each patch matrix.

3. Aggregate the patches to reconstruct the whole image.

Case study: image restoration

12 test images:



First 10 images 256×256; last 2 images 512×512.

Add $N(0, \sigma^2)$ noise, $\sigma = 20, 50, 70, 100$.

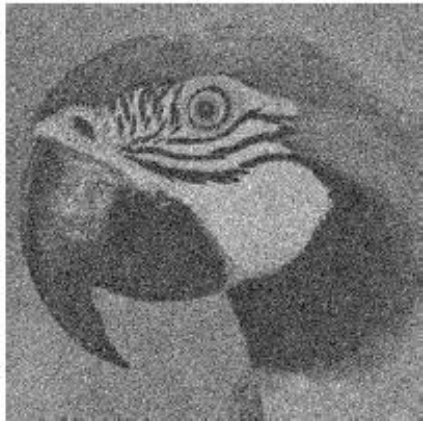
Patch sizes set to $6 \times 6, \ldots, 9 \times 9$; 70, 90, 120, 140.

No side information about similarity between patches, so set K_U to be the Identity matrix.

Case study: results

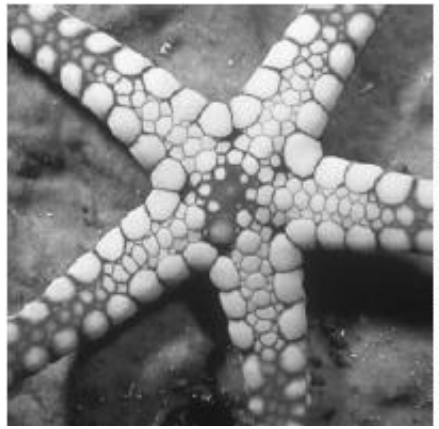
Original

Noisy

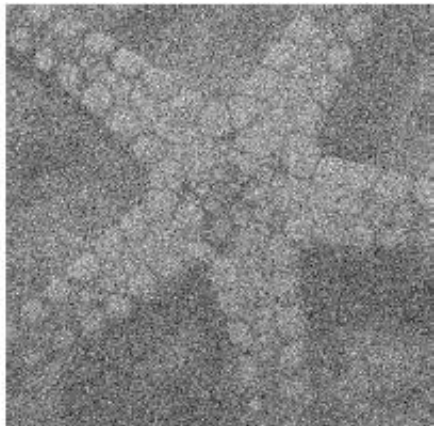


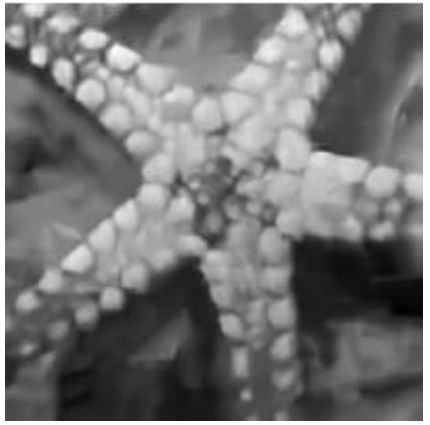
Reconstructed

Original



Noisy





Reconstructed

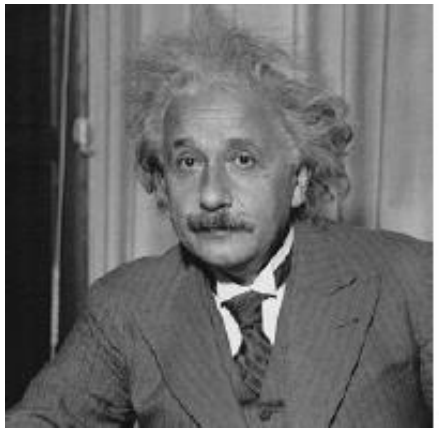
Case study: results

Original

Noisy

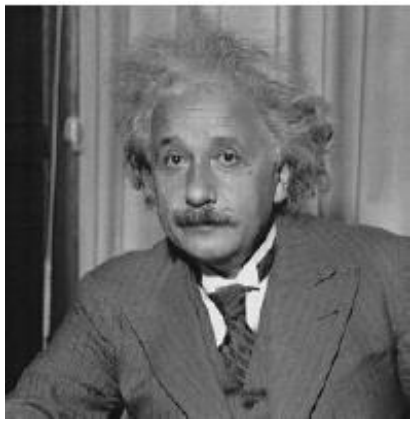
Reconstructed

Original



Noisy

Reconstructed



Summary

- New (KSBMF) model allows integration of side information.
- KSBMF automatically infers the parameters and latent variables including the reduced rank using variational Bayesian inference.
- The model simultaneously achieves low rank through sparsity induced by an enforced constraint on latent factor matrices.
- Experimental results demonstrate that KSBMF outperforms state-of-the-art approaches for image restoration tasks under various types and levels of corruption.

Case Study 2. Dimension reduction of tracking data via ID

- A small number of dimensions/variables is often sufficient to effectively describe high-dimensional data – “intrinsic dimension” (ID)
- ID can vary within the same dataset.

Examples:

- folded vs unfolded configurations in a protein molecular dynamics trajectory
- active vs inactive regions in brain imaging data
- organisations with different service delivery profiles
- sports teams with different strategies for attack/defence, home/away

- Aim is to cluster regions with the same local ID in a given data landscape.

Proposed approach: homogeneous model

- Data $x = (x_1, x_2, \ldots, x_N)$ sampled from a density $\rho(x)$ defined on a manifold with unknown ID, d, s.t. ρ is approximately constant in the region defined by the second neighbour of each point.
- Let r_{i1} and r_{i2} be the distances of the first and second neighbour of x_i.
- Then $\mu_i = r_{i2}/r_{i1}$ follows the Pareto distribution $f(\mu_i|d) = d{\mu_i}^{-(d+1)}$.
- Assuming $\mu = (\mu_1, \mu_2, \ldots \mu_N)$ are independent, then

$$P(\mu|d) = d^N \prod_{i=1}^{N} {\mu_1}^{-(d+1)} = d^N e^{-(d+1)V}$$

where $V = \sum_{i=1}^{N} \log(\mu_i)$

⟹ Place a suitable prior on *d* and estimate *d*.

Proposed approach: heterogeneous model

- Let x be sampled from a density $\rho(x)$ with support on the union of K manifolds with varying dimensions. Then

$$\rho(x) = \sum_{k=1}^{K} \rho_k(x)$$

  - each $\rho_k(x)$ has support on a manifold of dimension d_k
  - $p=(p_1, p_2,\ldots, p_K)$ are the *a priori* probabilities that a point belongs to the manifolds $1,\ldots,K$.
- Now the distribution of μ_i is a mixture of Pareto distributions:

$$P(\mu|d) = \sum_{k=1}^{K} p_k d_k \, \mu_i^{-(d+1)V}$$

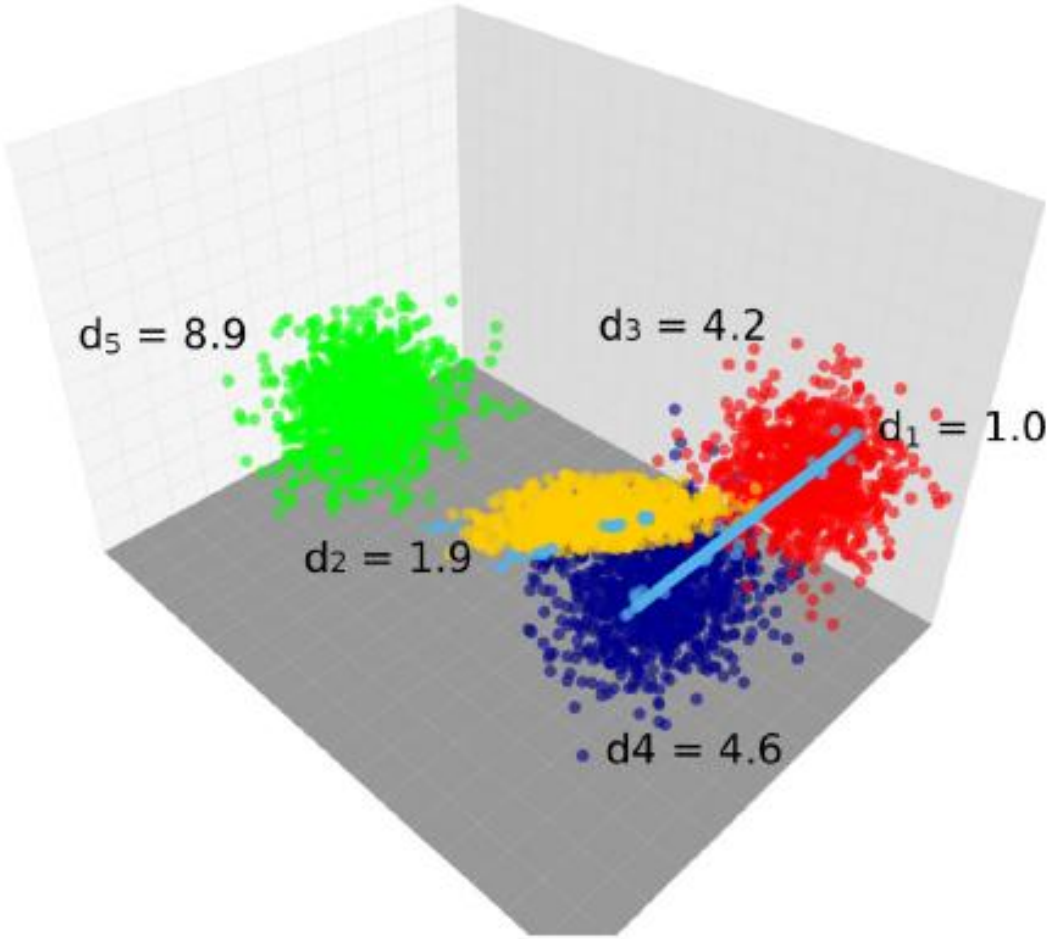
- Introduce latent variables $z = (z_1, z_2, \ldots, z_K)$ where $z_1 = k$ indicates that point i belongs to manifold k. Hence

$$P_{post}(z, d, p|\mu) \propto P(\mu|z, d)P(z|p)P(d)P(p)$$

- Priors $d_k \sim \text{Gamma}(a_k, b_k)$, $p \sim Dir(c_1, \ldots, c_K)$; $a_k, b_k, c_k = 1$

Simulation study

N=1000 points drawn from five Gaussians in dimensions
$d_1 = 1$, $d_2 = 2$, $d_3 = 4$, $d_4 = 5$, $d_5 = 9$



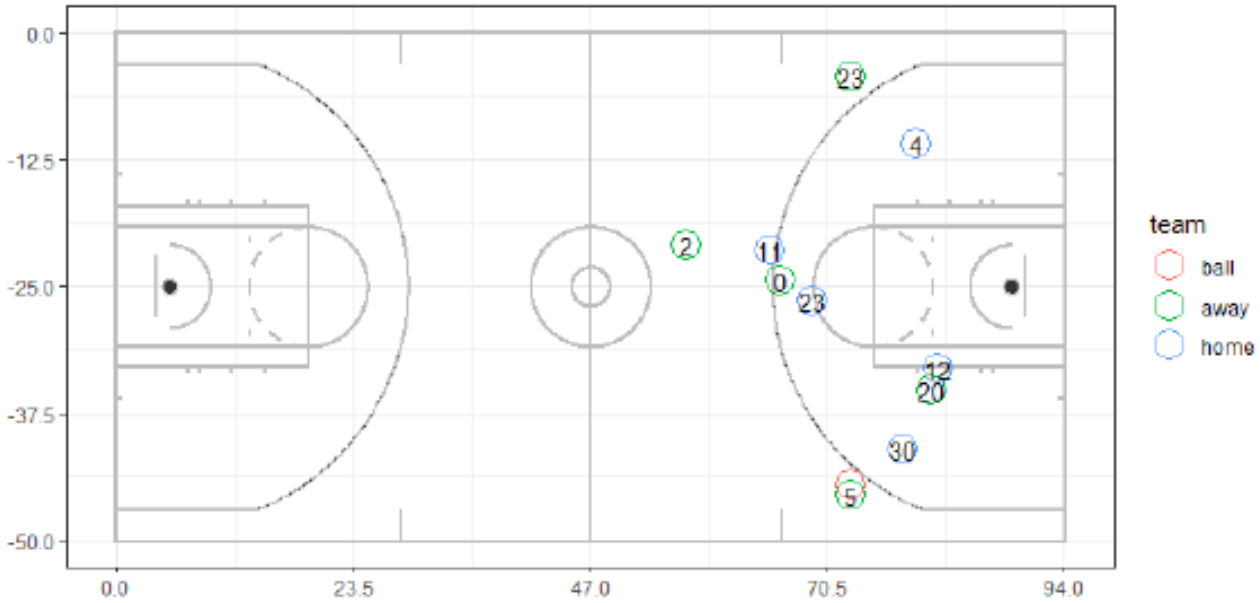
Case Study: ID in basketball

- SportVU NBA player tracking technology: player movement measurement 25 frames/sec.
- How is the placement of the players in attack and defence related to the success of a play?
- Do teams that have greater offensive ability produce successful shots from more unique locations in the court and do they create more shoot opportunities by passing the ball more effectively?
- What could ID tell us about the plays and teams?

Example

- High-resolution player tracking raw data from the NBA season 2015-16.
- Play-by-play events description and other statistics (https://stats.nba.com/).
- Verified via manual video annotation of the game (https://www.youtube.com/).
- From each play, we inferred the locations of the players at the moment of the shoot, and selected the events = {ShotMissed, ShotMadeg}.
- 15 games were included in the analysis.

Consider the game Cleveland Cavaliers (CLE) and the Golden State Warriors (GSW) from the 25th of December 2015.



Example

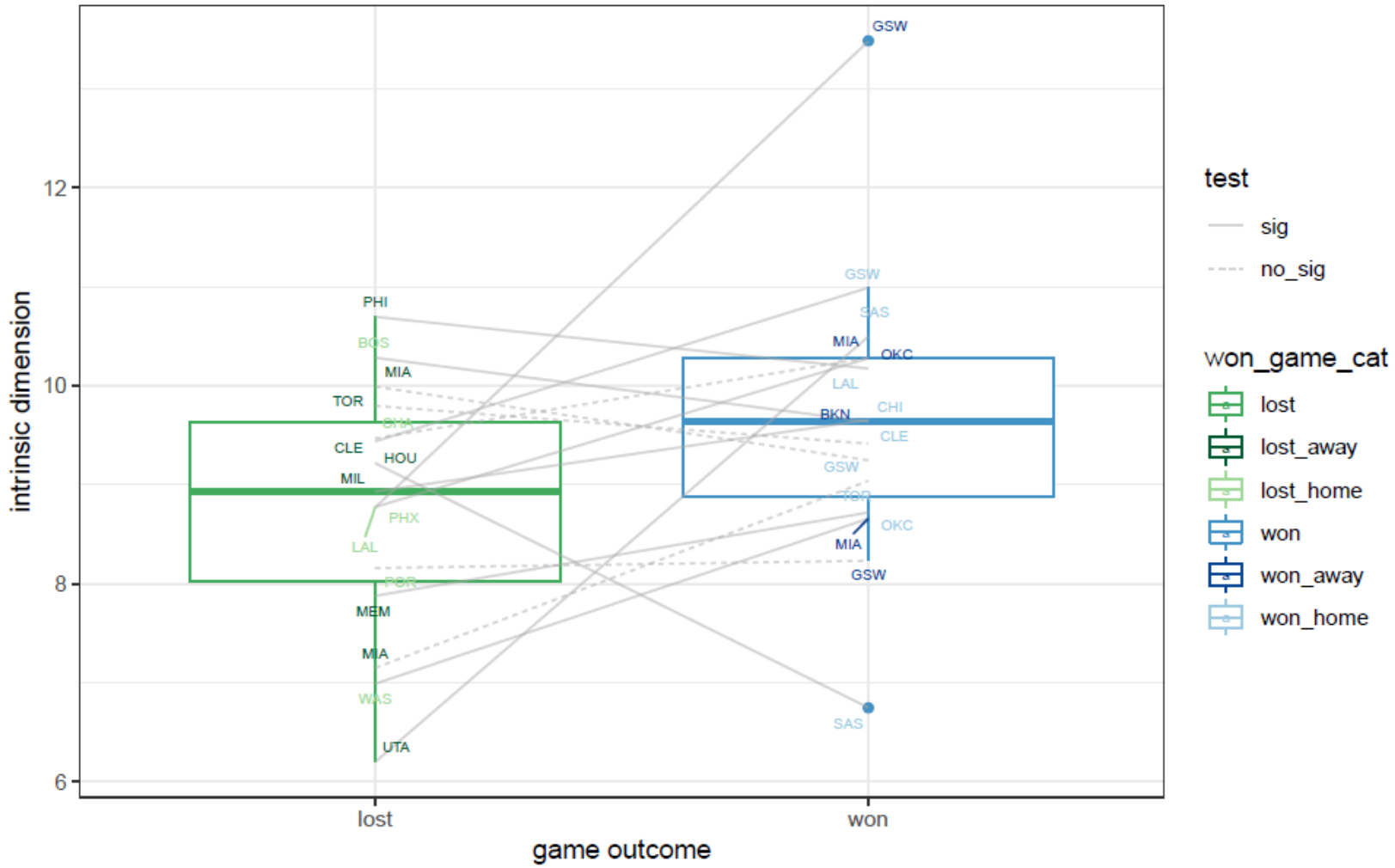
- Aim: compute the intrinsic dimension using the shot chart data from the home and away teams.

- Split the data into two sets: (1) field goals shots taken when the home team (GSW) is attacking, and (2) … when CLE is attacking.

- No. rows on each dataset = no. attempted field shots.
  No. columns = D = 20 (2 players' coordinates (x and y) 5 players 2 teams).

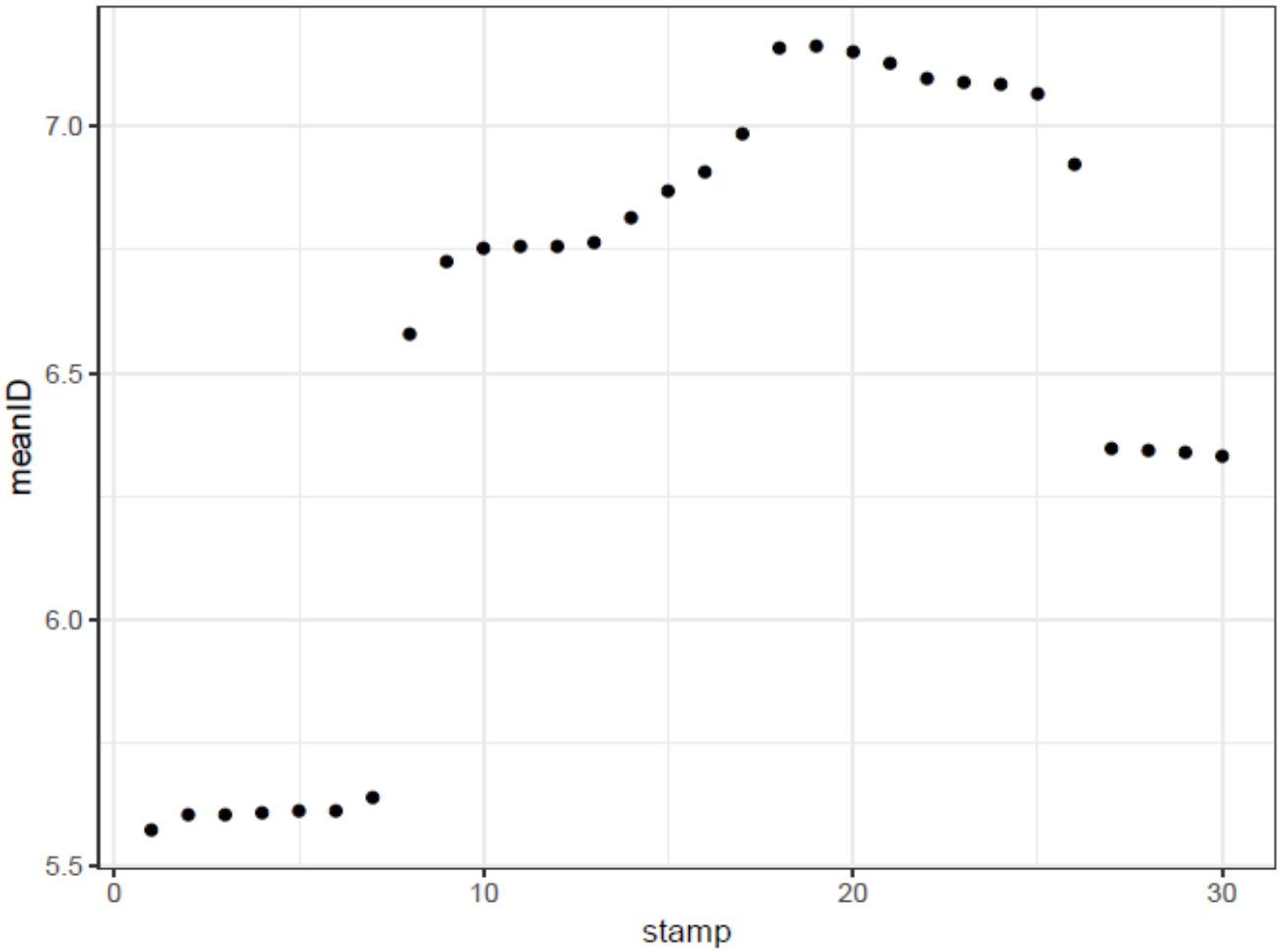
- The intrinsic dimension for the set of players (5 vs 5) corresponds to the number of independent directions in which the 20-dimensional points are embedded.

Results


intrinsic dimension
12
10
8
6
lost
won
game outcome
GSW
PHI
BOS
MIA
TOR
CLE
HOU
MIL
PHX
LAL
POR
MEM
MIA
WAS
UTA
GSW
SAS
MIA
OKC
LAL
BKN
CHI
CLE
GSW
TOR
OKC
MIA
GSW
SAS
test
sig
no_sig
won_game_cat
lost
lost_away
lost_home
won
won_away
won_home

Results

Posterior means of
The ID over the course
of play for the first
3-point field goal.



Case Study 3: High-dimensional variable selection

Challenge 1: wide, high-dimensional data

We have too many variables, prone to overfitting. Need to remove variable, or regularize, or both

- Main constraint: situation with $p > n$
- Strong colinearity among the variables.

Challenge 2: grouped data

- Genomics: genes within the same pathway have similar functions and act together in regulating a biological system.

$\hookrightarrow$ These genes can add up to have a larger effect

$\hookrightarrow$ can be detected as a group (i.e., at a pathway or gene set/module level).

We consider variables are divided into groups:

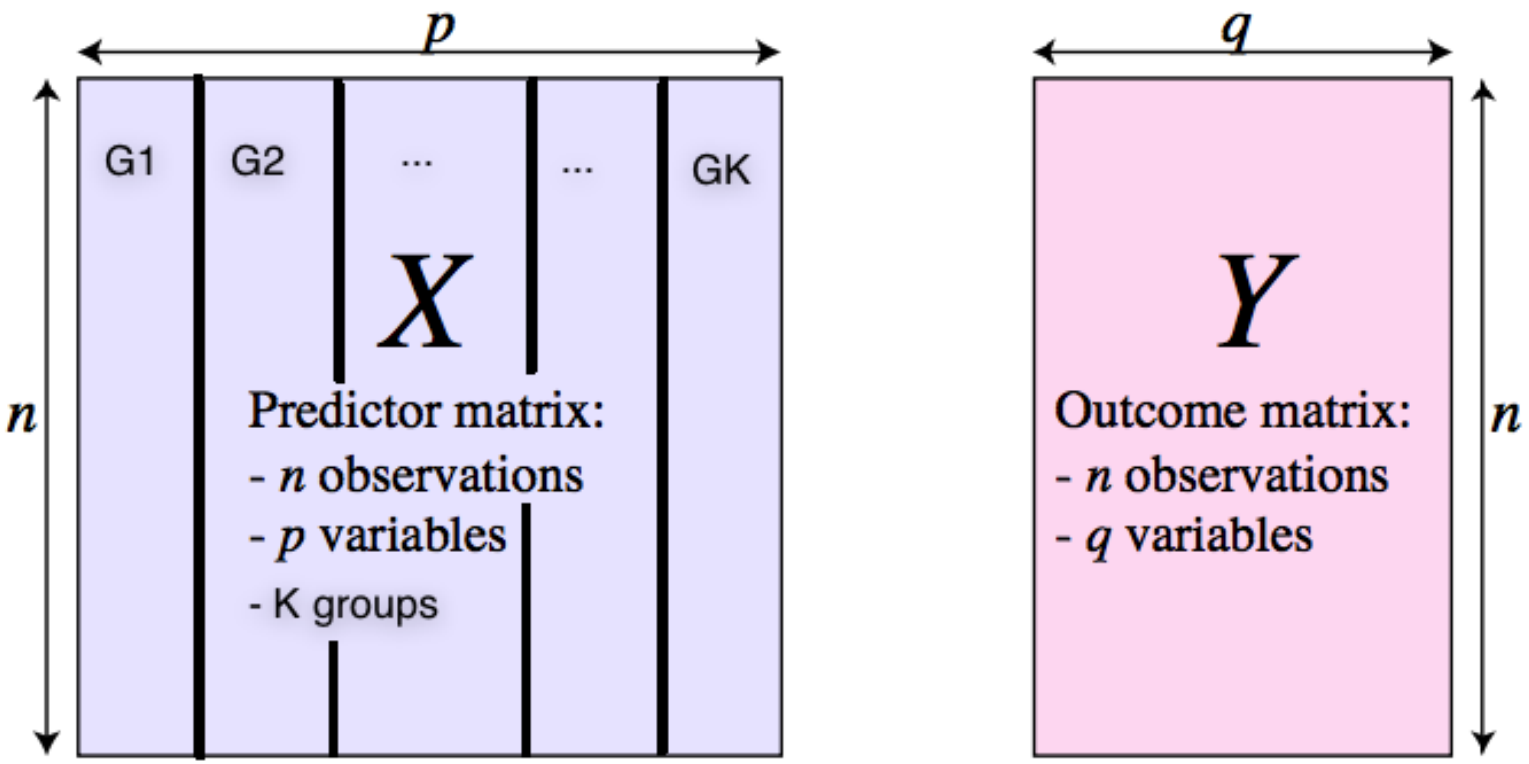
- Example p: SNPs grouped into K genes

$$\mathbf{X} = [\underbrace{SNP_1, \ldots + SNP_k}_{gene_1} \mid \underbrace{SNP_{k+1}, SNP_{k+2}, \ldots, SNP_h}_{gene_2} \mid \ldots \mid \underbrace{SNP_{l+1}, \ldots, SNP_p}_{gene_K}]$$

- Example p: genes grouped into K pathways/modules ($X_j = \text{gene}_j$)

$$\mathbf{X} = [\underbrace{X_1, X_2, \ldots, X_k}_{M_1} \mid \underbrace{X_{k+1}, X_{k+2}, \ldots, X_h}_{M_2} \mid \ldots \mid \underbrace{X_{l+1}, X_{l+2}, \ldots, X_p}_{M_K}]$$

Challenge 3: multivariate data



Solution?

Fully Bayesian sparse regression analysis for:

- Number of predictors (p) >> number observations (n)
- Multivariate response
- Covariates grouped by blocks
- Sparsity for blocks and within blocks

- Select group variables taking into account the data structures; all the variables within a group are selected, otherwise none of them are selected
- Combine both sparsity of groups and within each group; only relevant variables within a group are selected

Frequentist approaches: Partial Least Squares (PLS)

- Sparse Group PLS : SNP $\subset$ Gene or Gene $\subset$ Pathways

  Liquet B., Lafaye de Micheaux P., Hejblum B. and Thiebaut R., (2016) *Group and Sparse Group Partial Least Square Approaches Applied in Genomics Context.* **Bioinformatics**, 32(1), 35–42.

- Sparse Group subgroup PLS : SNP $\subset$ Gene $\subset$ Pathways

  M. Sutton, R. Thiebaut, and B. Liquet. (2018) *Sparse group subgroup Partial Least Squares with application to genomics data*. Statistics in Medicine.

  - combining L_1 and L_2 penalties into the optimization function
  - **Sparse Group Penalties:**

$$\lambda_1 \sum_{g=1}^{G} \sqrt{p_g} \|\beta_g\|_2 + \lambda_2 \|\beta\|_1$$

Multivariate Bayesian solution

Bayesian group lasso model with spike and slab priors

Liquet, Mengersen, Pettitt, Sutton (2017) Bayesian Analysis.

- **spike and slab** priors providing variable selection at the group level.
- hierarchical spike and slab prior structure to select variables both at the group level and within each group.

Model formulation

$$\mathbb{Y}|\mathbb{X}, \mathbb{B}, \Sigma \sim MN_{n\times q}(\mathbb{X}\mathbb{B}, \Sigma, \mathbb{I}_n),$$

$$Vec(\mathbb{B}_g^T|\Sigma, \tau_g, \pi_0) \overset{ind}{\sim} (1-\pi_0)N_{m_g q}(0, \mathbb{I}_{m_g} \otimes \tau_g^2\Sigma) + \pi_0\delta_0(Vec(\mathbb{B}_g^T))$$

$$\tau_g^2 \sim \text{Gamma}\left(\frac{m_g+1}{2}, \frac{\lambda_g^2}{2}\right), g = 1, \ldots, G$$

$$\Sigma \sim \text{IW}(d, Q)$$

$$\pi_0 \sim \text{Beta}(a, b)$$

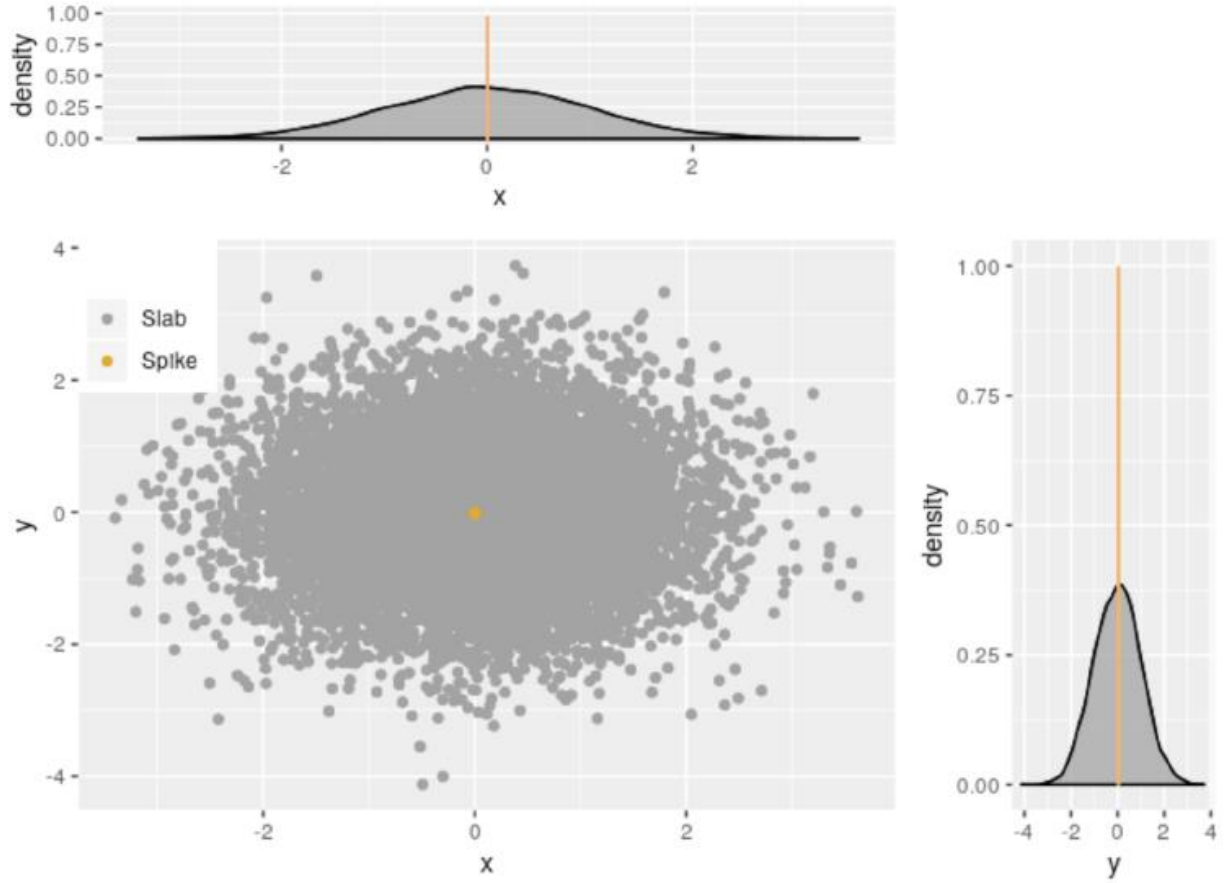
$\lambda_g = \sqrt{m_g}\lambda$: shrinkage for gth group
λ : global shrinkage parameter, estimated using Empirical Bayes
m_g : size of the group
λ_g is 'adaptive shrinkage', estimated using Monte Carlo EM

$\delta_0\left(Vec\left(\mathbb{B}_g^T\right)\right)$: point mass at $\mathbf{0} \in \mathbb{R}^{m_g q}$
$\mathbb{B}_{\boldsymbol{g}}$: $m_g \times q$ regression coefficient matrix for group g

d : d.f.
$Q = k\mathbb{I}_q$: positive finite scale matrix
$k \approx Var(\mathbb{Y}|\mathbb{X})$
$\text{E}(\Sigma) = Q/(d-2)$
$d = 3$: smallest integer ensuring existence of $\text{E}(\Sigma)$

a, b : prior information

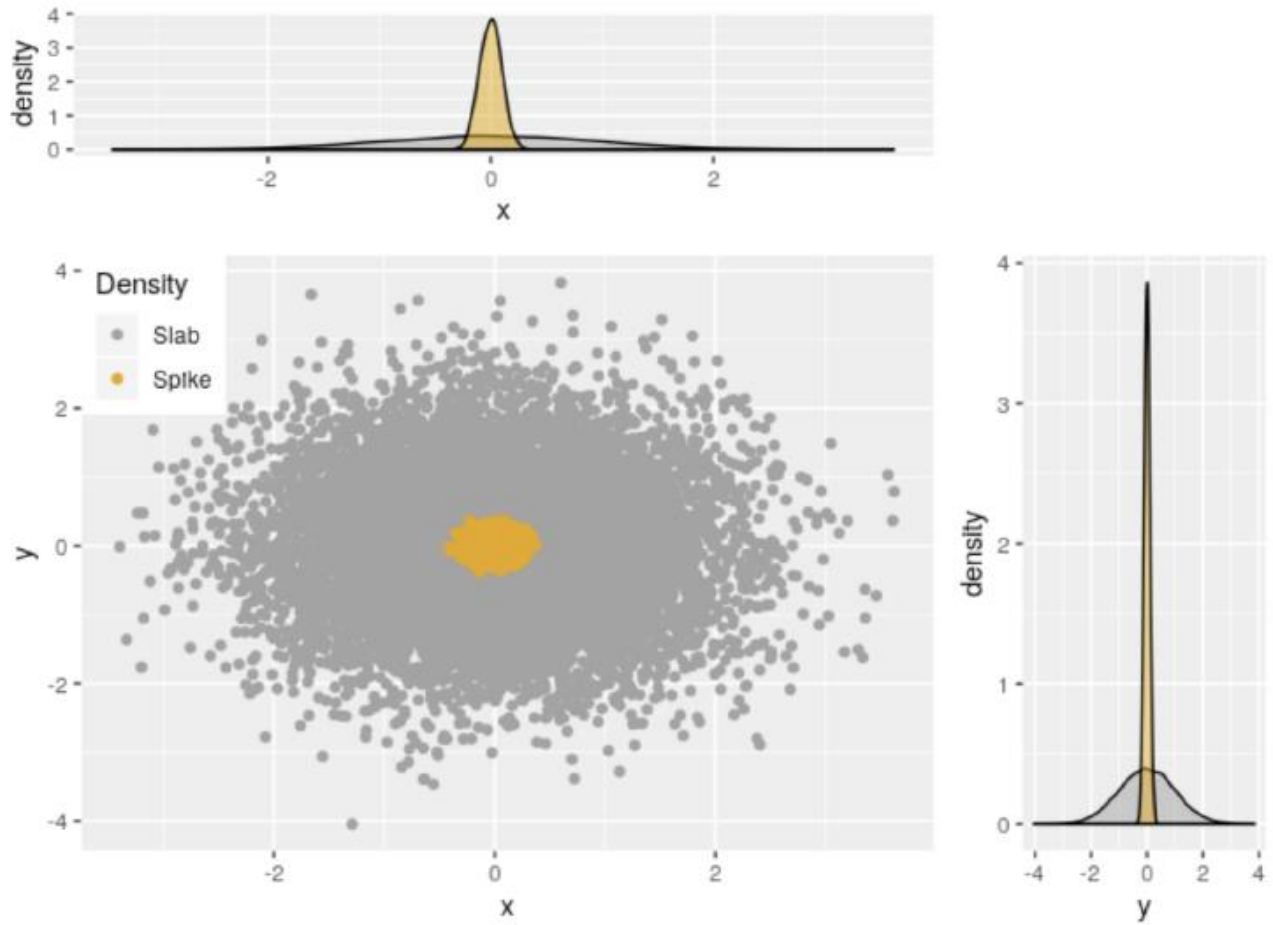
Spike and slab prior



Bivariate Dirac spike prior:

5000 samples at zero (spike)
and
5000 N(0,1) samples (slab)

Spike and slab prior



Bivariate Dirac spike prior:

5000 $N(0,10^{-2})$ samples (spike)
and
5000 $N(0,1)$ samples (slab)

Posterior median estimator

Use the posterior median estimator for both selection and estimation at the same time.

Benefits:

- Enables group variable selection by obtaining a zero coefficient for some groups
- Can be expressed as a soft thresholding estimator
- Consistent in model selection and has optimal asymptotic estimation rate.

Analysis via Gibbs sampler

$$p(\mathbb{B}, \boldsymbol{\tau}^2, \Sigma, \pi_0 | \mathbb{Y}, \mathbb{X}) \propto p(\mathbb{Y} | \mathbb{B}, \boldsymbol{\tau}^2, \Sigma, \pi_0) \times p(\mathbb{B} | \boldsymbol{\tau}^2, \Sigma, \pi_0) \times p(\boldsymbol{\tau}^2) \times p(\Sigma) \times p(\pi_0)$$

$$Vec(\mathbb{B}_g^T) | \text{rest} \sim (1 - l_g) N_{m_g q} \left(Vec(\mathbb{M}_g^T), \Sigma_g \otimes \Sigma \right) + l_g \delta_0 (Vec(\mathbb{B}_g^T)), \quad g = 1, \ldots, G$$

$$\alpha_g^2 | \text{rest} \sim \begin{cases} \text{Inverse Gamma} \left(\text{shape} = \frac{m_g q + 1}{2}, \text{scale} = \frac{\lambda_g^2}{2} \right), & \text{if } \mathbb{B}_g = 0 \\ \text{Inverse Gaussian} \left(\frac{\lambda_g}{(Tr[\mathbb{B}_g \Sigma^{-1} \mathbb{B}_g^T])^{-1/2}}, \lambda_g^2 \right), & \text{if } \mathbb{B}_g \neq 0 \end{cases}$$

$\alpha_g^2 = 1/\tau_g^2$

$$\Sigma | \text{rest} \sim \text{IW} \left(d + n + \sum_{g=1}^{G} m_g Z_g, (\mathbb{Y} - \mathbb{X}\mathbb{B})^T (\mathbb{Y} - \mathbb{X}\mathbb{B}) + \mathbb{B}^T \mathbb{D}_{\boldsymbol{\tau}} \mathbb{B} + Q \right)$$

$$Z_g = \begin{cases} 1 & \text{if } \mathbb{B}_g \neq 0, \\ 0 & \text{if } \mathbb{B}_g = 0 \end{cases}$$

$$\pi_0 | \text{rest} \sim \text{Beta} \left(a + G - \sum_{g=1}^{G} Z_g, b + \sum_{g=1}^{G} Z_g \right)$$

Sparsity of groups and within each group

1. Reparametrise the coefficient matrices to tackle the two kinds of sparsity separately

2. Assume a multivariate spike and slab prior:

   - for each group-specific vector $\mathbb{B}_g$ to choose groups
   - for each τ_{gj} to choose variables within a group

Case study

Expression Quantitative Trait Loci (eQTL):

- Discover the genetic causes of variation in the expression (i.e. transcription) of genes
- gene expression data are treated as a quantitative phenotype
- genotype data (SNPs) are used as predictors

Example:

- Hopx genes, part of a larger study (Heinig et al., 2010; Liquet et al., 2016)
- Identify a parsimonious set of predictors that explains the joint variability of gene expression in four tissues (adrenal gland, fat, heart, and kidney).

Case study - data

- Predictor matrix: 770 SNPs in 29 inbred rats ($n = 29$, $p = 770$)
- Outcome: the 29 measured expression levels in the 4 tissues ($q = 4$)

	Correlation				Summary statistics	
	ADR	Fat	Heart	Kidney	Mean	Variance
ADR	1.00	0.46	0.44	0.70	4.72	0.07
Fat		1.00	0.24	0.42	8.23	0.09
Heart			1.00	0.44	8.79	1.61
Kidney				1.00	6.65	0.07

Chromosome	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Group size	74	67	63	60	39	45	52	43	31	51	21	26	33	22	15	27	18	30	34	19

Repartition of the SNPs along the chromosomes, defines the group structure

Results

- 32 SNPs which belong to 6 groups/chromosomes
- Empirical estimates of the probability of inclusion of each chromosome

Chromosome	1	2	3	4	5	6	7	8	9	10
EPI	0.00	1.00	1.00	1.00	0.72	0.00	1.00	0.00	0.00	1.00
Chromosome	11	12	13	14	15	16	17	18	19	20
EPI	0.83	0.12	0.46	1.00	0.93	0.88	0.79	0.59	0.89	0.40

- Posterior median estimates

Chromosome	SNP Name	ADR	Fat	Heart	Kidney
2	D2Rat147	0.00553	0.00238	-	0.00329
2	D2Rat222	0.00442	0.00116	-	0.00305
2	D2CebrP476s2	0.00123	-	-	-
2	D2Rat69	0.00715	0.01748	0.00730	0.00620
2	D4Ucsf2	0.00054	-	-	-
2	D7Cebr205s3	0.00246	-	0.00950	0.00461
3	D7Cebr14C16s2	0.00209	0.00326	-	0.00049
4	D7Rat112	0.00035	0.00001	-	-
4	D7Rat19	0.01113	0.01800	0.03680	0.01828
4	Cyp11b2	0.00075	0.00374	-	0.00394
7	D10Ntr32	0.00123	-	0.01112	0.00143
7	D10Rat31	0.00031	0.00573	0.00442	0.00316
7	D10Cebr39s2	0.00280	0.00490	0.00821	0.00586
7	Es13	0.00539	-	0.00924	0.00419
7	D10Rat226	0.00415	0.00006	0.00987	0.00372
7	D14Rat36	0.00036	-	0.03076	-
7	D14Cebrp312s2	0.00004	-	0.05427	-
10	**D14Mit3**	**0.04963**	**0.05415**	**0.33434**	**0.07491**
10	D15Rat21	0.00937	0.00569	0.03140	0.01704
10	D19Utr1	0.00149	0.00297	0.00251	0.00487
10	Ednra	0.00026	-	-	-
10	D2Mit16	-	0.00077	-	-
10	D2Rat70	-	0.00190	-	-
10	D3Cebr204s4	-	0.00042	-	-
14	D4Rat49	-	0.00102	0.00092	0.00401
14	D7Mit6	-	0.00002	-	-
14	D10Rat102	-	0.00112	-	-
14	D4Rat252	-	-	-0.00184	-
14	Myc	-	-	0.00669	-
15	D10Mit3	-	-	0.00104	-
19	D14Rat8	-	-	0.00058	-
19	D14Rat52	-	-	0.00361	-

Overall:

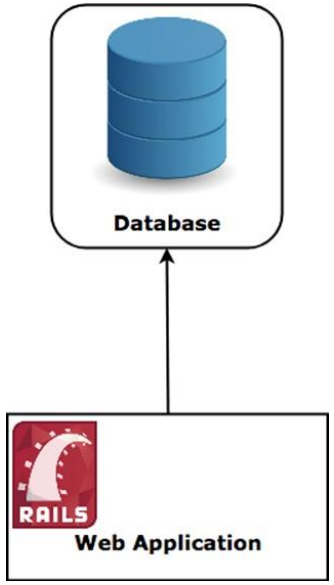
New Bayesian methods for group-sparse modelling of a multivariate correlated response variable.

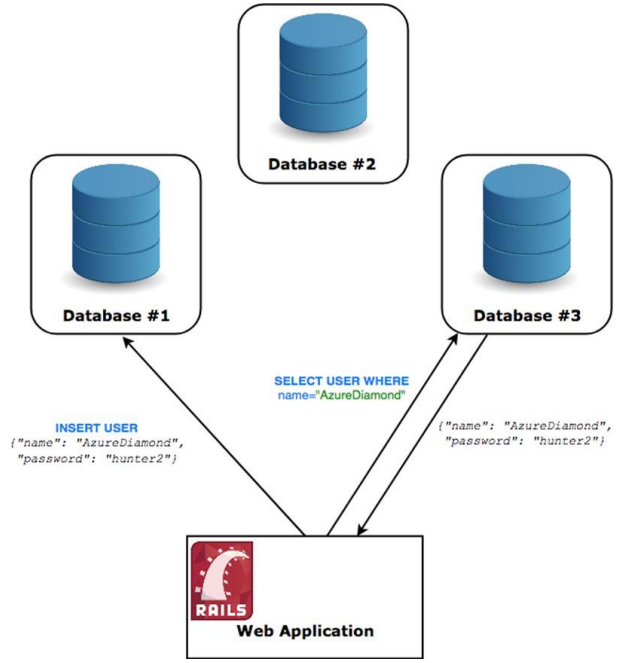
Highlights:

- Spike and slab type priors can facilitate both group and variable selection
- Posterior median estimator can both select and estimate the regression coefficients, and produce sparsity
- Simulation results showed excellent performance of the model and superior performance compared to other approaches
- Computation time is quite reasonable (minutes for case study)

Concern: (Not) Aggregating Data

Options





Single Database

Distributed Computing

Distributed Computing:

- Horizontal scaling
- Fault tolerance
- Low latency
- Sharding

MapReduce, Apache Spark, Hadoop, etc

[A Thorough Introduction to Distributed Systems (freecodecamp.org)]

Options

Decentralised Computing

- Information processed in the cloud
- Not owned by one actor
- Dapps
- Edge Computing

Edge Computing

• Information not processed on the cloud filtered through remote data centres; instead, the cloud comes to the centres

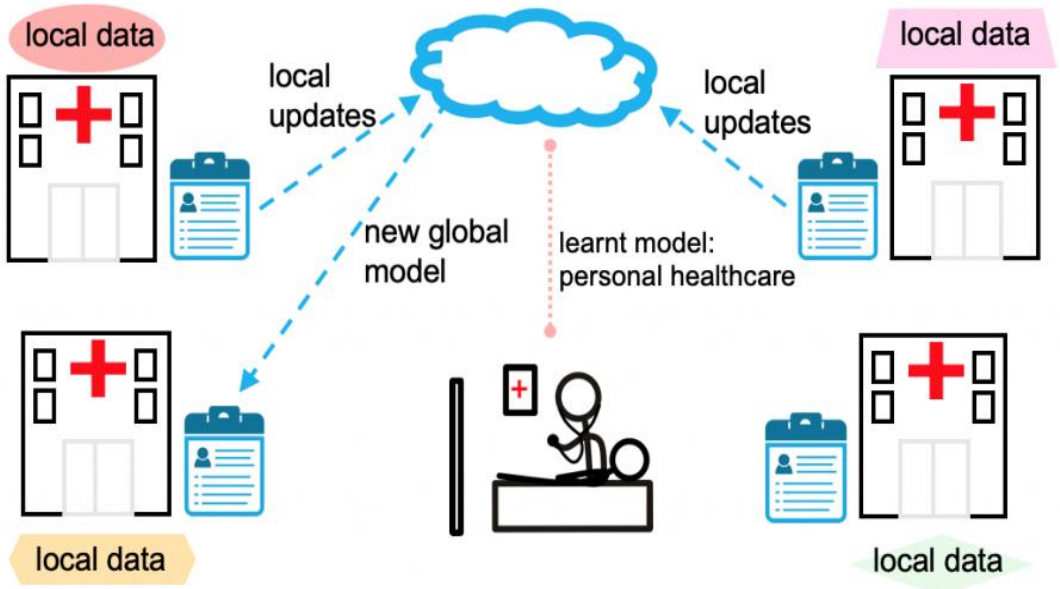
• Federated learning, Federated analysis

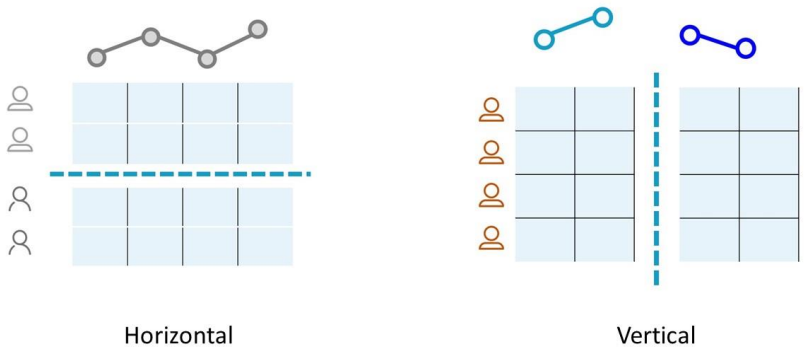
8bitmen.com

Federated Learning: Overview

Li *et al.* (2020)

- Groups:
  - the parties (e.g., clients)
  - the manager (e.g., server)
  - the communication-computation framework
- Components:
  - Data partitioning
  - Model
  - Privacy mechanism
  - Communication architecture
- Modelling approaches:
  - deep neural networks, gradient boosted decision trees, linear and logistic regression, support vector machines.





Cmu ML blog

Vantage6

Federated Learning Approaches

- Commentary:
  - Kairous *et al.* (2019) Advances and open problems in Federated Learning. *Arxiv.*
  - Li *et al.* (2020) A survey on federated learning systems: vision, hope and reality for data privacy and protection. *Arxiv*
- Repository with data-sharing agreements:
  - Canakoglu, A., *et al.* (2020) Federated sharing and processing of genomic datasets for tertiary data analysis. *Briefings in Bioinformatics*
- Bayesian Networks:
  - Jochems (2017) Survival prediction model through distributed learning across 3 countries. *Radiotherapy and Oncology.*
- Neural Networks:
  - Yurochkin (2019) Bayesian nonparametric federated learning of neural networks. *ICML.*
- Sequential and Hierarchical Bayesian models for time series data:
  - Fang *et al.* (2020) Bayesian Inference Federated Learning for Heart Rate Prediction. *International Conference on Intelligent Computing.*
- Communication-efficient surrogate likelihood (CSL):
  - Jordan, Lee, Yang (2018) Communication-efficient distributed statistical inference. *JASA.*

Case Study 1: Federated Matrix Factorisation

- Observation matrix $\boldsymbol{Y} \epsilon R^{i \times j}$
- MF: decompose into two latent matrices $\boldsymbol{U} \epsilon R^{i \times r}$ and $\boldsymbol{V} \epsilon R^{j \times r}$
- Solve $\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{V}^T + \boldsymbol{E}$
- SGD algorithm:
  - Formulate the regularised sum error (RSE) between $\widehat{\boldsymbol{Y}}$ and $\boldsymbol{Y}$ as

$$RSE = \sum_{i,j} [(y_{i,j} - u_i v_j^T)^2 + \lambda(\|u_i\|^2 + \|v_j\|^2)]$$

λ : regularisation coefficient

  - Update parameters by moving in the opposite direction of the gradient for each entry.
  - Represent error at step t-1 as

$$e_{i,j}^{t-1} = y_{i,j} - u_i^{t-1}(v_j^{t-1})^T$$

  - Obtain the gradient of the RSE over u_i^{t-1} and v_j^{t-1}
  - Update rule:

$$u_i^t = u_i^{t-1} + \beta g_i^{t-1}$$
$$v_j^t = v_j^{t-1} + \beta g_j^{t-1}$$

β : learning rate

Xie *et al.* (2021)

Adaptive learning rate

- Use the cosine of the angle between the learning directions of two consecutive epochs as an index to adaptively adjust the learning rate.
- For u_i (similarly for v_j):

$$g^t_{i(y_{i,j})} = -\frac{1}{2}\nabla_{u_i} = e^t_{i,j} v^{t-1}_j - \lambda u^{t-1}_i,$$

$$\cos\omega^t_{i(y_{i,j})} = (g^t_{i(y_{i,j})} \cdot g^{t-1}_{i(y_{i,j})}) \Big/ \left(\left\| g^t_{i(y_{i,j})} \right\| \cdot \left\| g^{t-1}_{i(y_{i,j})} \right\| \right),$$

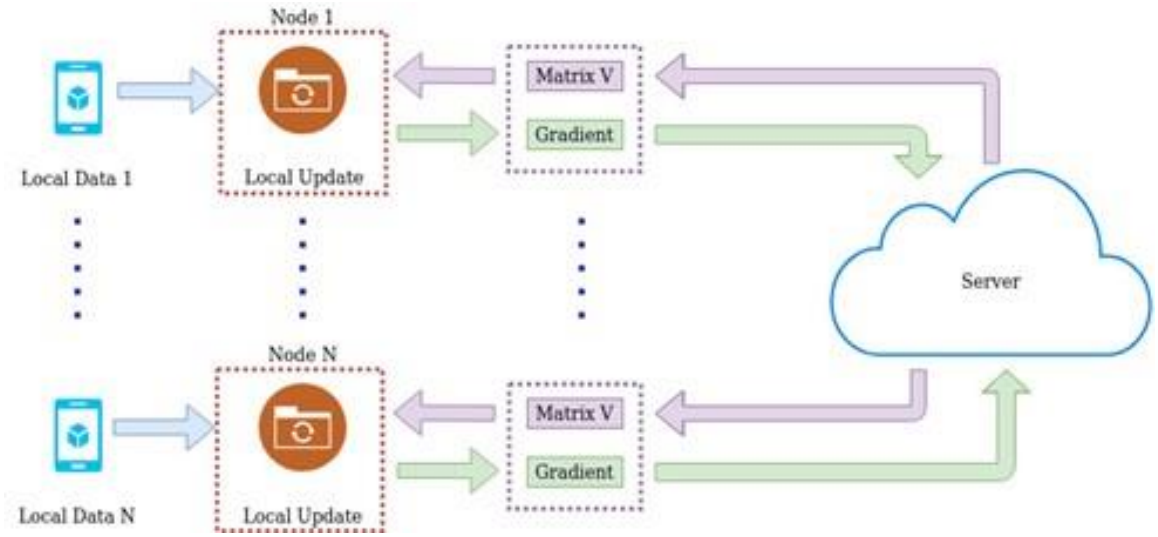
$$\beta^t_{i(y_{i,j})} = \beta^{t-1}_{i(y_{i,j})} (1 + \rho \cdot \cos\omega^t_{i(y_{i,j})}),$$

ρ : adjusts the fluctuation of the learning rate

$$u^t_i = u^{t-1}_i + \beta^t_{i(y_{i,j})} \cdot g^{t-1}_{i(y_{i,j})},$$

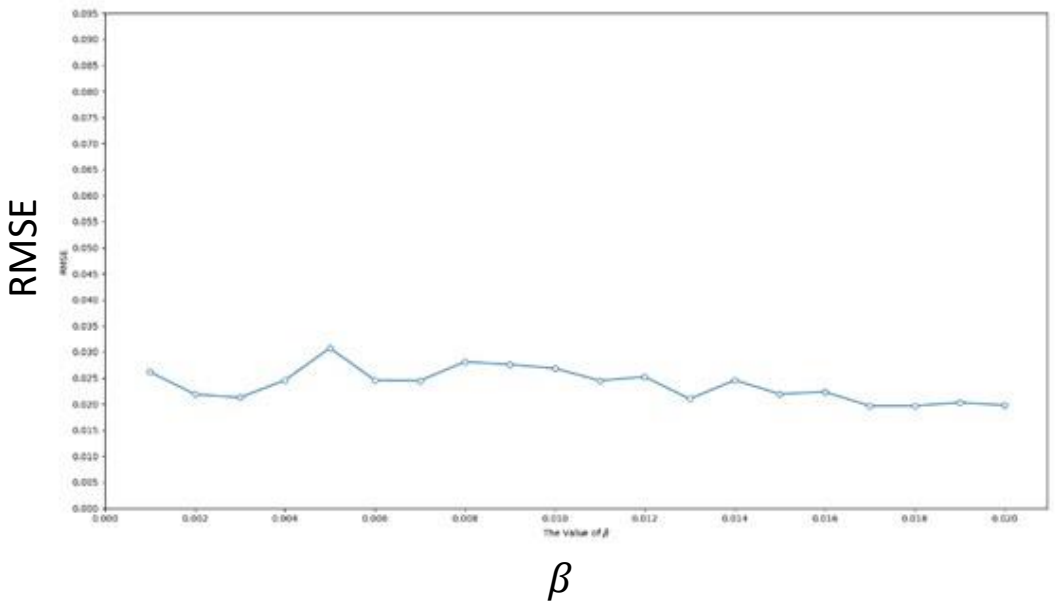
Federated Matrix Factorization

1. Node update: update u_i and gradient $g'_{j(y_{ij})}$ on each node *i*.
2. Upload information: upload encrypted $g'_{j(y_{ij})}$ to update v_j (two steps).
3. Server update: update **V** (encrypted).
4. Download information: each node downloads encrypted **V** from server and decrypts it to perform a new node update.
5. Stop when converged (small RSE).

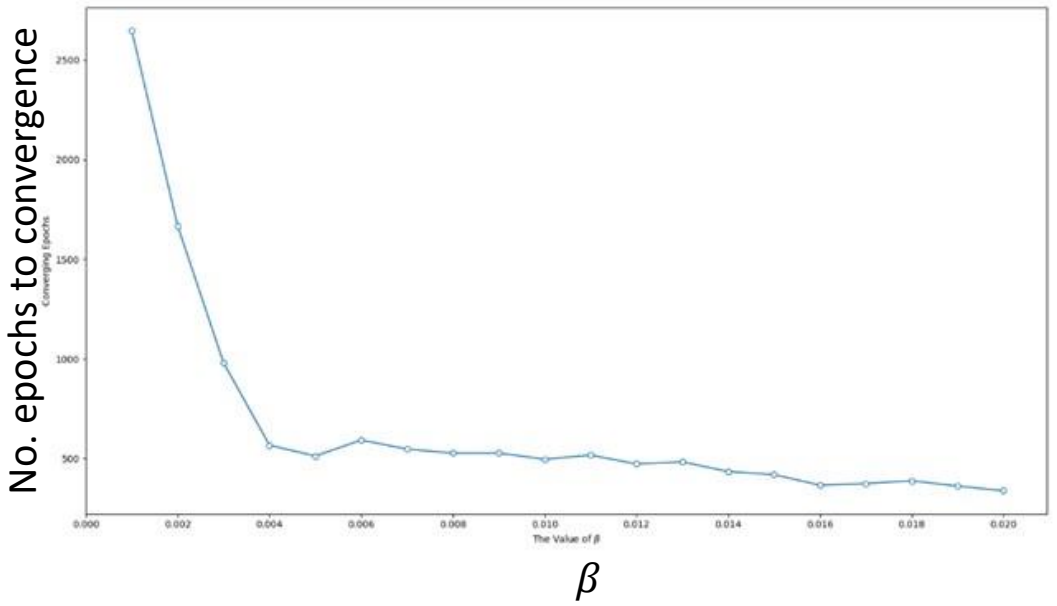


Results

RMSE vs learning rate

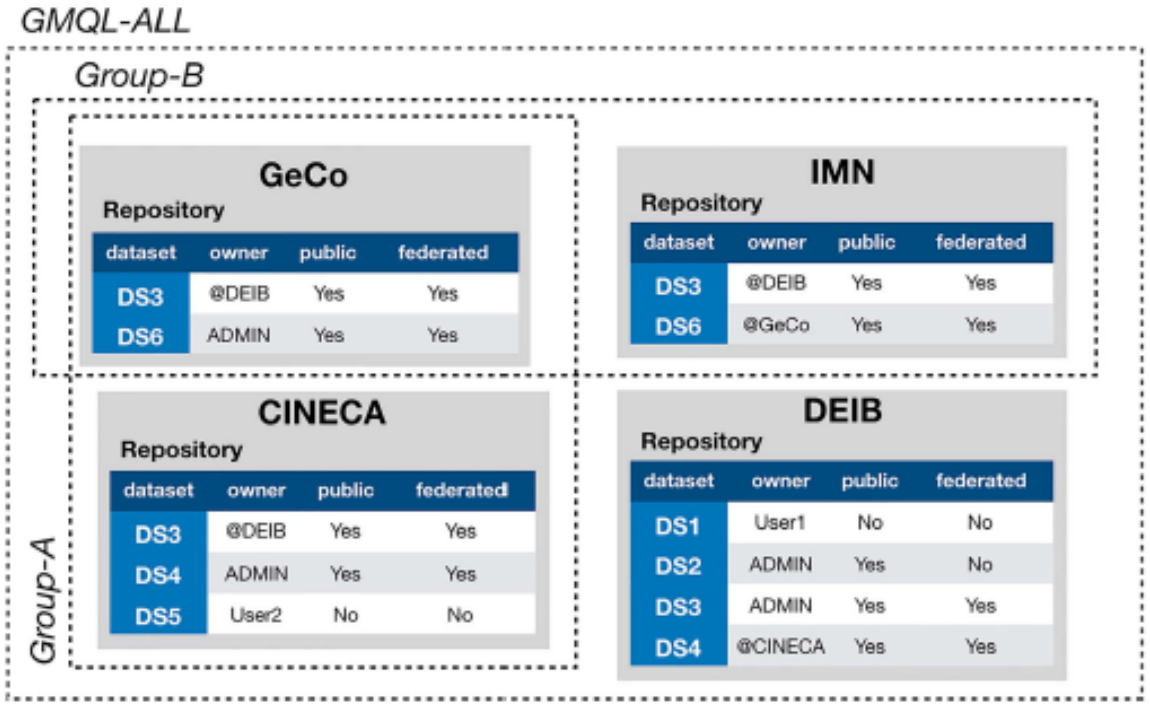


Convergence epochs vs learning rate



Case Study 3: Federated sharing of genomic datasets

- Huge growth of genomics data over last decade
- Many independent consortia and institutes



- Federated GMQL: web-based system for querying distributed genomics datasets across many instances connected through the Web.
- Automatically distribute the computation while preserving privacy constraints.

- Data sharing agreements across the Repository
- Facilitated by GMQL groups, each controlled by an administrator, who can dynamically add (or drop) GMQL instances to (from) the group and make agreements about the group

Canakoglu *et al.* (2020)

Summary

- Data are changing! Size, privacy, provenance, quality, diversity, …
- Federated analysis / federated learning offers some solutions.
- Current interest is in federated estimation of intrinsic dimension
- Statistical methods *and their implementation* are required!

References

CL Alston-Knox *et al.* (2018) Modelling habitat and planning surveillance using landsat imagery: a case study using imported red fire ants. Biological Invasions 20 (5), 1349 -1367.

H Battey *et al.* (2015) Distributed estimation and inference with statistical guarantees. arXiv preprint arXiv:1509.05457.

TS Brisimi *et al.* (2017) Federated learning of predictive models from federated Electronic Health Records. *International Journal of Medical Informatics.*

A Canakoglu *et al.* (2020) Federated sharing and processing of genomic datasets for tertiary data analysis. *Briefings in Bioinformatics.*

K Chang *et al.* (2018) Distributed deep learning networks among institutions for medical imaging. *Journal of the American Medical Informatics Association.*

M Deist *et al.* (2017) Infrastructure and distributed learning methodology for privacy-preserving multi-centric rapid learning health care: *euroCAT. Clinical and Translational Radiation Oncology.*

L Feng *et al.* (2020) Bayesian inference federated learning for heart rate prediction. *Conference paper.*

A Jochems *et al.* (2016) Distributed learning: developing a predictive model based on data from multiple hospitals without data leaving the hospital – a real life proof of concept. *Radioth. Onc.*

M Jordan *et al.* (2018) Communication-efficient distributed statistical inference. *JASA*114, 526, 68-681.

P Kairous *et al.* (2019) Advances and open problems in federated learning. arXiv:1912.04977.

JD Lee *et al.* (2015) Communication-efficient sparse regression: a one-shot approach. arXiv:1503.04337.

C Li *et al.* (2019) Image denoising based on nonlocal Bayesian singular value thresholding and Stein's unbiased risk estimator. *IEEE Transactions on Image Processing 28, 4899-4911*.

Q Li *et al.* (2020) A survey on federated learning systems: vision, hope and reality for data privacy and protection. *Arxiv*

C-L Lu *et al.* (2015) WebDISCO: a web service for distributed Cox model learning without patient-level data sharing. *J Am Med Inform Assoc* 22, 1212-1219.

A Moncada-Torres *et al.* (2020) VANTAGE6: an open source priVAcy preserviNg federaTed leArninG infrastructurE for Secure Insight eXchange. AMIA Annual Symp. Proc. vantage6

E Santos-Fernandez *et al.* (2020) Correcting misclassification errors in crowdsourced ecological data: a Bayesian perspective. *JRSS C – under revision.*

E Santos-Fernandez, KM (2020) Bayesian item response models for citizen science ecological data. *Methods in Ecology and Evolution – under revision.*

Z Stark *et al.* (2019) Australian genomics: a federated model for integrating genomics into healthcare. *The American Journal of Human Genetics.*

I Ullah, KM (2019) Bayesian mixture models and their Big Data implementations with application to invasive species presence-only data. *Journal of Big Data* 6 (1), 29.

M Yurochkin (2019) Bayesian nonparametric federated learning for neural networks. *ICML*. arXiv:1905.12022.

F Zerka *et al.* (2020) Systematic review of privacy-preserving distributed machine learning from federated databases in health care. *Clinical Cancer Informatics.*

References

M Allegra, E Facco, A Laio, A Mira (2019) Clustering by the local intrinsic dimension: the hidden structure of real-world data. Submitted.

CL Alston, KL Mengersen (2010) Allowing for the effect of data binning in a Bayesian Normal mixture model. Computational Statistics & Data Analysis, 54(4):916–923.

CL Alston, KL Mengersen, CP Robert, JM Thompson, PJ Littlefield, D Perry, AJ Ball (2007) Bayesian mixture models in a longitudinal setting for analysing sheep CAT scan images. Computational Statistics & Data Analysis, 51(9):4282–4296.

CL Alston, KL Mengersen, JM Thompson, PJ Littlefield, D Perry, AJ Ball (2005) Extending the Bayesian mixture model to incorporate spatial information in analysing sheep CAT scan images. Australian Journal of Agricultural Research, 56(4):373–388.

CL Alston, KL Mengersen, JM Thompson, PJ Littlefield, D Perry, AJ Ball (2004) Statistical analysis of sheep CAT scan images using a Bayesian mixture model. Australian Journal of Agricultural Research, 55(1):57–68.

CL Alston-Knox, KL Mengersen, R Denham, CM Strickland (2018) Modelling habitat and planning surveillance using landsat imagery: a case study using imported red fire ants. Biological Invasions 20 (5), 1349 -1367.

M Cespedes, J McGree, CC Drovandi, K Mengersen, LB Reid, JD Doecke, J Fripp (2017) A Bayesian hierarchical approach to jointly model structural biomarkers and covariance networks. Arxiv.

M Cespedes, J Fripp, JM McGree, CC Drovandi, K Mengersen, JD Doecke (2017) Comparison of neurodegeneration over time between health ageing and Alzheimer's disease cohorts via Bayesian inference. BMJ Open, 7(2).

MG Falk, CL Alston, CA McGrory, SJ Clifford, EA Heron, D Leonte, M Moores, C Walsh, TN Pettitt, KL Mengersen (2015) Recent Bayesian approaches for spatial analysis of 2-D images with application to environmental modelling, Environmental and Ecological Statistics, 22 (3), p571-600.

M Moores, G Nicholls, A Pettitt, K Mengersen (2018) Scalable Bayesian Inference for the Inverse Temperature of a Hidden Potts Model. Bayesian Analysis.

MT Moores, CE Hargrave, T Deegan, M Poulsen, F Harden, K Mengersen (2015) An external field prior for the hidden Potts model with application to cone-beam computed tomography. Computational Statistics and Data Analysis 86 27-41.

MT Moores, C Drovandi, K Mengersen, CP Robert (2015) Pre-processing for approximate Bayesian computation in image analysis. Statistics and Computing, 25(1), pp. 23-33.

CM Strickland, RL Burdett, R Denham, KL Mengersen (2014) PySSM : a Python module for Bayesian inference of linear Gaussian state space models. Journal of Statistical Software, 57(6), 1-37 J Holloway, KJ Helmstedt, K Mengersen, M Schmidt (2019) A decision tree approach for spatially interpolating missing land cover data and classifying satellite images. Remote Sensing, 11(15).

CM Strickland, DP Simpson, IW Turner, R Denham, KL Mengersen (2011) Fast Bayesian analysis of spatial dynamic factor models for multitemporal remotely sensed imagery. Journal of the Royal Statistical Society Series C–Applied Statistics, 60(Part 1):109–124.

CM Strickland, IW Turner, R Denham, KL Mengersen (2009) Efficient Bayesian estimation of multivariate state space models. Computational Statistics & Data Analysis, 53(12):4116–4125.

I Ullah, K Mengersen (2019) Bayesian mixture models and their Big Data implementations with application to invasive species presence-only data. Journal of Big Data 6 (1), 29.

C Li, H Xie, X Fan, RY Da Xu, S Van Huffel, SA Sisson, K Mengersen (2019) Image denoising based on nonlocal Bayesian singular value thresholding and Stein’s unbiased risk estimator. IEEE Transactions on Image Processing 28, 4899-4911.