Machine Learning 2: advanced

Andre F. Marquand
a.marquand@donders.ru.nl

Outline

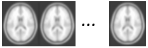
Outline

Types of pattern recognition

Supervised learning involves learning a mapping between input and output:

Input

Training examples
$\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$



$f(\mathbf{x},\mathbf{w}) \longrightarrow y$

Output

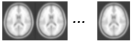
Target variables
$\{y_1, y_2, \ldots y_N\}$

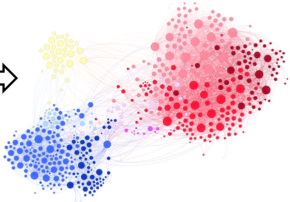
$y_i \in R$: regression
$y_i \in \{0,1\}$: classification

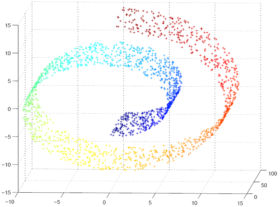
In **Unsupervised** learning, algorithms are not provided with output labels and must learn to structure the data by applying heuristics

Input

Training examples
$\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$

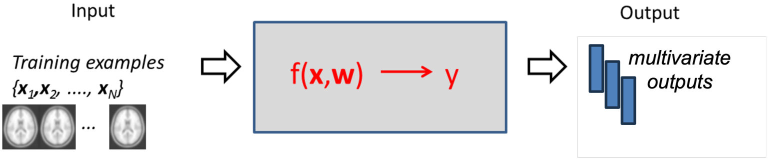






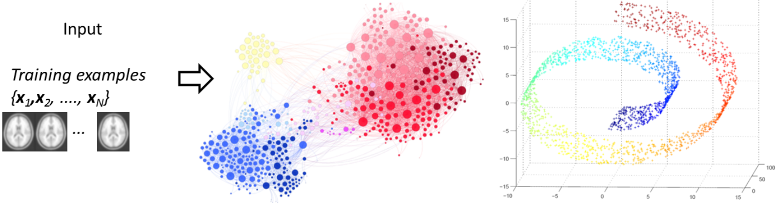
Types of pattern recognition

Supervised learning involves learning a mapping between input and output:



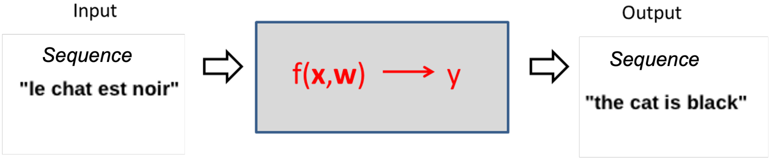
In **Unsupervised** learning, algorithms are not provided with output

labels and must learn to structure the data by applying heuristics



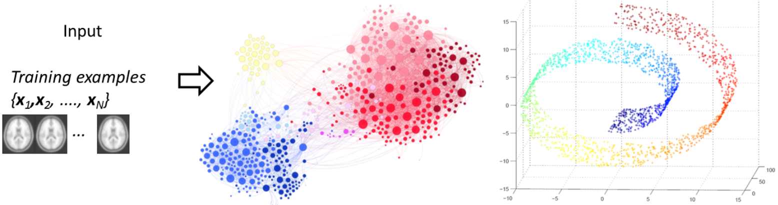
Types of pattern recognition

Supervised learning involves learning a mapping between input and output:

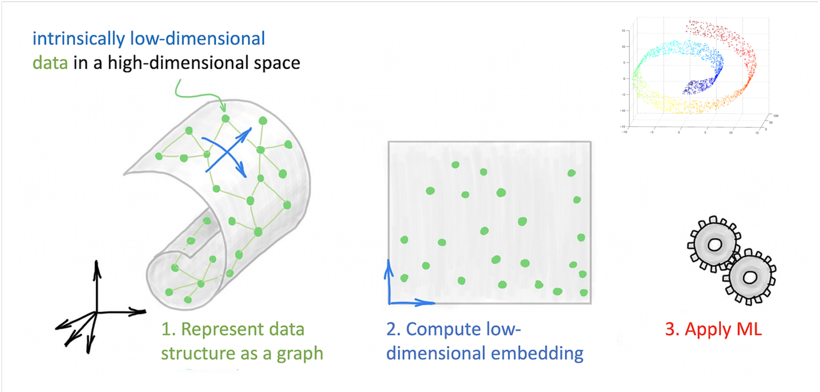


In **Unsupervised** learning, algorithms are not provided with output

labels and must learn to structure the data by applying heuristics



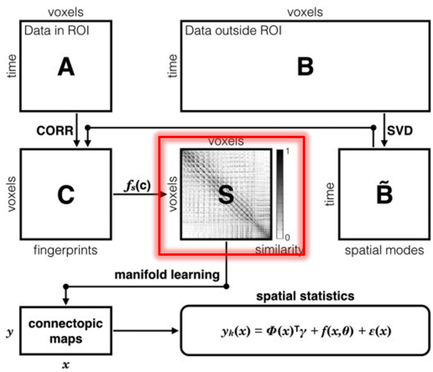
Manifold learning

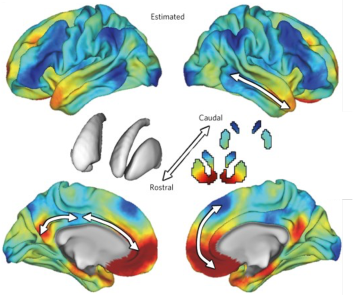


- Many different algorithms, diffusion embeding, isomap, Laplacian eigenmaps, autoencoders
- Most use a graph to represent relationships between data points

Image courtesy of Michael Bronstein

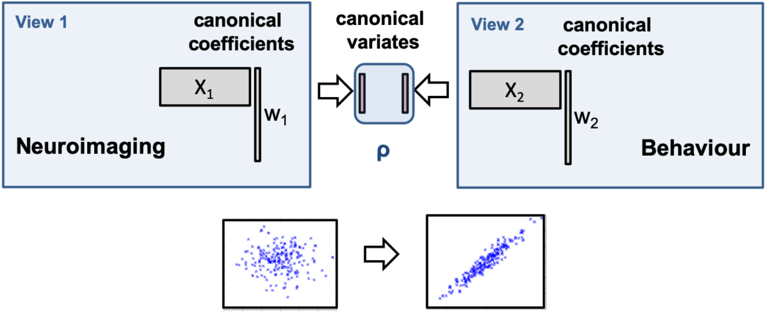
Connectopic mapping





Haak et al. (2016); Marquand et al. (2017)

- **Canonical Correlation Analysis** is a standard statistical tool for finding multivariate relationships between datasets
- Generalises Pearson correlation to multiple variables
- Finds projections of the data that maximise the correlation between "views" of the data



Canonical Correlation Analysis

- CCA is related to techniques such as partial least squares
- Formally, CCA solves the following objective function:

$$\max_{\mathrm{w}_1,\mathrm{w}_2} corr(\mathrm{X}_1\mathrm{w}_1, \mathrm{X}_2\mathrm{w}_2)$$

$$\text{subject to } ||\mathrm{w}_1^T\mathrm{X}_1^T\mathrm{X}_1\mathrm{w}_1|| \leq 1 \text{ and } ||\mathrm{w}_2^T\mathrm{X}_2^T\mathrm{X}_2\mathrm{w}_2|| \leq 1$$

where the constraint is sometimes amended to:

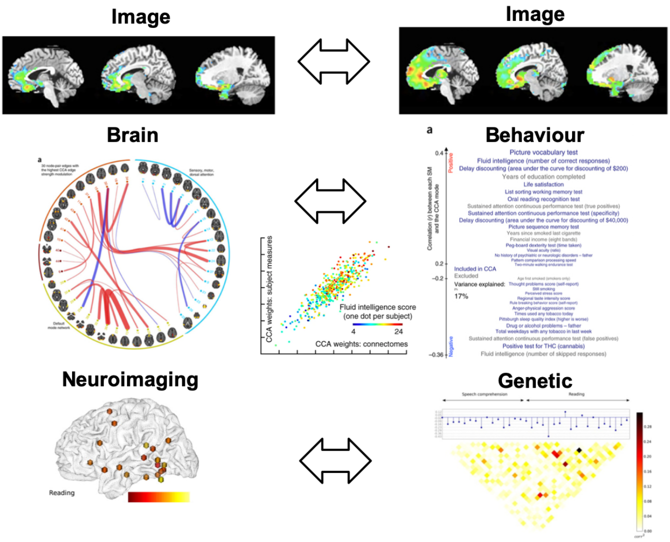
$$\text{subject to } ||\mathrm{w}_1||^2 \leq 1 \text{ and } ||\mathrm{w}_2||^2 \leq 1$$

...and other constraints can be added (e.g. to promote sparsity)

$$P(\mathrm{w}_1) < c_1 \text{ and } P(\mathrm{w}_2) < c_2$$

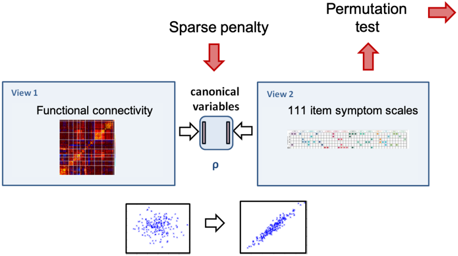
- if $n > p_1$ and p_2, an analytical solution is available
- There are many variants (kernel, Bayesian, deep CCA...)

Applications of CCA

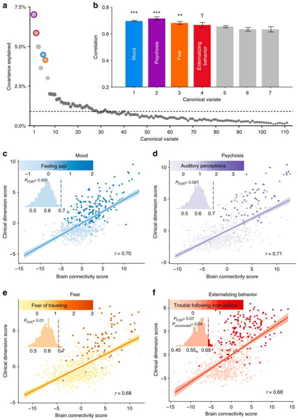


Smith et al. (2015); Floch et al. (2012)

Applications of CCA

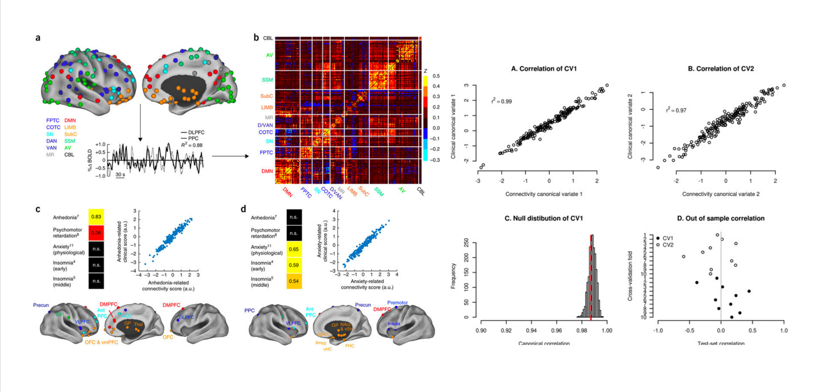


4 Linked dimensions



Xia et al. (2018)

Overfitting in CCA

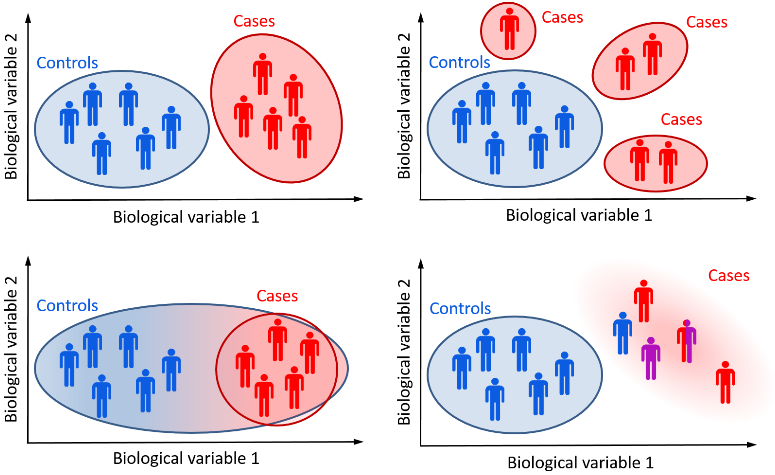


- CCA easily overfits, even when $N > p$
- In-sample canonical correlation is often high under the null!
- Regularization and/or feature selection is very important
- Statistical evaluation should include the whole pipeline

Drysdale et al. (2017); Dinga et al. (2019)

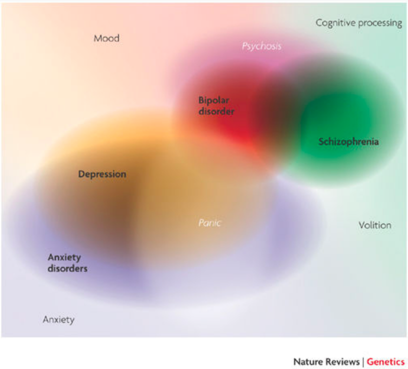
Outline

Many types of heterogeneity



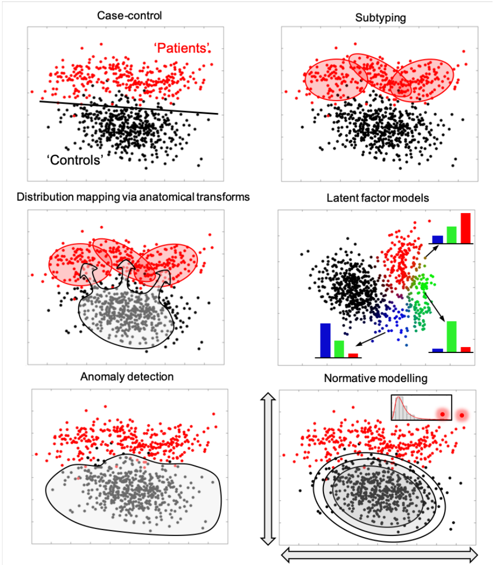
Marquand et al. (2016)

Many types of heterogeneity



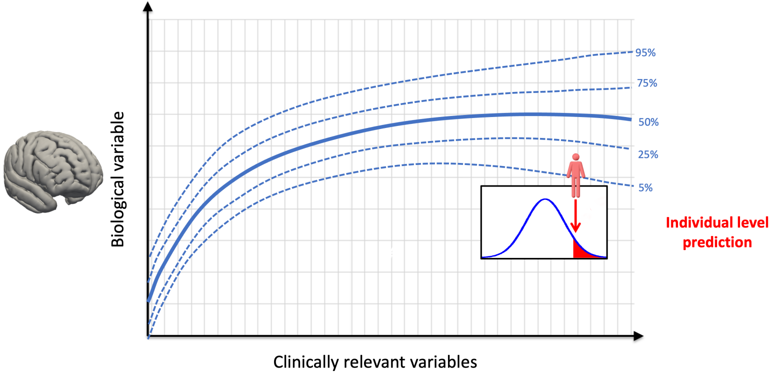
Burmeister et al. (2008)

Methods for addressing heterogeneity



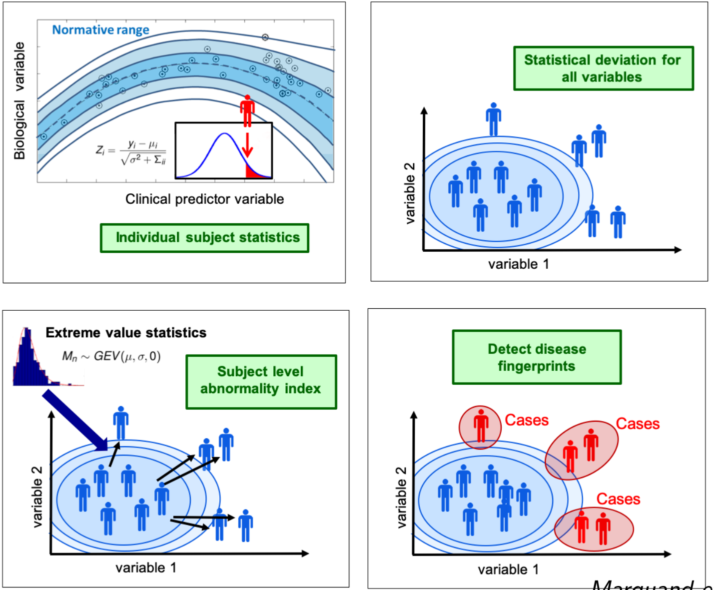
Dong et al. (2016); Zhang et al. (2016); Mourao-Miranda et al. (2011)

Normative modelling



Marquand et al. (2016)

Normative modelling

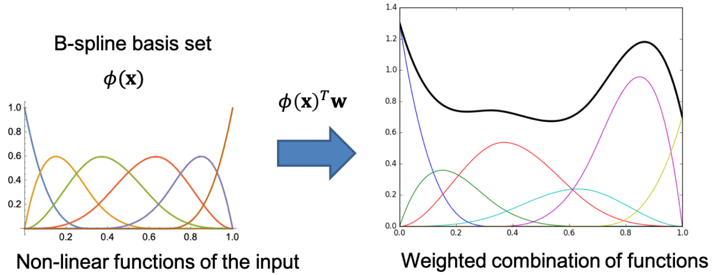


Marquand et al. (2016)

Modelling non-linearity using Bayesian 'linear' regression

- A simple way to model non-linearity is to use a pre-specified basis expansion Φ (e.g. polynomials, RBF, B-spline)

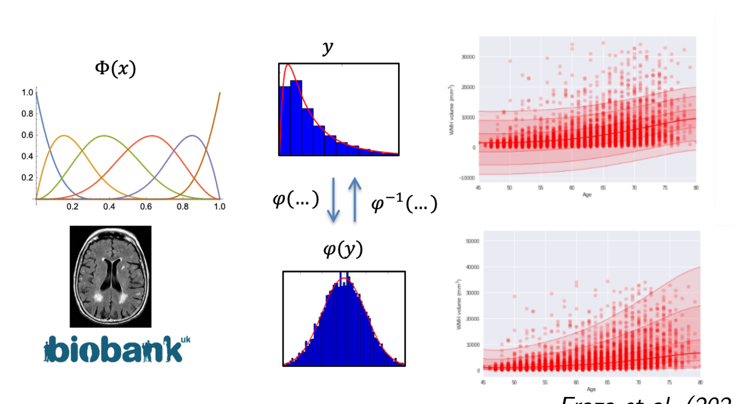
$$p(y|\boldsymbol{w}, \phi(\boldsymbol{x}), \sigma) \propto \mathcal{N}(y|\boldsymbol{\phi}(\boldsymbol{x})^T\boldsymbol{w}, \sigma)\mathcal{N}(\boldsymbol{w}|0, \boldsymbol{\Sigma_\theta})$$



Modelling non-Gaussianity

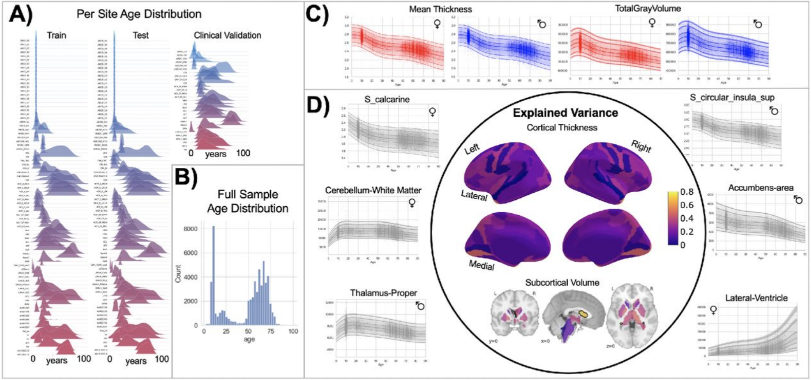
- One way to model non-Gaussian data is to 'warp' the data to a Gaussian space using a non-linear function $\varphi()$

$$p(\varphi(y)|\boldsymbol{w}, \phi(\boldsymbol{x}), \sigma) \propto \mathcal{N}(\varphi(y)|\phi(\boldsymbol{x})^T \boldsymbol{w}, \sigma)\mathcal{N}(\boldsymbol{w}|0, \boldsymbol{\Sigma_\theta})$$



Fraza et al. (2021)

Braincharts across the lifespan

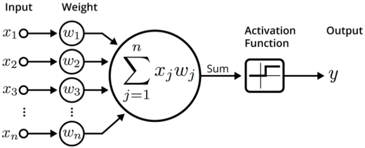


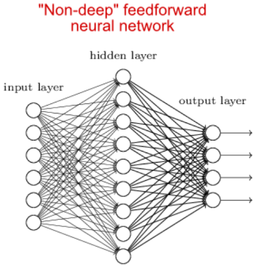
Tutorial: https://github.com/saigerutherford/CPC_ML_tutorial

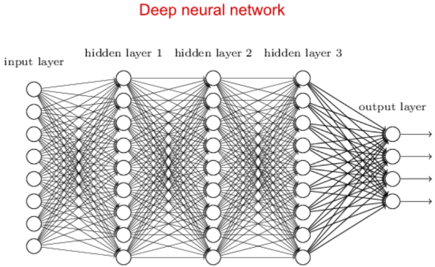
Rutherford et al. (2021a,b)

Outline

Deep Learning





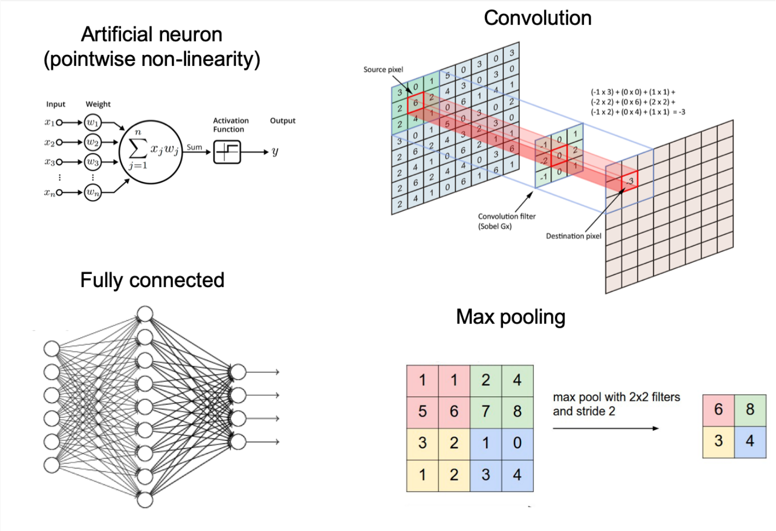


- Many variants but "convolutional" networks are popular
- Predominantly supervised learning
- Usually many parameters to optimise

Deep Learning

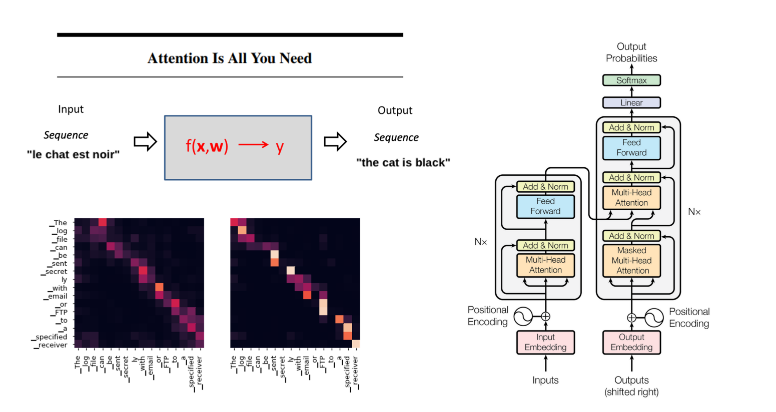
It is helpful to think of deep learning as combining matix products with point-wise linearity, e.g.:

$$f(x) = \sigma\left(xW_1 + b_1\right)W_2$$



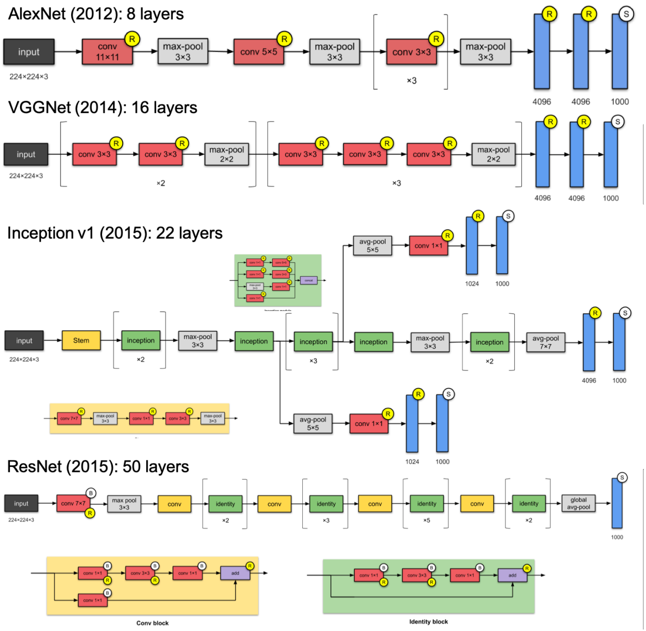
Transformers

- Transformers can model long range dependencies in sequence to sequence prediction
- Uses "attention" to focus on parts of the sequence



Transformers are at the heart of models like chatGPT

Neural Network Architectures



GPT-4 Technical Report

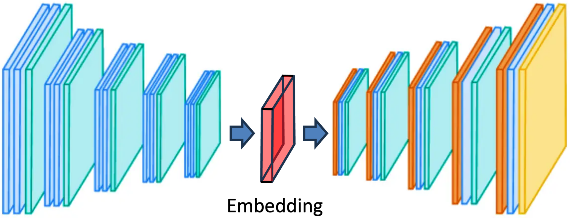
OpenAI*

Given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar.

Encoder-Decoder architectures

Encoder

Decoder



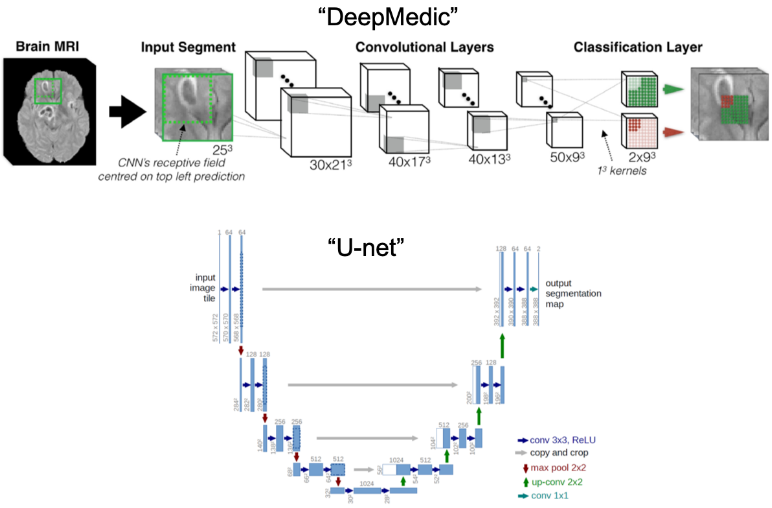
“Le chat is noir”

“The cat is black”

“A cute kitten riding a skateboard”

Zabihi et al. (2024); Serin et al. (2024)

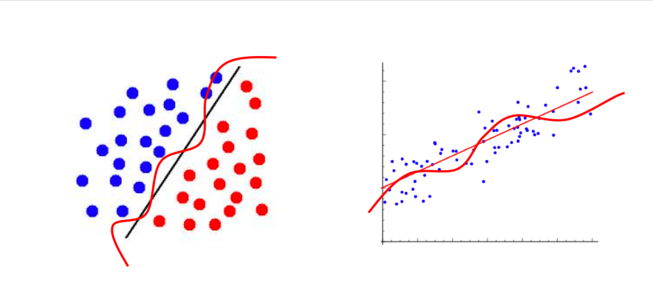
Deep Learning for Segmentation



Kamnitsas et al. (2017); Ronneberger et al. (2015)

Overfitting (again)

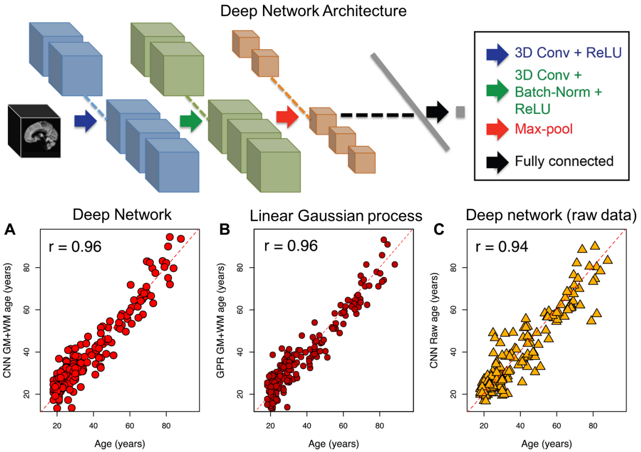
- But if your problem is linear, your fancy nonlinear algorithm will just overfit



- The more complex the model, the easier it is to overfit
- In complex (deep) models it is often not possible to properly optimise all parameters
- This makes validation extremely important!

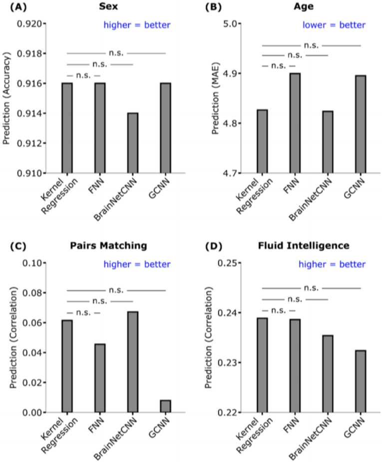
Deep Learning in Neuroscience?

- Predict age from N= 2001 structural MRI images



- Similar performance to a linear model on preprocessed data
- Better performance on minimally processed data *Cole et al. (2017)*

Deep Learning in Neuroscience?



He et al. (2019)

Deep Learning in Neuroscience?

New Results

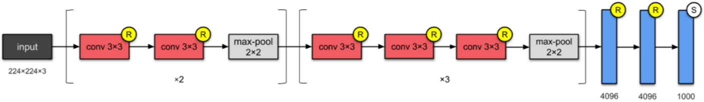
Comment on this paper

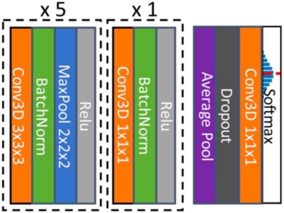
Accurate brain age prediction with lightweight deep neural networks

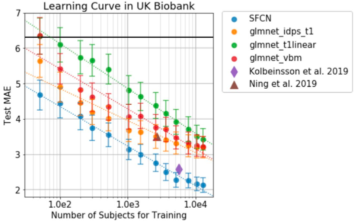
Han Peng, Weikang Gong, Christian F. Beckmann, Andrea Vedaldi, Stephen M. Smith

doi: https://doi.org/10.1101/2019.12.17.879346

This article is a preprint and has not been certified by peer review [what does this mean?].



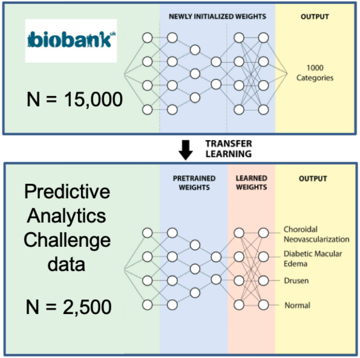




Won the 2019 PAC 'brainage' challenge

Peng et al. (2019)

Transfer Learning



- Basic idea: transfer knowledge (i.e. weights) from a large dataset to a small one (where it is harder to learn)
- Different variants depending on whether the targets are the same, similiar or different

Deep Learning in Medicine?

- Predict mortality from electronic health records

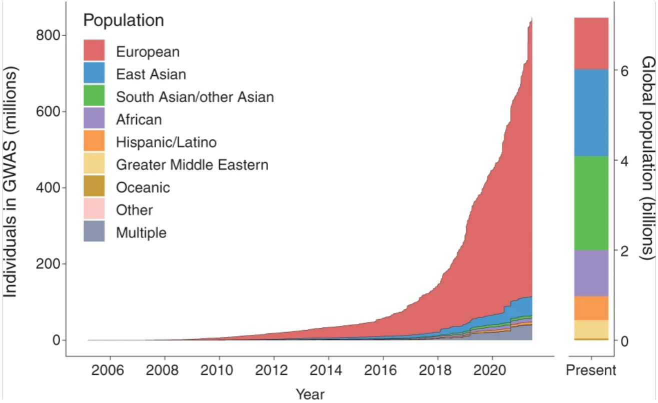
ARTICLE OPEN

Scalable and accurate deep learning with electronic health records

Alvin Rajkomar[1,2], Eyal Oren[1], Kai Chen[1], Andrew M. Dai[1], Nissan Hajaj[1], Michaela Hardt[1], Peter J. Liu[1], Xiaobing Liu[1], Jake Marcus[1], Mimi Sun[1], Patrik Sundberg[1], Hector Yee[1], Kun Zhang[1], Yi Zhang[1], Gerardo Flores[1], Gavin E. Duggan[1], Jamie Irvine[1], Quoc Le[1], Kurt Litsch[1], Alexander Mossin[1], Justin Tansuwan[1], De Wang[1], James Wexler[1], Jimbo Wilson[1], Dana Ludwig[2], Samuel L. Volchenboum[3], Katherine Chou[1], Michael Pearson[1], Srinivasan Madabushi[1], Nigam H. Shah[4], Atul J. Butte[2], Michael D. Howell[1], Claire Cui[1], Greg S. Corrado[1] and Jeffrey Dean[1]

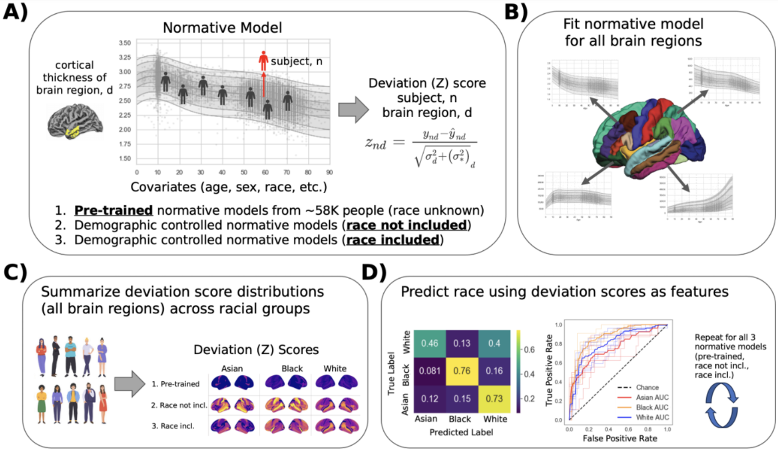
	Hospital A	Hospital B
Inpatient Mortality, AUROC[1](95% CI)		
Deep learning 24 hours after admission	**0.95**(0.94-0.96)	**0.93**(0.92-0.94)
Full feature enhanced baseline at 24 hours after admission	0.93 (0.92-0.95)	0.91 (0.89-0.92)
Full feature simple baseline at 24 hours after admission	0.93 (0.91-0.94)	0.90 (0.88-0.92)
Baseline (aEWS[2]) at 24 hours after admission	0.85 (0.81-0.89)	0.86 (0.83-0.88)
30-day Readmission, AUROC (95% CI)		
Deep learning at discharge	**0.77**(0.75-0.78)	**0.76**(0.75-0.77)
Full feature enhanced baseline at discharge	0.75 (0.73-0.76)	0.75 (0.74-0.76)
Full feature simple baseline at discharge	0.74 (0.73-0.76)	0.73 (0.72-0.74)
Baseline (mHOSPITAL[3]) at discharge	0.70 (0.68-0.72)	0.68 (0.67-0.69)
Length of Stay at least 7 days AUROC (95% CI)		
Deep learning 24 hours after admission	**0.86**(0.86-0.87)	**0.85**(0.85-0.86)
Full feature enhanced baseline at 24 hours after admission	0.85 (0.84-0.85)	0.83 (0.83-0.84)
Full feature simple baseline at 24 hours after admission	0.83 (0.82-0.84)	0.81 (0.80-0.82)
Baseline (mLiu[4]) at 24 hours after admission	0.76 (0.75-0.77)	0.74 (0.73-0.75)

Be careful about hidden bias!



- There are many examples in healthcare showing biased predictions for marginalised groups
  (e.g. African-Americans having a lower probablity of being assigned pain medications)

Be careful about hidden bias!



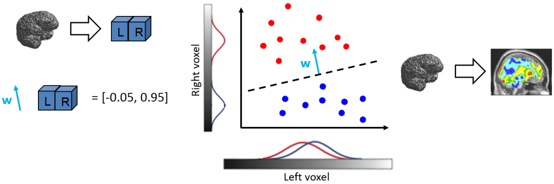
Rutherford et al. (2024)

Outline

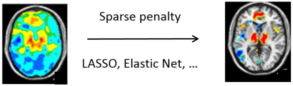
Mapping the discriminative pattern

For clinical applications it is crucial to infer which variables drive the predictions. There are multiple options:

- Regional classification accuracy as a proxy (searchlight)
- Explainable AI (relevance propagation, deconvolution, salience, occlusion, ...)
- For linear models, the weights can be directly visualised



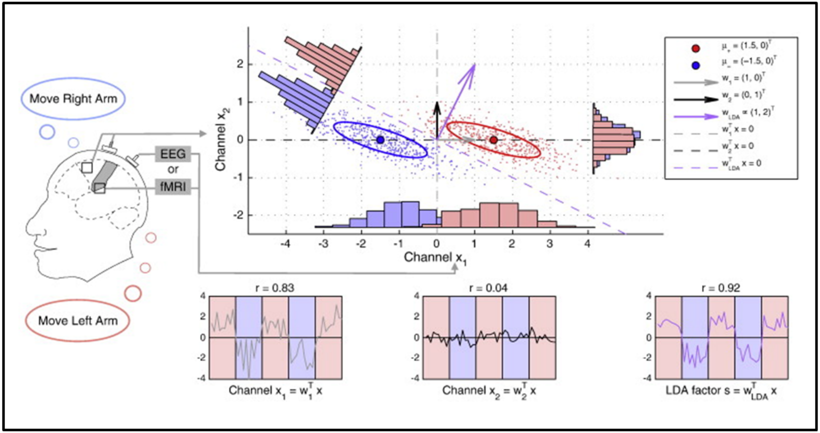
- Can also use regularization to enforce sparsity



- Weights are often regarded as being difficult to interpret, but this is not always true

Weights do not reflect univariate differences

One proposal is to consider the the weights from a forward model



Construct 'forward maps' by premultiplying by the data covariance

$$a = \frac{1}{\sigma_y^2} \Sigma_x w$$

Haufe et al. (2014)

Understanding weights of discriminative models

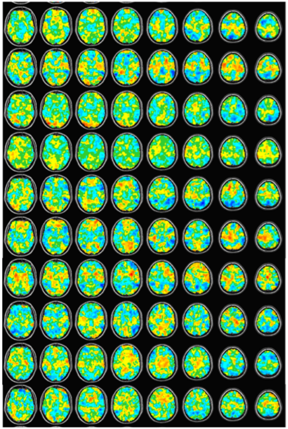
- The correct interpretation of the weights is the contribution of each feature to the predictions. This is the same as in a GLM
- Difficulty arises only due to multicollinearity between predictor variables which inflates the variance of the weights
- A variable can have a high weight because:
  1. It is associated with the response variable
  2. It acts as a 'supressor' variable that helps to cancel out noise or mismatch in other covariates
- To distinguish between these possibilities, we can compute:

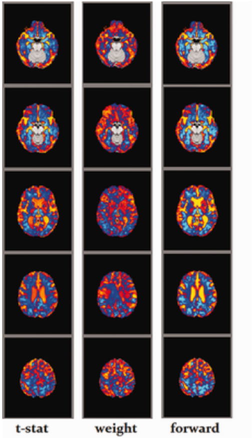
$$\mathrm{a} \propto \Sigma_x \mathrm{w} = \mathrm{cov}\left[\mathrm{X}, \hat{\mathrm{y}}\right] = \mathrm{corr}\left[\mathrm{X}, \hat{\mathrm{y}}\right]$$

- These are **structure coefficients** from multivariate statistics
- These measure the univariate association between covariate p and the predictions, i.e.: $\rho(\mathrm{x}_p, \hat{\mathrm{y}})$
- The weights are influenced by collinearity which impact the magnitude and sign of coefficients

Kraha et al. (2012)

Examples of weights





Aksman et al. (2016)

Outline

Take home messages

- Machine learning provides powerful tools for single subject inference and detect spatially distributed effects
- Many different approaches beyond simple notions such as 'classification' or 'clustering'

Recommendations

- Linear models are often sufficient: they are fast, interpretable and often perform as well as non-linear methods
- Careful validation is extremely important for all methods to guard agains overfitting
- Machine learning can be easily integrated with neurocognitive models (e.g. to assess candidate models)
- Be aware of potential biases! With complex models it is not possible to assess this definititvely

Acknowledgments

Mariam Zabihi

Richard Dinga

Barbora Rehák Bučková

Saige Rutherford

Charlotte Fraza

Predictive Clinical Neuroscience Lab

Hannah Savage

Tutorial on Normative modelling:

`https://github.com/saigerutherford/CPC_ML_tutorial`

References I

Leon Aksman, David J. Lythgoe, Steven C.R. Williams, Martha Jokisch, Christoph Moenninghoff, Johannes Streffer, Karl Heinz Joeckel, Christian Weimar, and Andre F. Marquand. Making use of longitudinal information in pattern recognition. *Human Brain Mapping*, 37(12):4385–4404, 2016. doi: 10.1002/hbm.23317. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/hbm.23317.

James Cole, Rudra Poudel, Dimosthenis Tsagkasoullis, Matthan Caan, Claire Steves, Tim Spector, and Giovanni Montana. Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker. *Neuroimage*, (In Press), 2017.

Richard Dinga, Lianne Schmaal, Brenda W.J.H. Penninx, Marie Jose van Tol, Dick J. Veltman, Laura van Velzen, Maarten Mennes, Nic J.A. van der Wee, and Andre F. Marquand. Evaluating the evidence for biotypes of depression: Methodological replication and extension of drysdale et al. (2017). *NeuroImage: Clinical*, 22: 101796, 2019. ISSN 2213-1582. doi: https://doi.org/10.1016/j.nicl.2019.101796. URL http://www.sciencedirect.com/science/article/pii/S2213158219301469.

A. Dong, N. Honnorat, B. Gaonkar, and C. Davatzikos. Chimera: Clustering of heterogeneous disease effects via distribution matching of imaging patterns. *IEEE Transactions on Medical Imaging*, 35(2):612–621, Feb 2016. ISSN 0278-0062. doi: 10.1109/TMI.2015.2487423.

Andrew T Drysdale, Logan Grosenick, Jonathan Downar, Katharine Dunlop, Farrokh Mansouri, Yue Meng, Robert N Fetcho, Benjamin Zebley, Desmond J Oathes, Amit Etkin, Alan F Schatzberg, Keith Sudheimer, Jennifer Keller, Helen S Mayberg, Faith M Gunning, George S Alexopoulos, Michael D Fox, Alvaro Pascual-Leone, Henning U Voss, BJ Casey, Marc J Dubin, and Conor Liston. Resting-state connectivity biomarkers define neurophysiological subtypes of depression. *Nature Medicine*, (23):28–38, 2017.

Edith Le Floch, Vincent Guillemot, Vincent Frouin, Philippe Pinel, Christophe Lalanne, Laura Trinchera, Arthur Tenenhaus, Antonio Moreno, Monica Zilbovicius, Thomas Bourgeron, Stanislas Dehaene, Bertrand Thirion, Jean-Baptiste Poline, and Edouard Duchesnay. Significant correlation between a set of genetic polymorphisms and a functional brain network revealed by feature selection and sparse partial least squares. *NeuroImage*, 63 (1):11 – 24, 2012. ISSN 1053-8119.

Charlotte J. Fraza, Richard Dinga, Christian F. Beckmann, and Andre F. Marquand. Warped bayesian linear regression for normative modelling of big data. *bioRxiv*, 2021. doi: 10.1101/2021.04.05.438429. URL https://www.biorxiv.org/content/early/2021/04/06/2021.04.05.438429.

K. V. Haak, A. F. Marquand, and C. F Beckmann. Connectopic mapping with resting fmri. 1602.07100, 2016.

References II

Stefan Haufe, Frank Meinecke, Kai Goergen, Sven Daehne, John-Dylan Haynes, Benjamin Blankertz, and Felix Biessmann. On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage*, 87(0):96 – 110, 2014.

Tong He, Ru Kong, Avram Holmes, Minh Nguyen, Mert Sabuncu, Simon Eickhoff, Danilo Bzdok, Jiashi Feng, and B. T. Thomas Yeo. Deep neural networks and kernel regression achieve comparable accuracies for functional connectivity prediction of behavior and demographics. *BioRxiv*, 2019.

Konstantinos Kamnitsas, Christian Ledig, Virginia F.J. Newcombe, Joanna P. Simpson, Andrew D. Kane, David K. Menon, Daniel Rueckert, and Ben Glocker. Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. *Medical Image Analysis*, 36:61–78, 2017. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2016.10.004. URL https://www.sciencedirect.com/science/article/pii/S1361841516301839.

Amanda Kraha, Heather Turner, Kim Nimon, Linda Zientek, and Robin Henson. Tools to support interpreting multiple regression in the face of multicollinearity. *Frontiers in Psychology*, 3:44, 2012. ISSN 1664-1078. doi: 10.3389/fpsyg.2012.00044. URL https://www.frontiersin.org/article/10.3389/fpsyg.2012.00044.

Andre Marquand, Koen Haak, and Christian Beckmann. Functional corticostriatal connection topographies predict goal-directed behaviour in humans. 1(8):0146, 2017. doi: 10.1038/s41562-017-0146DO-10.1038/s41562-017-0146.

Andre F. Marquand, Iead Rezek, Jan Buitelaar, and Christian F. Beckmann. Understanding heterogeneity in clinical cohorts using normative models: beyond case control studies. In Press, 2016.

Janaina Mourao-Miranda, David R. Hardoon, Tim Hahn, Andre F. Marquand, Steve C.R. Williams, John Shawe-Taylor, and Michael Brammer. Patient classification as an outlier detection problem: An application of the one-class support vector machine. *NeuroImage*, 58(3):793 – 804, 2011. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2011.06.042. URL http://www.sciencedirect.com/science/article/pii/S1053811911006872.

Han Peng, Weikang Gong, Christian F. Beckmann, Andrea Vedaldi, and Stephen M. Smith. Accurate brain age prediction with lightweight deep neural networks. *bioRxiv*, 2019. doi: 10.1101/2019.12.17.879346. URL https://www.biorxiv.org/content/early/2019/12/18/2019.12.17.879346.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

Saige Rutherford, Charlotte Fraza, Richard Dinga, Seyed Mostafa Kia, Thomas Wolfers, Mariam Zabihi, Pierre Berthet, Amanda Worker, Serena Verdi, Derek Andrews, Laura Han, Johanna Bayer, Paola Dazzan, Phillip McGuire, Roel T. Mocking, Aaart Schene, Brenda W. Pennix, Chandra Sripada, Ivy F. Tso, Elizabeth R. Duval, Soo-Eun Chang, Mary Heitzeg, S. Alexandra Burt, Luke Hyde, David Amaral, Christine Wu Nordahl, Ole A. Andreasssen, Lars T. Westlye, Roland Zahn, Henricus G. Ruhe, Christian Beckmann, and Andre F. Marquand. Charting brain growth and aging at high spatial precision. *bioRxiv*, 2021a. doi: 10.1101/2021.08.08.455487. URL https://www.biorxiv.org/content/early/2021/08/09/2021.08.08.455487.

Saige Rutherford, Seyed Mostafa Kia, Thomas Wolfers, Charlotte Fraza, Mariam Zabihi, Richard Dinga, Pierre Berthet, Amanda Worker, Serena Verdi, Henricus G. Ruhe, Christian F. Beckmann, and Andre F. Marquand. The normative modeling framework for computational psychiatry. *bioRxiv*, 2021b. doi: 10.1101/2021.08.08.455583. URL https://www.biorxiv.org/content/early/2021/08/10/2021.08.08.455583.

Saige Rutherford, Thomas Wolfers, Charlotte Fraza, Nathaniel G. Harrnet, Christian F. Beckmann, Henricus G. Ruhe, and Andre F. Marquand. To which reference class do you belong? measuring racial fairness of reference classes with normative modeling, 2024. URL https://arxiv.org/abs/2407.19114.

Emin Serin, Kerstin Ritter, Gunter Schumann, Tobias Banaschewski, Andre Marquand, and Henrik Walter. Generating synthetic task-based brain fingerprints for population neuroscience using deep learning. *bioRxiv*, 2024. doi: 10.1101/2024.08.03.606469. URL https://www.biorxiv.org/content/early/2024/08/08/2024.08.03.606469.

Stephen M Smith, Thomas E Nichols, Diego Vidaurre, Anderson M Winkler, Timothy E J Behrens, Matthew F Glasser, Kamil Ugurbil, Deanna M Barch, David C Van Essen, and Karla L Miller. A positive-negative mode of population covariation links brain connectivity, demographics and behavior. 18:1565–1567, 2015.

Cedric Xia, Zongming Ma, Rastko Ciric, Shi Gu, Richard F. Betzel, Antonia N. Kaczkurkin, Monica E. Calkins, Philip A. Cook, Angel Garcia de la Garza, Simon N. Vandekar, Zaixu Cui, Tyler M. Moore, David R. Roalf, Kosha Ruparel, Daniel H. Wolf, Christos Davatzikos, Ruben C. Gur, Raquel E. Gur, Russell T. Shinohara, Danielle S. Bassett, and Theodore D. Satterthwaite. Linked dimensions of psychopathology and connectivity in functional brain networks. *Nature Communications*, 9, 2018.

References IV

Mariam Zabihi, Seyed Mostafa Kia, Thomas Wolfers, Stijn de Boer, Charlotte Fraza, Richard Dinga, Alberto Llera Arenas, Danilo Bzdok, Christian F. Beckmann, and Andre Marquand. Nonlinear latent representations of high-dimensional task-fmri data: Unveiling cognitive and behavioral insights in heterogeneous spatial maps. *PLOS ONE*, 19(8):1–21, 08 2024. doi: 10.1371/journal.pone.0308329. URL https://doi.org/10.1371/journal.pone.0308329.

Xiuming Zhang, Elizabeth C. Mormino, Nanbo Sun, Reisa A. Sperling, Mert R. Sabuncu, and B. T. Thomas Yeo. Bayesian model reveals latent atrophy factors with dissociable cognitive trajectories in alzheimer's disease. *Proceedings of the National Academy of Sciences*, 113(42):E6535–E6544, 2016. ISSN 0027-8424. doi: 10.1073/pnas.1611073113. URL https://www.pnas.org/content/113/42/E6535.

But what about the penalty?

- Collinearity is well-known in classical GLM settings, where advice is often given to avoid collinearity
- It is true that collinearity impacts on efficiency, but models with collinear predictors are still interpretable
- Collinearity also impacts penalised regression. Recall that:

$$f(\mathrm{x}_i, \mathrm{w}) = \mathrm{x}_i^T \mathrm{w} \qquad \Rightarrow \hat{\mathrm{w}} = \ \min_{\mathrm{w}} \sum_{i=1}^{n} \ell(y_i, f_i) + \lambda J(\mathrm{w})$$

- Considering ridge regression, where the objective function is:

$$\hat{\mathrm{w}} = \ \min_{\mathrm{w}} \sum_{i=1}^{n} \mathcal{N}(\mathrm{w}^T \mathrm{x}_i, \sigma^2) + \frac{\lambda}{2} ||\mathrm{w}||_2^2$$

- This is eqivalent to:

$$\min_{\mathrm{w}} -\frac{1}{2\sigma^2} \left(\sum_{i=1}^{n} (y_i - \mathrm{x}_i^T \mathrm{w})^2 + \lambda \mathrm{w}^T \mathrm{w} \right)$$

But what about the penalty?

- This is exactly equivalent to finding the MAP estimate of a posterior distribution over w, with prior:

$$\mathcal{N}(0, \sigma^2/\lambda \mathsf{I})$$

So what does this mean?

- Collinearity influence the magnitude and sign of coefficients
- Including/excluding variables can change both magnitude and sign of coefficients
- When $p > n$, the problem is ill-posed (multiple ways the same prediction can be achieved).
- Regularisation helps to stabilise coefficients, but does not eliminate the problem