Extracting Features from sMRI

John Ashburner

Wellcome Centre for Human Neuroimaging,
UCL Queen Square Institute of Neurology,
12 Queen Square,
London WC1N 3BG, UK.

Feature engineering

First-timers are often surprised by how **little time in a machine learning project is spent actually doing machine learning**. *But it makes sense if you consider how time-consuming it is to gather data, integrate it, clean it and pre-process it, and how much* **trial and error can go into feature design**. *Also, machine learning is not a one-shot process of building a data set and running a learner, but rather an iterative process of running the learner, analyzing the results, modifying the data and/or the learner, and repeating. Learning is often the quickest part of this, but that's because we've already mastered it pretty well!* **Feature engineering is more difficult because it's domain-specific, while learners can be largely general-purpose.** *However, there is no sharp frontier between the two, and this is another reason* **the most useful learners are those that facilitate incorporating knowledge**.

Domingos, Pedro. "A few useful things to know about machine learning." Communications of the ACM 55, no. 10 (2012): 78-87.

Accuracy

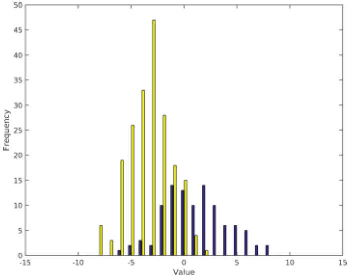
- Proportion of guesses that are correct. Assessed by cross-validation.
- A very simple measure of generalisation. Very noisy.

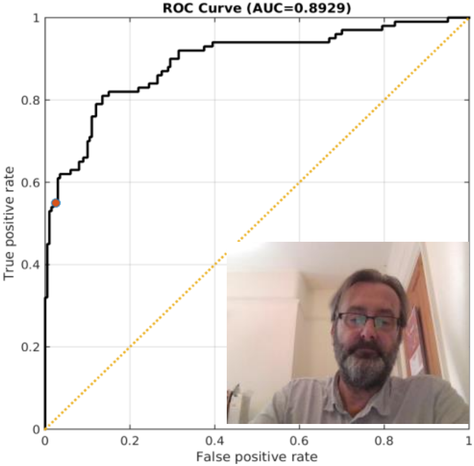
If 90% of subjects are controls and 10% are patients, then guessing that everyone is a control will give 90% accuracy.

Area Under the Curce (AUC)

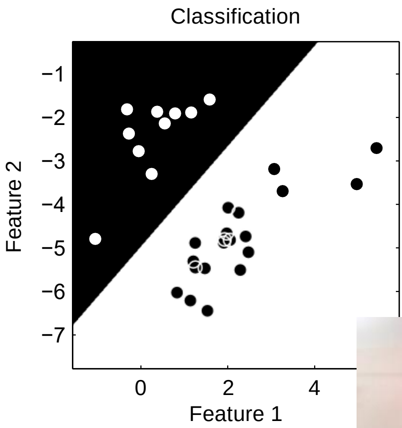
Area under the Receiver Operating Characteristic (ROC) curve.

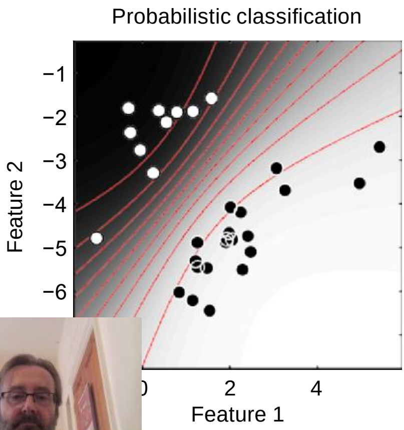
Assessed by cross-validation.





Hard v Probabilistic Classification





Target Information

Using cross-validation with binary classification, the average number of *bits* of information obtained for each subject is:

$$I = \frac{1}{N}\sum_{n=1}^{N}\left(t_n \log_2 p_n + (1-t_n)\log_2(1-p_n)\right) - \left(\bar{t}\log_2 \bar{t}^* + (1-\bar{t})\log_2(1-\bar{t}^*)\right)$$

where t_n is the label of the nth test subject (0 or 1)

p_n is the predicted probability for the nth test subject

$\bar{t}^*$ is the average of the labels of the training data.

A similar scheme may be used for regression, where information is given in *nats* (used $\log_e$, rather than $\log_2$).

Log Marginal Likelihood (ELBO)

Bayesian methods give a measure known as log marginal likelihood.

$$P(\mathbf{y}|\mathbf{X}) = \int_{\mathbf{w}} P(\mathbf{y}|\mathbf{X}, \mathbf{w})p(\mathbf{w})d\mathbf{w}$$

- An established Bayesian model selection approach (see papers by David MacKay and others).
- Does not involve cross-validation.
- Not trusted by some machine learning people.

No Free Ducklings

No Free Lunch theorem says that learning is impossible without prior knowledge.
http://en.wikipedia.org/wiki/No_free_lunch_in_search_and_optimization

Ugly Duckling theorem says that things are all equivalently similar to each other without prior knowledge.
http://en.wikipedia.org/wiki/Ugly_duckling_theorem

By Ryan Ebert from Portland, US (Flickr) [CC BY 2.0], via Wikimedia Commons. https://creativecommons.org/licenses/by/2.0/

What prior knowledge do we have about variability among people that can be measured using MRI?
How do we use this knowledge?

Incorporating Prior Knowledge into Kernels

Linear kernel matrices are often computed from the raw features:

$$\mathbf{K} = \mathbf{X}\mathbf{X}^T$$

A simple spatial feature selection may be considered as the following, where $\boldsymbol{\Sigma}_0$ is a (scaled) diagonal matrix of ones and zeros:

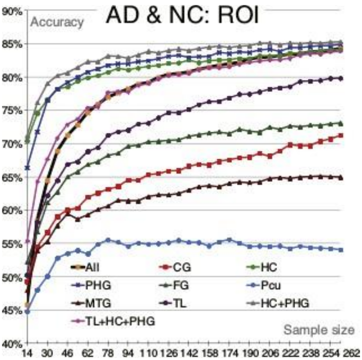
$$\mathbf{K} = \mathbf{X}\boldsymbol{\Sigma}_0\mathbf{X}^T$$

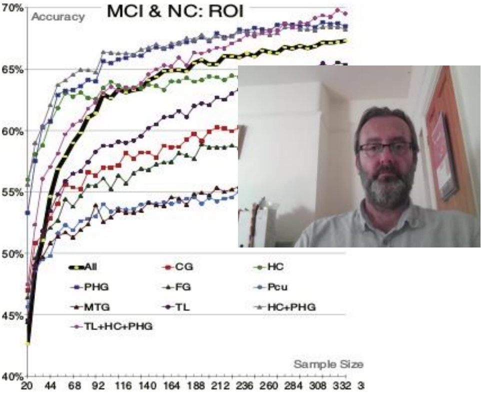
$\boldsymbol{\Sigma}_0$ may be more complicated, for example encoding spatial smoothing, high-pass filtering or any number of other things.

Weighting Suspected Regions More Heavily

- The best way would be to augment the training data with data from previous studies.
- Lack of data-sharing means this is generally not possible, so we need to extract information from publications.
- The neuroimaging literature is mostly blobs.
- These give pointers about how best to weight the data ($\boldsymbol{\Sigma}_0 = diag(\mathbf{s}), s_i \in \mathbb{R}^+$).

Weighting suspected regions more heavily





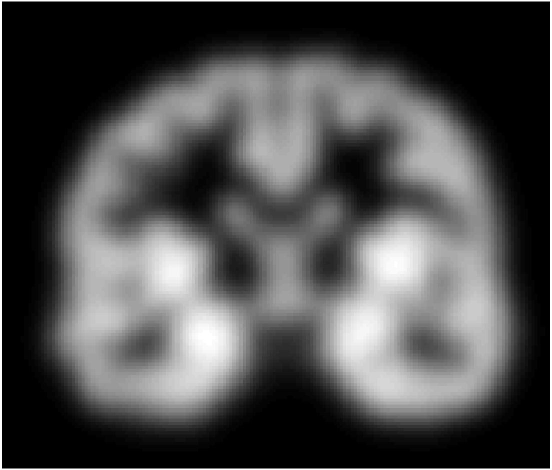
Chu et al. "Does feature selection improve classification accuracy? Impact of sample size and feature selection on classification using anatomical magnetic resonance images". NeuroImage 60:59–70 (2012).

Smoothing

If we know that higher frequency signal is more likely to be noise.

$$\mathbf{K} = \mathbf{X}\boldsymbol{\Sigma}_0\mathbf{X}^T$$

$\boldsymbol{\Sigma}_0$ no longer diagonal.



"Data-driven Feature Selection"

Two main approaches:

- **Non-embedded feature selection**, where approaches such as t- or F-tests, or *recursive feature elimination* are used to switch off certain features. Not very principled, but can save computation time.

  > *We should only do feature selection if there is a cost associated with measuring features or predicting with many features.*
  > *Note: Radford Neal won the NIPS feature selection competition using Bayesian methods that used 100% of the features.*
  >
  > — Zoubin Ghahramani

- **Embedded feature selection**, where features are weighted differently as part of the machine learning model. Works best when features are of different types so need different weighting (*a priori*).

"Data-driven Feature Selection"

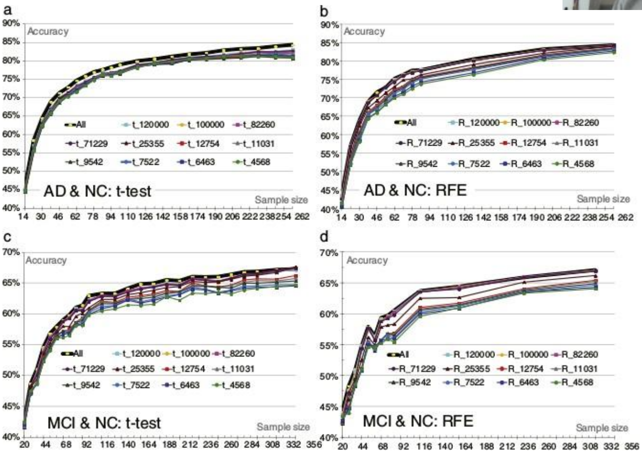
- Lots of effort on data-driven feature selection methods.
  - Involves estimating $\mathbf{\Sigma}_0 = diag(\mathbf{s}), s_i \in \{0, w\}$, where $w \in R^+$.
  - Lots of parameters needed to achieve this.
- Many papers claim excellent results.
- Little evidence to suggest that most voxel-based feature selection methods help.
  - Little or no increase in predictive accuracy.
  - Commonly perceived as being more "interpretable".

"Data-driven Feature Selection"

"In our evaluation, two methods included a feature selection step: Voxel-STAND and Voxel-COMPARE. Overall, these methods did not perform substantially better than simpler ones... … A more robust way to decrease the dimensionality of the features way would be to use more prior knowledge of the disease."

Cuingnet et al. "Automatic classification of patients with Alzheimer's disease from structural MRI: A comparison of ten methods using the ADNI database". NeuroImage 56(2):766–781 (2011).

"Data-driven Feature Selection"



Chu et al. "Does feature selection improve classification accuracy? Impact of sample size and feature selection on classification using anatomical magnetic resonance images". NeuroImage 60:59–70 (2012).

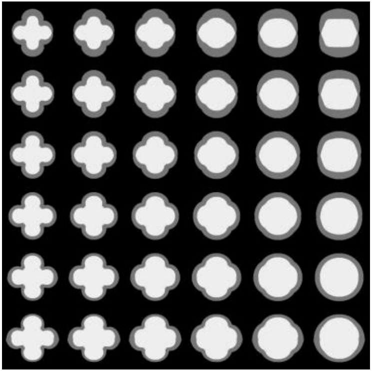
Removing nonlinearities

Instead of using nonlinear pattern recognition methods, we can...

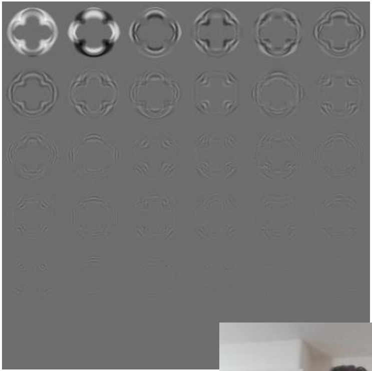
- Capture nonlinearities by appropriate preprocessing.
  - Accurate nonlinear registration can remove much of the nonlinearity.
- Allows nonlinear effects to be modelled by a linear classifier.
- Gives more interpretable characterisations of differences.
- May lead to more accurate predictions – particularly with smaller amounts of training data.

Removing nonlinearities

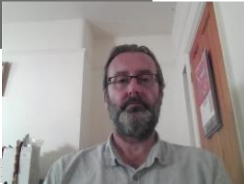
Simulated images



Principal components



A suitable model would reduce this variability to two dimensions.

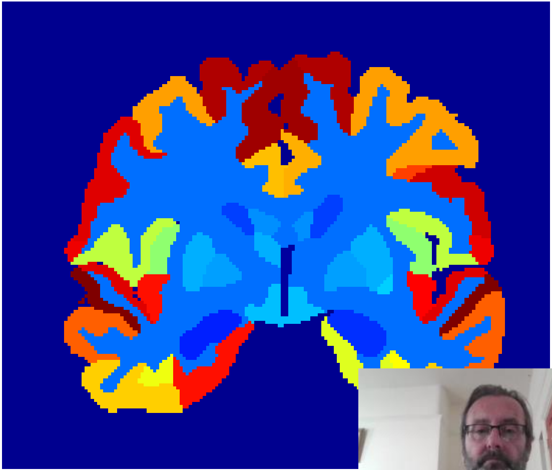


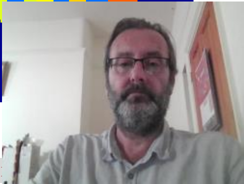
Raw pixel values

Raw pixel data could be another option. Data needs to be "spatially normalised" (and possibly skull-stripped). Results may not generalise well to data from other scanners.

Region volumes

Label propagation or other methods can be used to subdivide brain into regions.

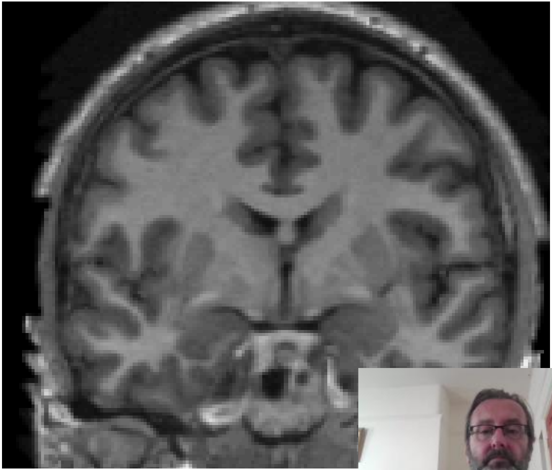




Other features

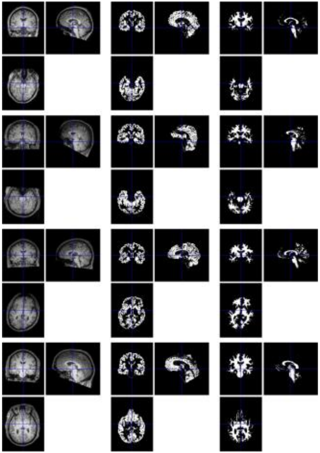
Other features include:

- Cortical thickness.
- Shape features.
- PCA/ICA weights.
- Lesion maps.
- etc

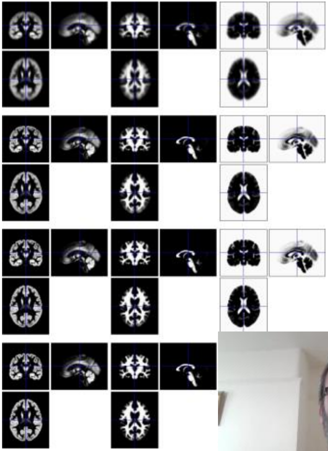


SPM12 Processing

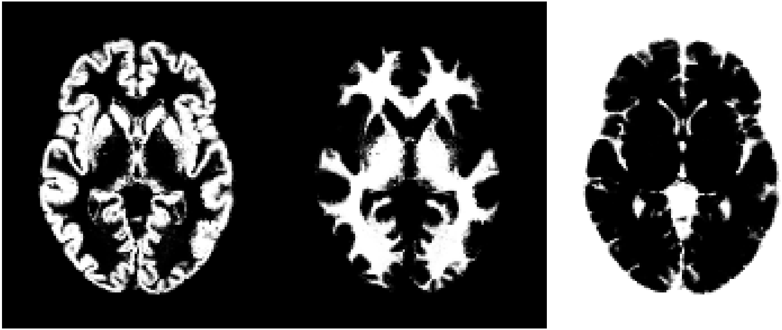
Tissue class segmentation



Alignment with Shoot

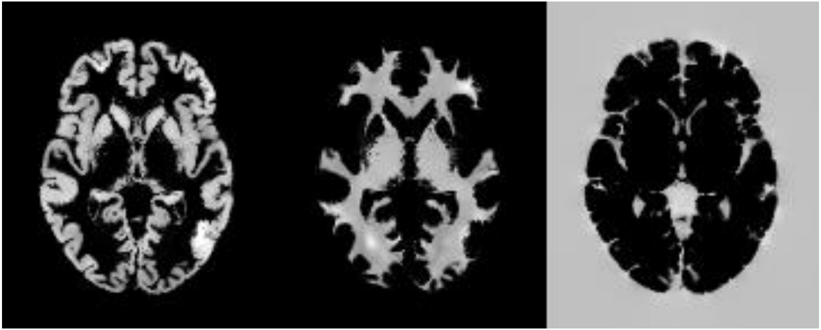


"Unmodulated" GM, WM & BG



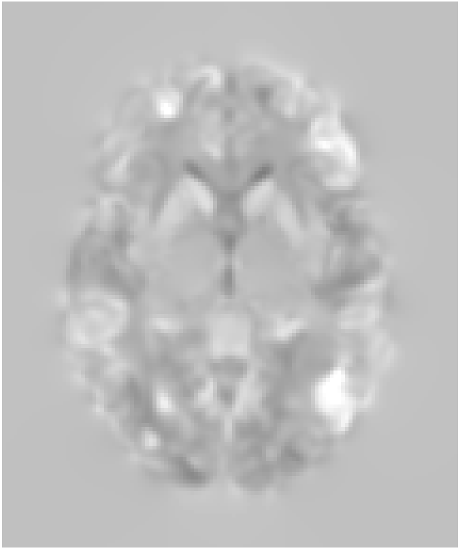
Pattern recognition run using: GM alone; WM alone; BG alone; GM + WM; GM + WM + BG.

"Modulated" GM, WM & BG



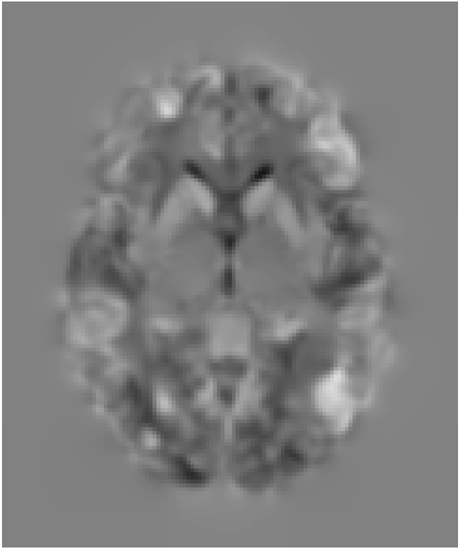
Pattern recognition run using: GM alone; WM alone; BG alone; GM + WM; GM + WM + BG.

Jacobian determinants

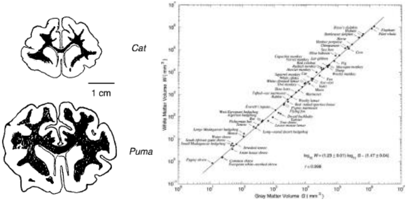


Encodes relative volumes before and after warping.

Logarithms of Jacobian determinants

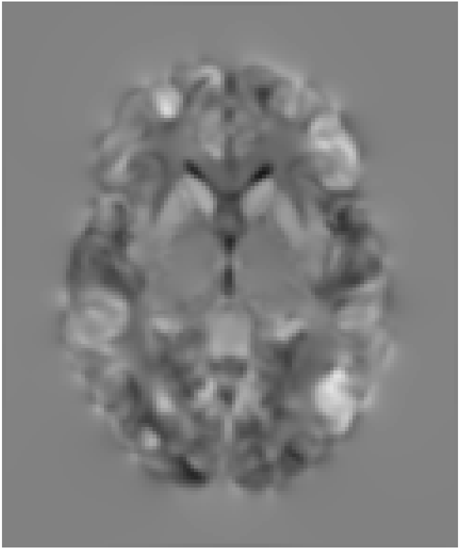


There are sometimes simple logarithmic relationships among volumes.



Zhang and Sejnowski. "A universal scaling law between gray matter and white matter o
Proceedings of the National Ac
97(10):5621–5626 (2000).

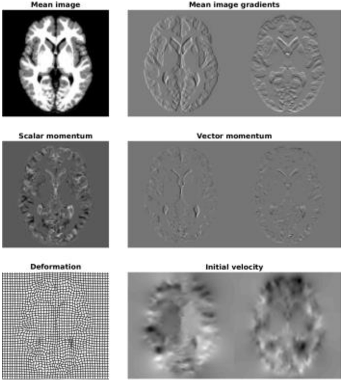
Divergences of Velocity Fields

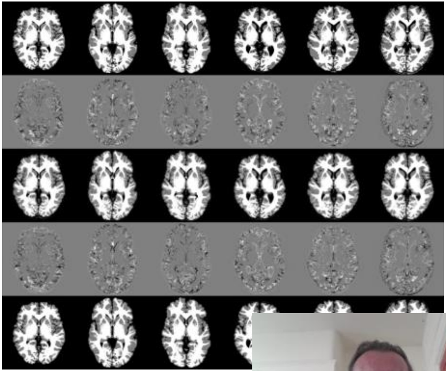


Very similar to logarithms of Jacobians.
Not easy to explain.

Scalar Momentum

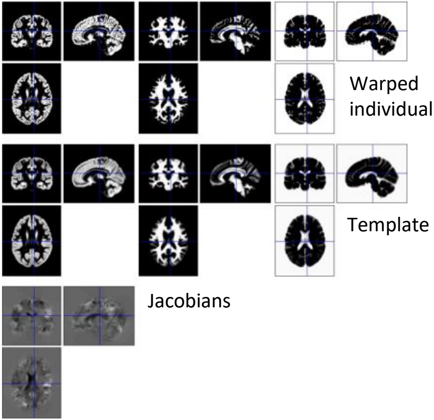
$$\mathbf{a} = |D\varphi|(\boldsymbol{\mu} - \mathbf{c}(\varphi))$$



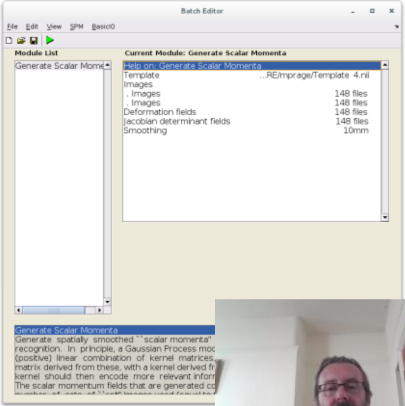


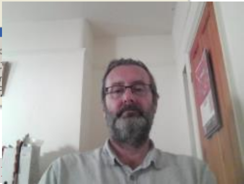
Scalar Momentum

$$\mathbf{a} = |D\varphi|(\boldsymbol{\mu} - \mathbf{c}(\varphi))$$



SPM12 GUI for scalar momentum.



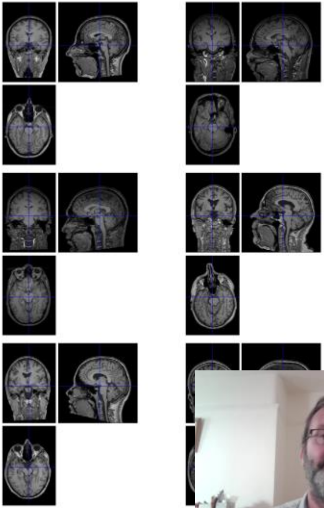


IXI: Dataset

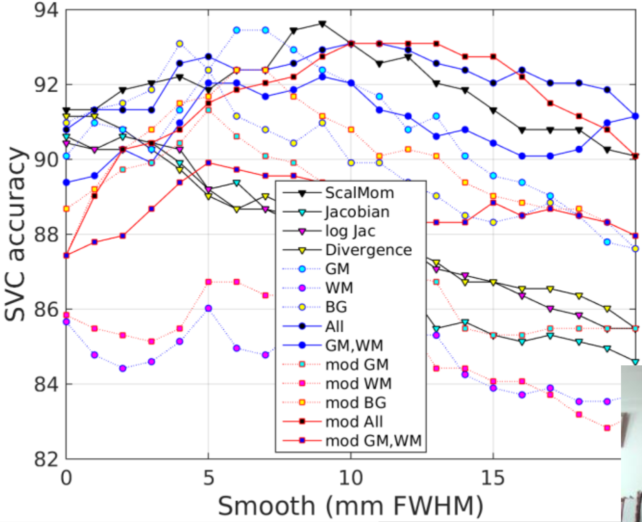
580 T1w brain MRI from IXI (Information eXtraction from Images) dataset.
http://www.brain-development.org/
Data from three different hospitals in London:

- Hammersmith Hospital using a Philips 3T system
- Guy's Hospital using a Philips 1.5T system
- Institute of Psychiatry using a GE 1.5T system

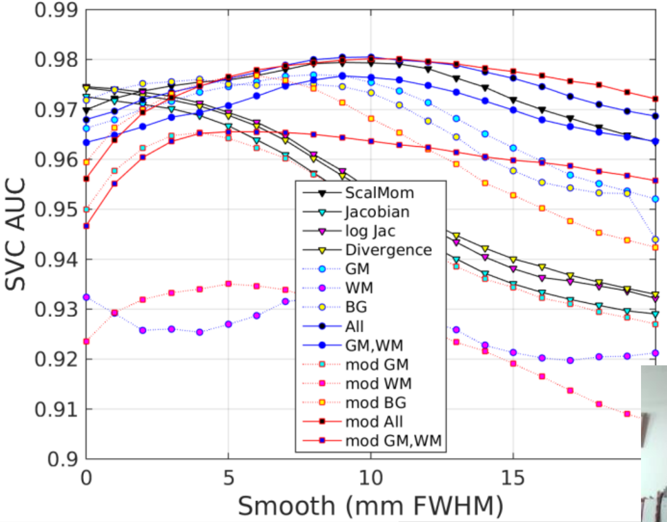
10-fold cross-validation.



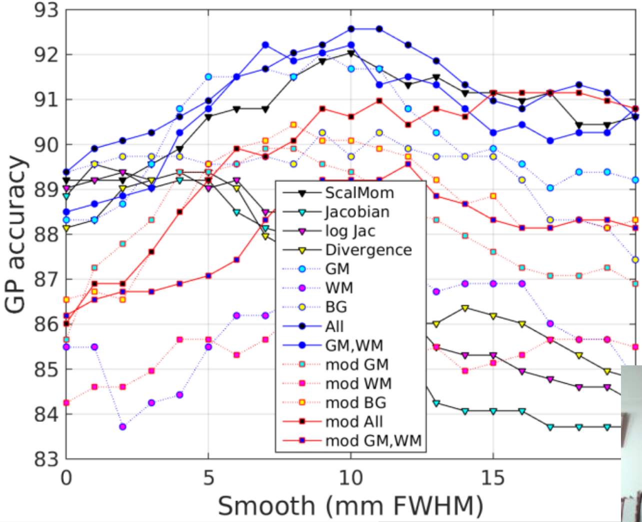
IXI: Gender Classification (SVC)



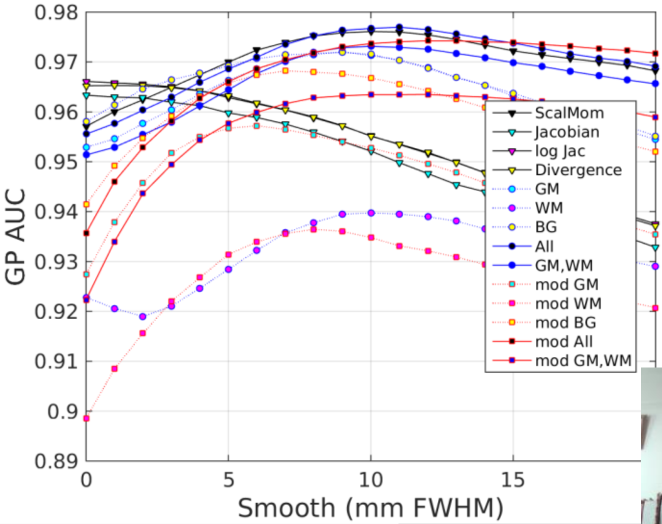
IXI: Gender Classification (SVC)



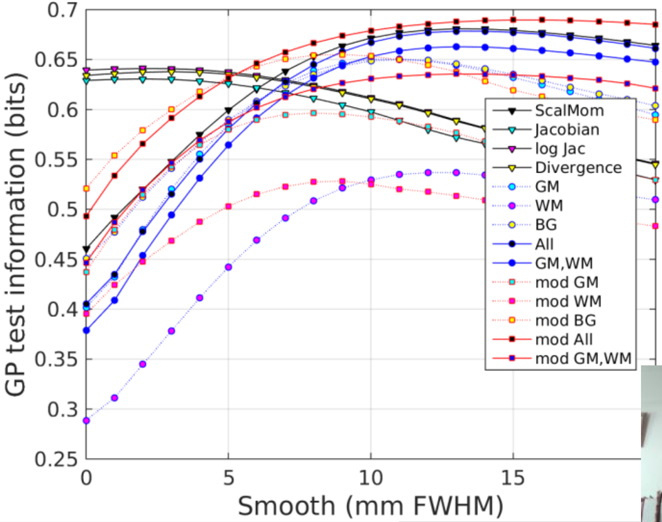
IXI: Gender Classification (GPC)



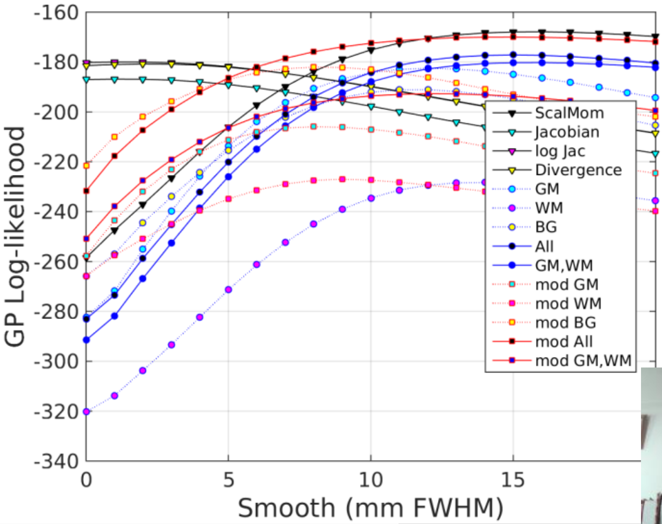
IXI: Gender Classification (GPC)



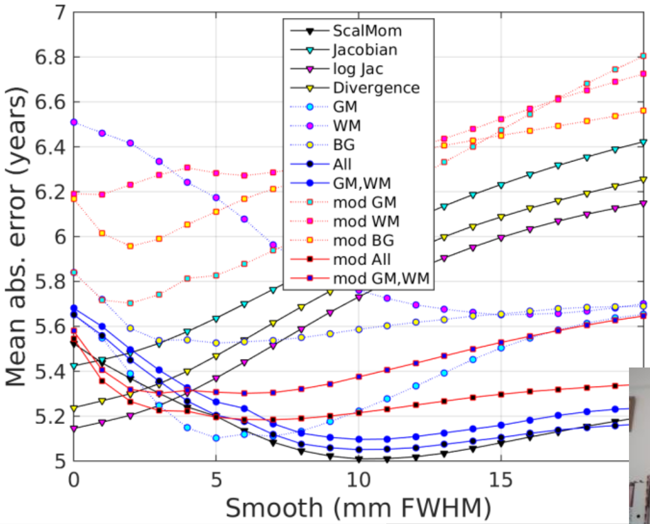
IXI: Gender Classification (GPC)



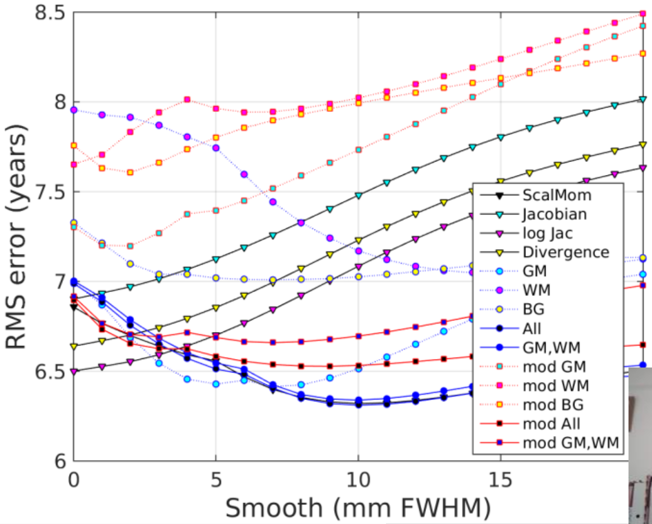
IXI: Gender Classification (GPC)



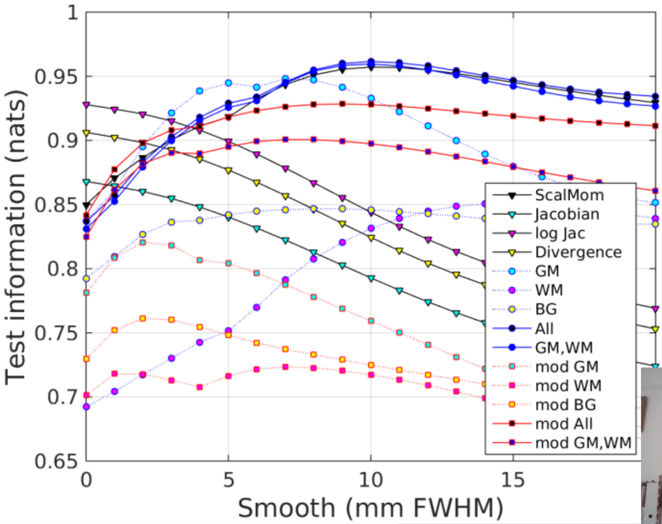
IXI: Age Regression (GPR)



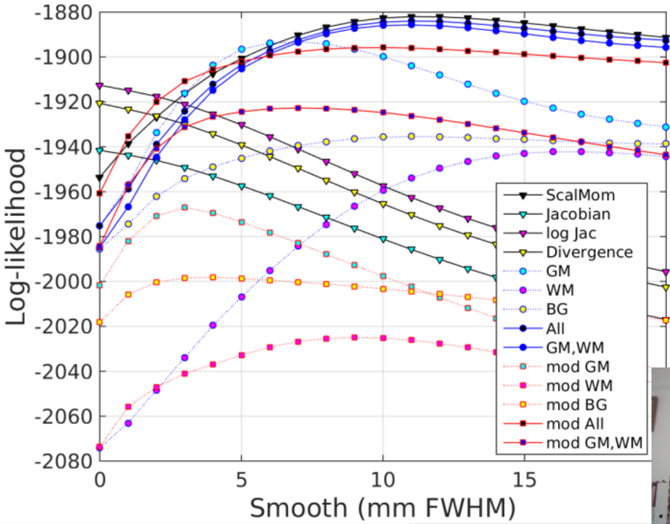
IXI: Age Regression (GPR)



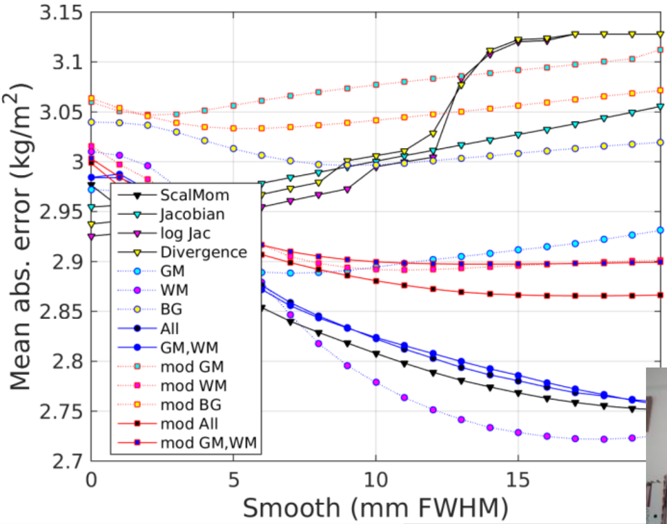
IXI: Age Regression (GPR)



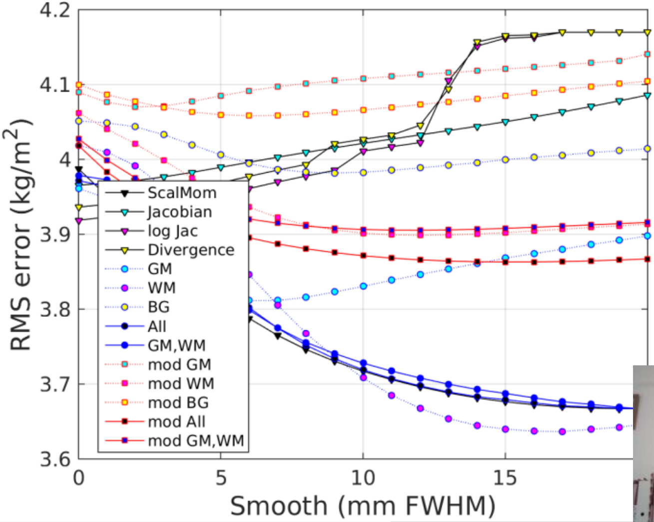
IXI: Age Regression (GPR)



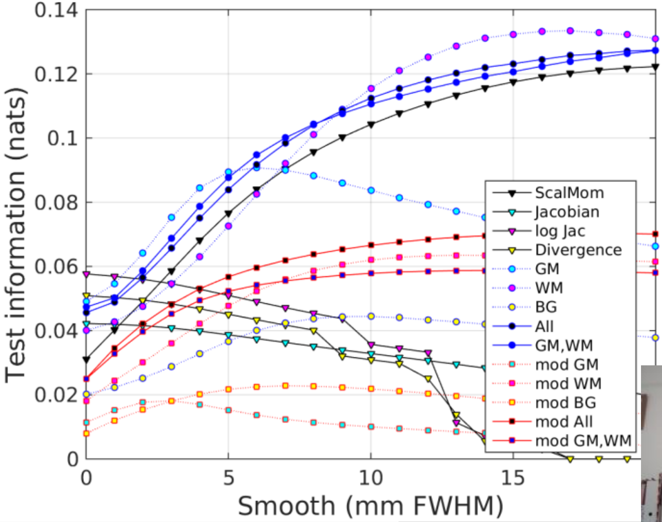
IXI: BMI Regression (GPR)



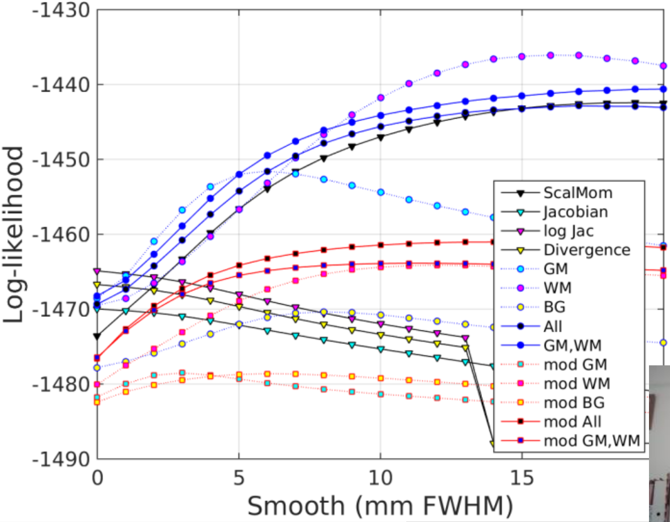
IXI: BMI Regression (GPR)



IXI: BMI Regression (GPR)



IXI: BMI Regression (GPR)



ABIDE: Dataset

The **Autism Brain Imaging Data Exchange** initiative.
http://fcon_1000.projects.nitrc.org/indi/abide/.

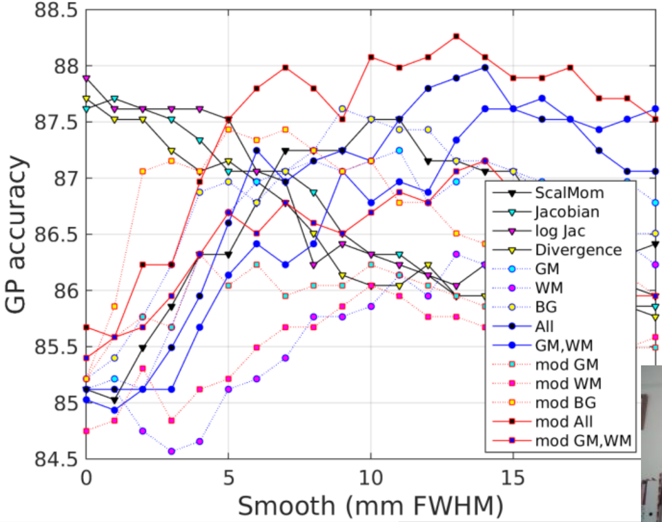
T1w brain MRI from 1,102 subjects.

- 531 with Autism Spectrum Disorder (Gender ratio: 64:467).
- 571 controls (Gender ratio: 99:472).

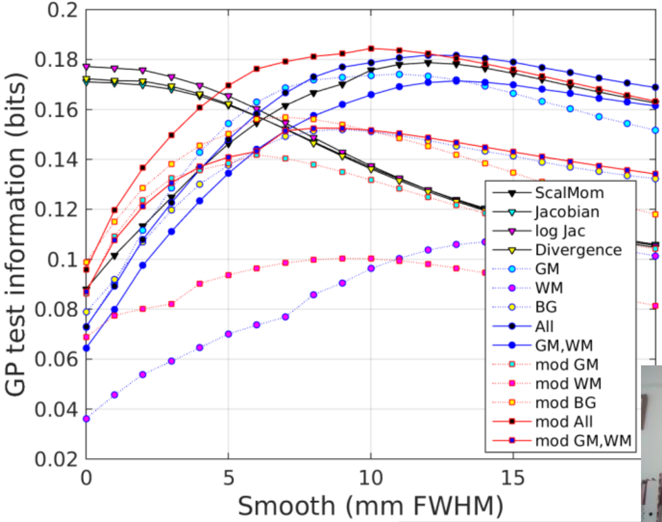
Data from 17 international sites.
The 20 greatest outliers were excluded.

5-fold cross-validation.

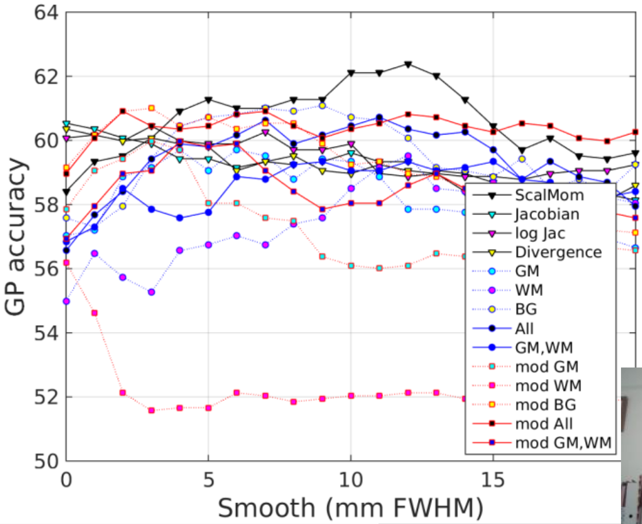
ABIDE: Gender Classification (GPC)



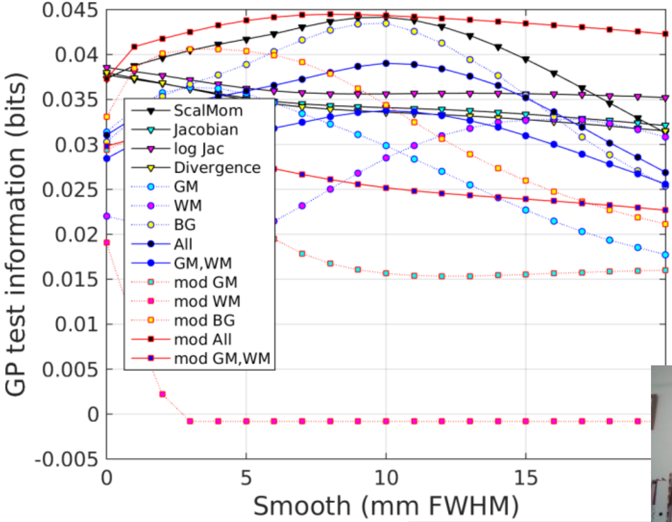
ABIDE: Gender Classification (GPC)



ABIDE: ASD v. Control (GPC)



ABIDE: ASD v. Control (GPC)



COBRE: Dataset

Centre for Biomedical Research Excellence
http://fcon_1000.projects.nitrc.org/indi/retro/cobre.html.

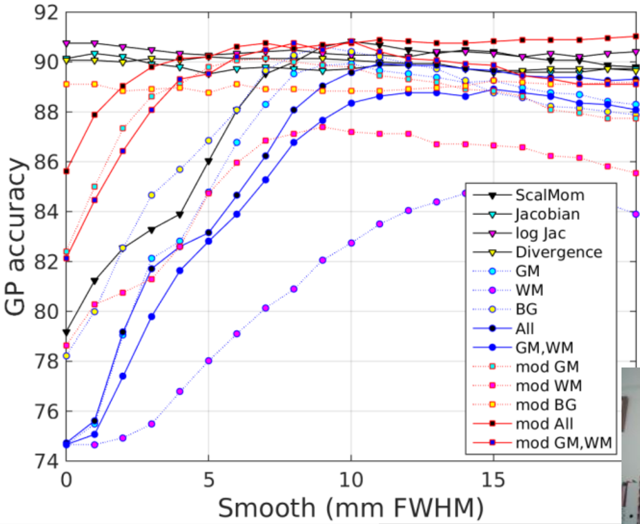
T1w brain MRI from 146 subjects.

- 72 with schizophrenia (14 male : 58 female).
- 74 controls (23 male : 51 female).

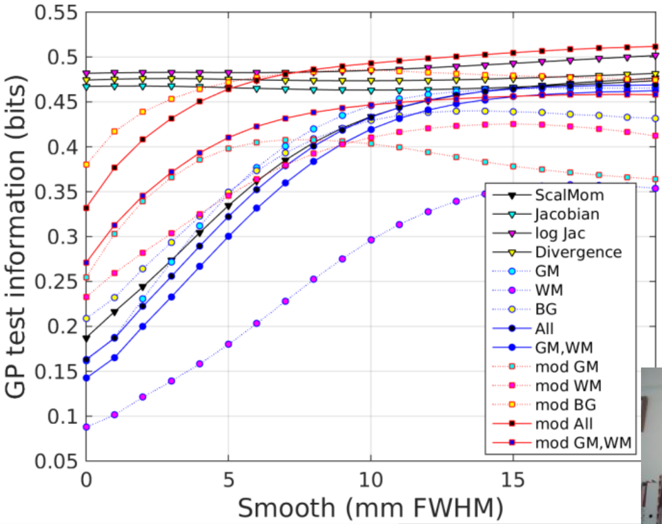
All from a single scanner.

5-fold cross-validation, repeated 10 times.

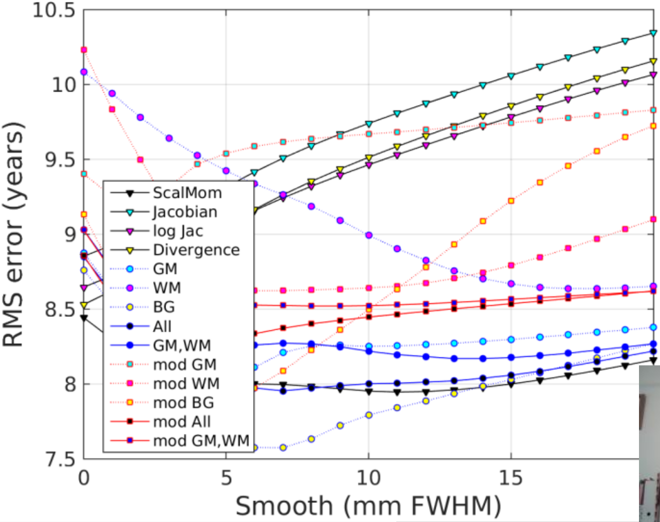
COBRE: Gender Classification (GPC)



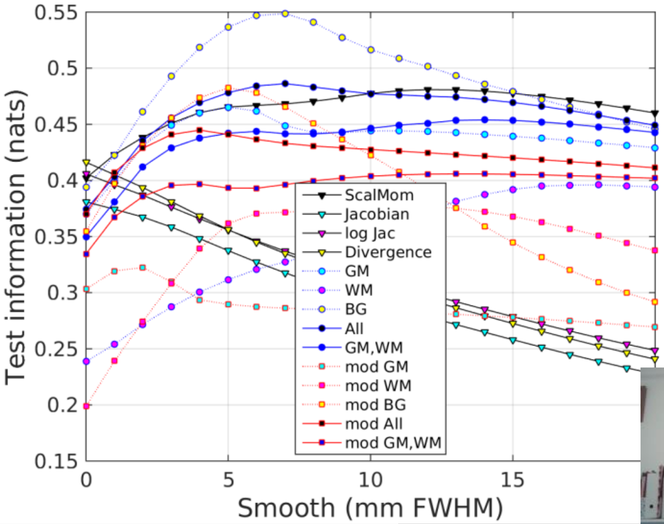
COBRE: Gender Classification (GPC)



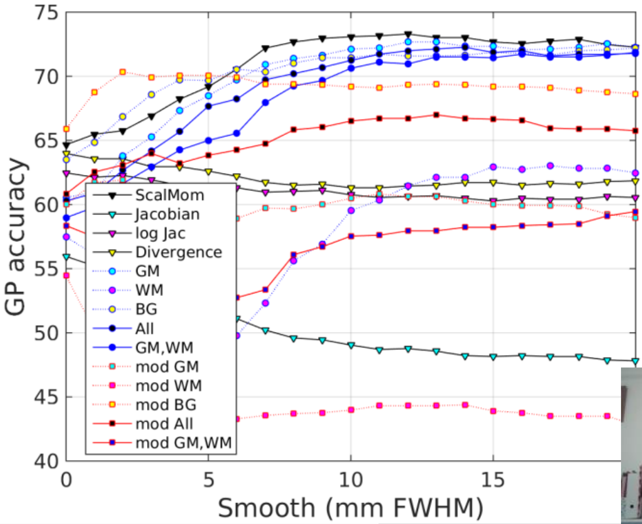
COBRE: Age Regression (GPR)



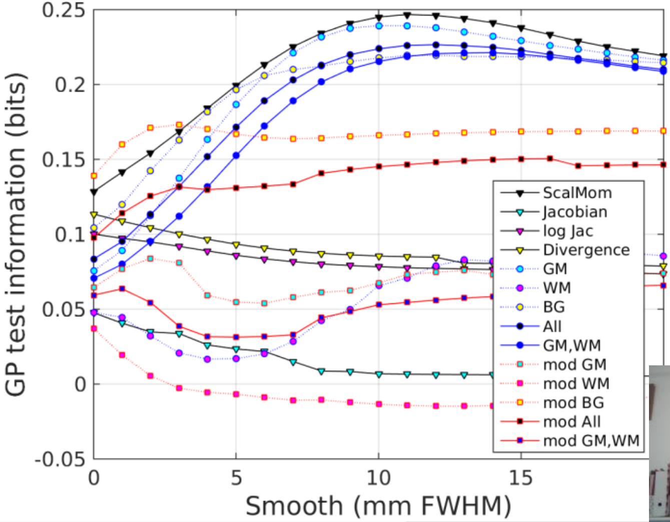
COBRE: Age Regression (GPR)



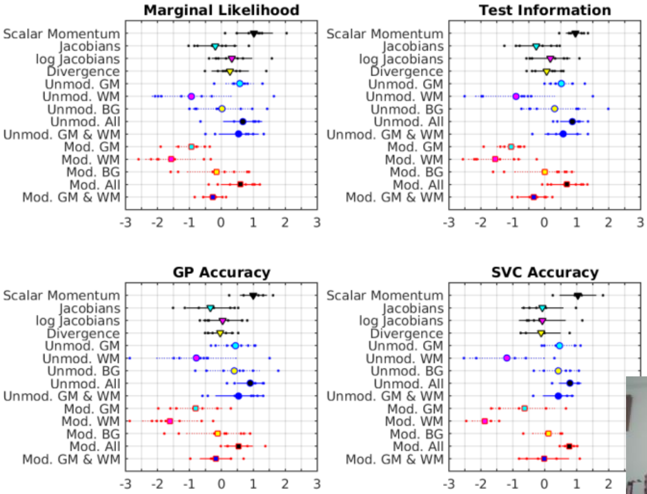
COBRE: Schiz. v. Control (GPC)



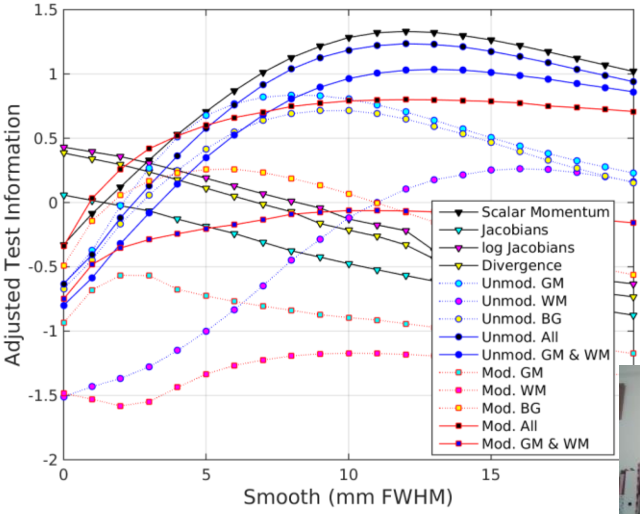
COBRE: Schiz. v. Control (GPC)



Overall Scores



Overall Scores



Conclusions

- No feature set was best in all situations (no free lunch).
- Scalar momentum appears to be a useful feature set, although its effectiveness was not statistically significantly better than other methods that also considered the BG class.
- Jacobian-scaled warped GM alone, or with WM, is surprisingly poor.
- Amount of spatial smoothing makes a difference, with the best results from smoothing of about 12mm FWHM.
- Further dependencies on the details of the registration still need exploring.

Monté-Rubio GC, Falcón C, Pomarol-Clotet E, Ashburner J.
A comparison of various MRI feature types for characterizing whole brain anatomical differences using linear pattern recognition methods.
NeuroImage. 2018 Sep 1;178:753-68.