

GENERATIVE MODELS FOR MEDICAL IMAGING

John Ashburner

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

- 1 INTRODUCTION
 - Pipelines v Models
 - Probability Theory
 - Medical image computing
- 2 A SIMPLE(ISH) MODEL
- 3 LEARNING SHAPE AND APPEARANCE

MORAVEC'S PARADOX

Rodney Brooks explains that, according to early AI research, intelligence was “best characterized as the things that highly educated male scientists found challenging”, such as chess, symbolic integration, proving mathematical theorems and solving complicated word algebra problems. “The things that children of four or five years could do effortlessly, such as visually distinguishing between a coffee cup and a chair, or walking around on two legs, or finding their way from their bedroom to the living room were not thought of as activities requiring intelligence.”

Moravec's paradox. (2015, April 25). In Wikipedia, The Free Encyclopedia. Retrieved 14:46, June 17, 2015, from https://en.wikipedia.org/w/index.php?title=Moravec%27s_paradox&oldid=659139375



IN CS, IT CAN BE HARD TO EXPLAIN THE DIFFERENCE BETWEEN THE EASY AND THE VIRTUALLY IMPOSSIBLE.

<https://xkcd.com>

WHY IMAGE PROCESSING SEEMS EASY

Neurons for visual processing take up 30% of the brain's cortex (as opposed to about 8 % for touch and 3 % for hearing).

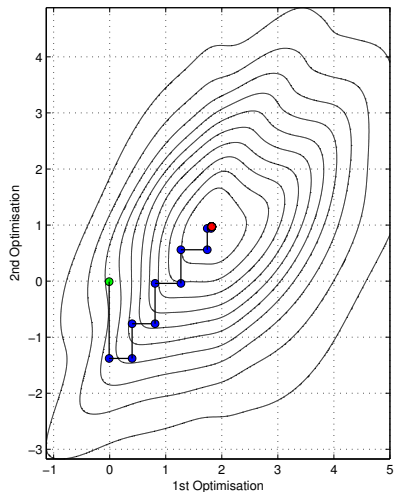
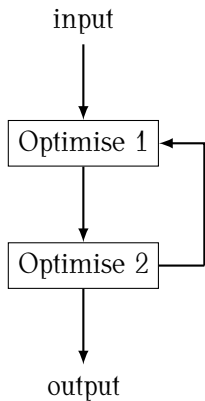


PIPELINES

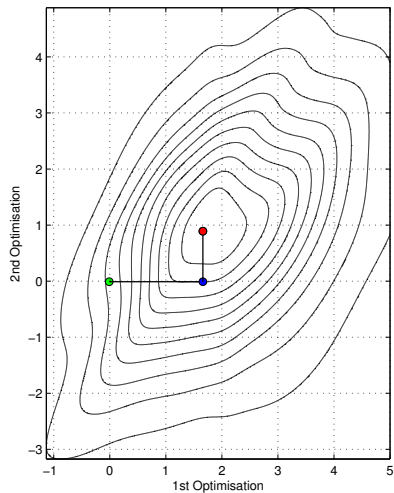
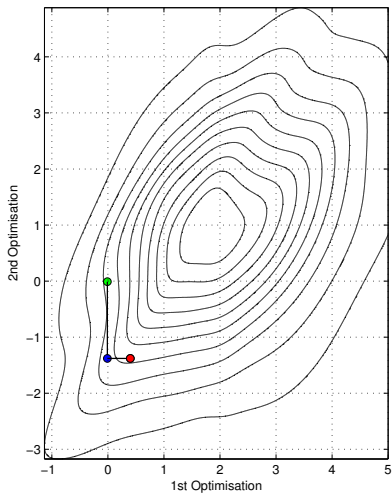
In software engineering, a pipeline consists of a chain of processing elements (processes, threads, coroutines, functions, etc.), arranged so that the output of each element is the input of the next

Pipeline (software). (2015, May 1). In Wikipedia, The Free Encyclopedia. Retrieved 16:50, June 17, 2015, from [https://en.wikipedia.org/w/index.php?title=Pipeline_\(software\)&oldid=660291081](https://en.wikipedia.org/w/index.php?title=Pipeline_(software)&oldid=660291081)

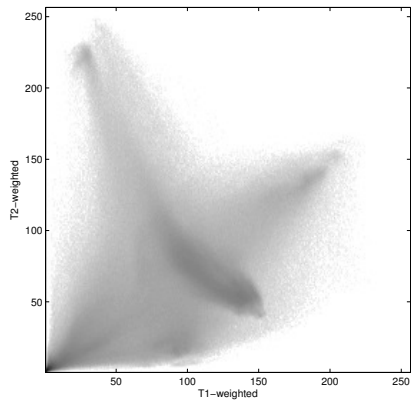
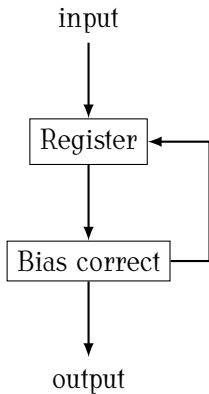
OPTIMISING TWO PARAMETERS



SINGLE PASS



OPTIMISING TWO FUNCTIONS



BOTTOM-UP & TOP-DOWN PROCESSING IN THE BRAIN

- Pipelines are a purely bottom up approach, with no top-down control.
- Data processing in the brain involves both top-down and bottom-up processing.
- Can not expect to achieve optimal understanding from a purely bottom-up approach.

GENERATIVE MODELS

A generative model is a model for randomly generating observable data, typically given some hidden parameters. It specifies a joint probability distribution over observation and label sequences.

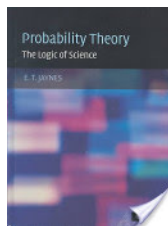
Generative models are used in machine learning for either modeling data directly (i.e., modeling observations draws from a probability density function), or as an intermediate step to forming a conditional probability density function. A conditional distribution can be formed from a generative model through Bayes' rule.

Generative model. (2015, April 30). In Wikipedia, The Free Encyclopedia. Retrieved 16:46, June 17, 2015, from https://en.wikipedia.org/w/index.php?title=Generative_model&oldid=660109222

PROBABILITY THEORY

“Probability theory is nothing but common sense reduced to calculation.”

— Laplace



Desiderata of probability theory:

- ➊ Representation of degree of plausibility by real numbers.
- ➋ Qualitative correspondence with common sense.
- ➌ Consistency.

Jaynes, Edwin T. *Probability theory: the logic of science*. Cambridge university press, 2003.

PRODUCT AND SUM RULES

Product Rule

$$\begin{aligned} p(\mathbf{y}, \mathbf{x}) &= p(\mathbf{y}|\mathbf{x})p(\mathbf{x}) \\ &= p(\mathbf{x}|\mathbf{y})p(\mathbf{y}) \end{aligned}$$

Sum Rule

$$p(\mathbf{y}) = \sum_{\mathbf{x}} p(\mathbf{y}, \mathbf{x})$$

or for continuous \mathbf{x}

$$p(\mathbf{y}) = \int_{\mathbf{x}} p(\mathbf{y}, \mathbf{x}) d\mathbf{x}$$

$p(\mathbf{x})$ is the probability of \mathbf{x} .

$p(\mathbf{x}, \mathbf{y})$ is the joint probability of \mathbf{x} and \mathbf{y} .

$p(\mathbf{x}|\mathbf{y})$ is the probability of \mathbf{x} conditional on \mathbf{y} .

BAYES RULE

Combining the sum and product rules, gives Bayes rule:

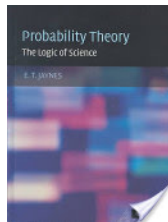
$$p(\theta|\mathbf{X}) = \frac{p(\mathbf{X}|\theta)p(\theta)}{p(\mathbf{X})} = \frac{p(\mathbf{X}|\theta)p(\theta)}{\int_{\theta} p(\mathbf{X}|\theta)p(\theta)d\theta}$$

In words:

$$\text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Evidence}}$$

IGNORANCE PRIORS

- Sometimes we don't have previous observations to formulate priors.
- Jaynes suggests using a maximum entropy prior.
- An ignorance prior is a prior probability distribution where equal probability is assigned to all possibilities.
- Ignorance priors can be motivated via invariance/symmetry (transformation groups).

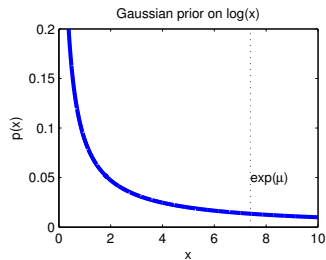
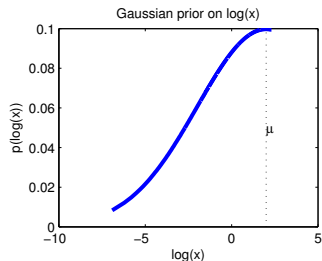


Jaynes, Edwin T. *Probability theory: the logic of science*. Cambridge university press, 2003.

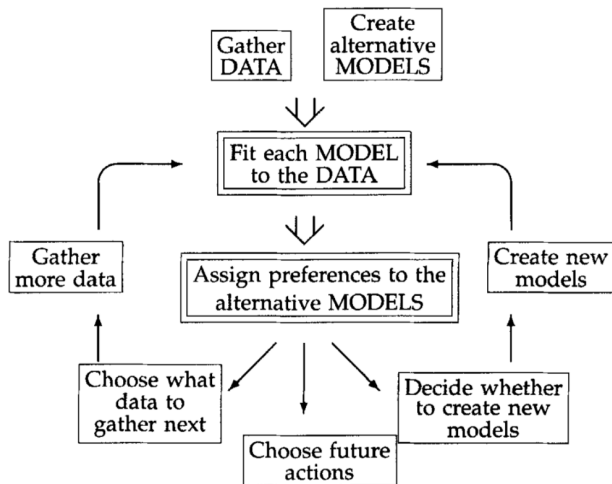
PRIORS FOR POSITIVE VALUES

- Some things can not be less than zero.
 - Counts of observed photons.
 - Multiplicative “bias” fields.
 - Lengths, areas, volumes, etc.
- Formulate the model via logarithms, and impose a prior on these.

Jeffreys, Harold. “An invariant form for the prior probability in estimation problems.” In Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, vol. 186, no. 1007, pp. 453-461. The Royal Society, 1946.

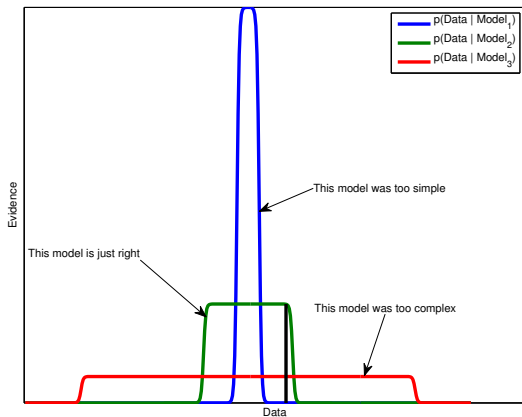


SCIENTIFIC PROCESS



MacKay, David JC. "Bayesian interpolation." *Neural computation* 4, no. 3 (1992): 415-447.

GOLDBLOCKS AND THE THREE BAYESIAN MODELS



“Everything should be made as simple as possible, but not simpler.”

— Einstein (possibly)

GENERAL AIM OF MEDICAL IMAGE COMPUTING

Given an image or a set of images \mathbf{x}^* , best predict \mathbf{y}^* .

Here, \mathbf{y} may be:

- A diagnosis.
- An optimal treatment decision.
- Another image, for example:
 - A cleaned up version of the same image.
 - A map of where a neurosurgeon should best avoid.
 - A map of gamma ray absorption for attenuation correction in MR/PET.
- etc

GENERAL AIM OF MEDICAL IMAGE COMPUTING

Often a collection of training data to work from (\mathbf{X} and \mathbf{Y}).
The aim becomes of of determining $p(\mathbf{y}^*|\mathbf{x}^*, \mathbf{Y}, \mathbf{X})$.

GENERAL AIM OF MEDICAL IMAGE COMPUTING

Predictions are based on some model, \mathcal{M} . Usually, a model has parameters, θ :

$$\begin{aligned} p(\mathbf{y}^* | \mathbf{x}^*, \mathbf{Y}, \mathbf{X}, \mathcal{M}) &= \int_{\theta} p(\mathbf{y}^*, \theta | \mathbf{x}^*, \mathbf{Y}, \mathbf{X}, \mathcal{M}) d\theta \\ &= \int_{\theta} p(\mathbf{y}^* | \mathbf{x}^*, \mathbf{Y}, \mathbf{X}, \theta, \mathcal{M}) p(\theta | \mathcal{M}) d\theta \end{aligned}$$

Predictions may also be made by averaging over models.

$$p(\mathbf{y}^* | \mathbf{x}^*, \mathbf{Y}, \mathbf{X}) = \sum_i p(\mathbf{y}^* | \mathbf{x}^*, \mathbf{Y}, \mathbf{X}, \mathcal{M}_i) P(\mathcal{M}_i)$$

UNEORTUNATELY...

"In theory, there is no difference between theory and practice. But, in practice, there is."

Many of the integrations needed to compute model evidence are not computationally feasible in medical image computing applications. Workarounds include:

- Use *maximum a posteriori* (MAP) estimation, and approximate probability distributions via a delta function.

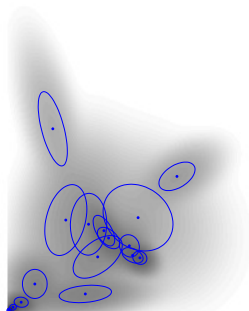
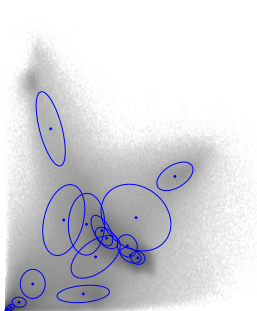
$$\hat{\theta} = \arg \max_{\theta} \log p(\mathbf{X}, \theta)$$

- Model selection via cross-validation.

- 1 INTRODUCTION
- 2 A SIMPLE(ISH) MODEL
 - Tissue appearance
 - Tissue priors
- 3 LEARNING SHAPE AND APPEARANCE

MIXTURE OF GAUSSIANS

$$\begin{aligned}\mathcal{E} &= -\log p(\mathbf{f}|\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\gamma}) \\ &= -\sum_{i=1}^I \log \left(\sum_{k=1}^K \frac{\gamma_k}{\sqrt{2\pi\sigma_k^2}} \exp \left(-\frac{(f_i - \mu_k)^2}{2\sigma_k^2} \right) \right)\end{aligned}$$

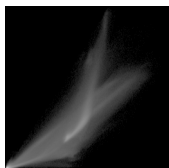


INCORPORATING “BIAS” CORRECTION

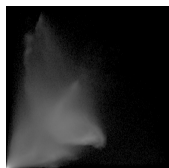
$$\mathcal{E} = - \sum_{i=1}^I \log \left(\sum_{k=1}^K \frac{\gamma_k}{\sqrt{2\pi \frac{\sigma_k^2}{\rho_i(\boldsymbol{\beta})^2}}} \exp \left(- \frac{\left(f_i - \frac{\mu_k}{\rho_i(\boldsymbol{\beta})} \right)^2}{2 \frac{\sigma_k^2}{\rho_i(\boldsymbol{\beta})^2}} \right) \right)$$



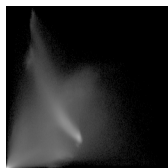
Original



Corrected



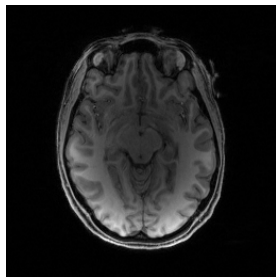
Original



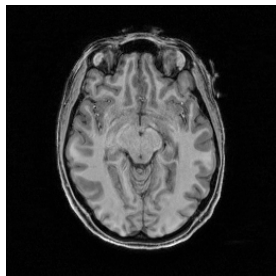
Corrected

INCORPORATING “BIAS” CORRECTION

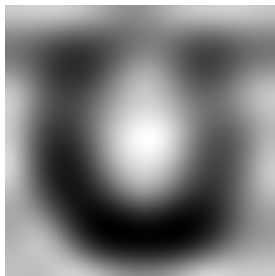
$$\mathcal{E} = - \sum_{i=1}^I \log \left(\rho_i(\boldsymbol{\beta}) \sum_{k=1}^K \frac{\gamma_k}{\sqrt{2\pi\sigma_k^2}} \exp \left(- \frac{(\rho_i(\boldsymbol{\beta}) f_i - \mu_k)^2}{2\sigma_k^2} \right) \right)$$



Original



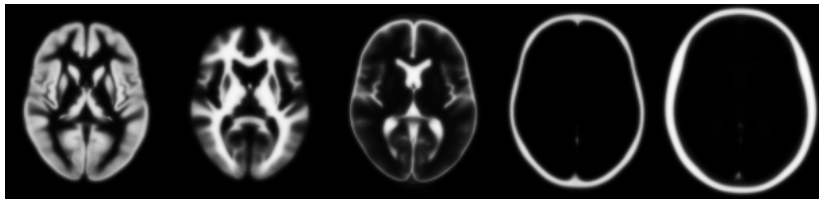
Corrected



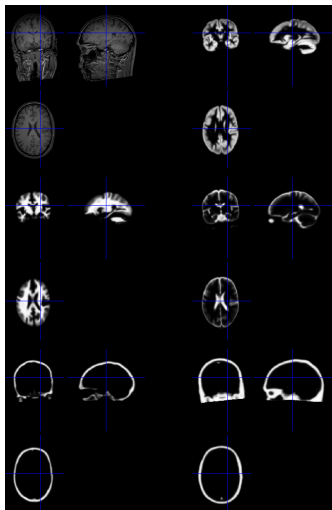
Field

INCORPORATING DEFORMABLE TISSUE PRIORS

$$\mathcal{E} = - \sum_{i=1}^I \log \left(\frac{\rho_i(\boldsymbol{\beta})}{\sum_{k=1}^K \gamma_k b_{ik}(\boldsymbol{\alpha})} \sum_{k=1}^K \frac{\gamma_k b_{ik}(\boldsymbol{\alpha})}{\sqrt{2\pi\sigma_k^2}} \exp \left(- \frac{(\rho_i(\boldsymbol{\beta}) f_i - \mu_k)^2}{2\sigma_k^2} \right) \right)$$

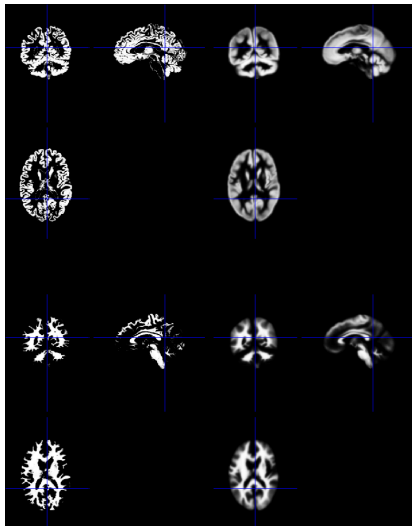


INCORPORATING DEFORMABLE TISSUE PRIORS



$$\mathcal{E} = - \sum_{i=1}^I \log \left(\frac{\rho_i(\beta)}{\sum_{k=1}^K \gamma_k b_{ik}(\alpha)} \sum_{k=1}^K \frac{\gamma_k b_{ik}(\alpha)}{\sqrt{2\pi\sigma_k^2}} \exp \left(- \frac{(\rho_i(\beta) f_i - \mu_k)^2}{2\sigma_k^2} \right) \right)$$

LATENT VARIABLES



Optimisation done via EM.

Marginalised with respect to latent variables (\mathbf{z}), which encode tissue class memberships.

$$p(\mathbf{f}, \theta) = \int_{\mathbf{z}} p(\mathbf{f}, \mathbf{z}, \theta) d\mathbf{z}$$

where

$$\theta = \{\mu, \sigma, \gamma, \beta, \alpha\}$$

- Ashburner, John, and Karl J. Friston. "*Unified segmentation.*" *Neuroimage* 26, no. 3 (2005): 839-851.
- http://www.fil.ion.ucl.ac.uk/spm/software/spm12/,spm12/spm_preproc_run.m.

- 1 INTRODUCTION
- 2 A SIMPLE(ISH) MODEL
- 3 LEARNING SHAPE AND APPEARANCE**
 - Equations
 - Examples

PRINCIPAL COMPONENT ANALYSIS

Given a $P \times N$ matrix \mathbf{F} , decompose it into a $P \times K$ matrix \mathbf{H} and a $K \times N$ matrix \mathbf{W} , such that:

$$\mathbf{F} \simeq \mathbf{H}\mathbf{W}$$

The EM algorithm is:

- E-step: $\mathbf{W} \leftarrow (\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T \mathbf{F}$
- M-step: $\mathbf{H} \leftarrow \mathbf{F}\mathbf{W}^T (\mathbf{W}\mathbf{W}^T)^{-1}$

Roweis, Sam. "EM algorithms for PCA and SPCA." Advances in neural information processing systems (1998): 626-632.

PRINCIPAL COMPONENT ANALYSIS

Minimise the following w.r.t. \mathbf{H} and \mathbf{W} :

$$\mathcal{E} = \sum_{n=1}^N \frac{1}{2} \|\mathbf{f}_n - \sum_{k=1}^K \mathbf{h}_k w_{kn}\|^2$$

Or this, w.r.t. $\boldsymbol{\mu}$, \mathbf{H} and \mathbf{W} :

$$\mathcal{E} = \sum_{n=1}^N \frac{1}{2} \|\mathbf{f}_n - \boldsymbol{\mu} - \sum_{k=1}^K \mathbf{h}_k w_{kn}\|^2$$

GENERALISED PRINCIPAL COMPONENT ANALYSIS

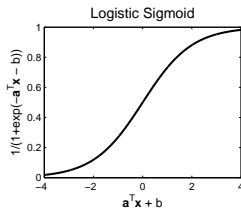
If \mathbf{F} is binary, we could fit a logistic version by minimising the following w.r.t. \mathbf{H} and \mathbf{W} :

$$\mathcal{E} = - \sum_{n=1}^N \sum_{p=1}^P \log(\sigma_{pn}) f_{pn} + \log(1 - \sigma_{pn})(1 - f_{pn})$$

where

$$\sigma_{pn} = \frac{1}{1 + \exp(\sum_{k=1}^K h_{pk} w_{kn})}$$

The EM algorithm involves logistic regression.



PRINCIPAL GEODESIC ANALYSIS

Could combine diffeomorphic registration with PCA by minimising:

$$\mathcal{E} = \sum_{n=1}^N \frac{\lambda}{2} \|\mathbf{f}_n - \boldsymbol{\mu} \circ \boldsymbol{\varphi}_n^{-1}\|^2 + \frac{1}{2} \|\mathbf{v}_n\|_{\mathbf{V}}^2$$

where \mathbf{H} encodes principal components of initial velocity for computing diffeomorphisms:

$$\begin{aligned} \mathbf{v}_n &= \sum_{k=1}^K \mathbf{h}_k w_{kn} \\ \boldsymbol{\varphi}_n &= \text{Exp}(\mathbf{v}_n) \text{ (via geodesic shooting)} \end{aligned}$$

Zhang, Miaomiao, and P. Thomas Fletcher. "Probabilistic principal geodesic analysis." In Advances in Neural Information Processing Systems, pp. 1178-1186. 2013.

Zhang, Miaomiao, and P. Thomas Fletcher. "Bayesian Principal Geodesic Analysis for Estimating Intrinsic Diffeomorphic Image Variability." Medical Image Analysis (2015).

COMBINED PCA/PGA MODEL

Could combine diffeomorphic registration with PCA by minimising the following w.r.t. $\boldsymbol{\mu}$, \mathbf{H} , \mathbf{A} and \mathbf{W} :

$$\mathcal{E} = \sum_{n=1}^N \frac{\lambda_1}{2} \|\mathbf{f}_n - (\boldsymbol{\mu} + \mathbf{r}_n) \circ \boldsymbol{\varphi}_n^{-1}\|^2 + \frac{\lambda_2}{2} \|\mathbf{r}_n\|^2 + \frac{1}{2} \|\mathbf{v}_n\|_V^2$$

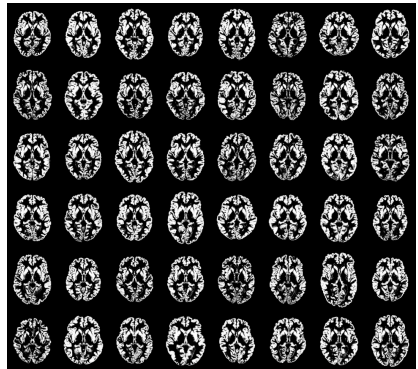
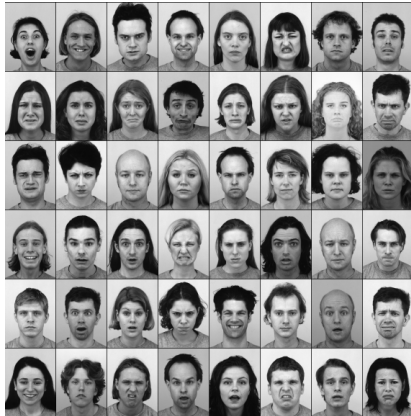
where:

$$\begin{aligned}\mathbf{v}_n &= \sum_{k=1}^K \mathbf{h}_k w_{kn} \\ \boldsymbol{\varphi}_n &= \text{Exp}(\mathbf{v}_n) \\ \mathbf{r}_n &= \sum_{k=1}^K \mathbf{a}_k w_{kn}\end{aligned}$$

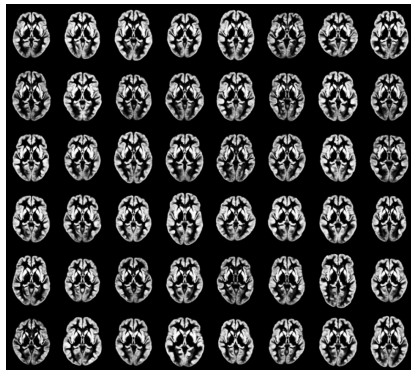
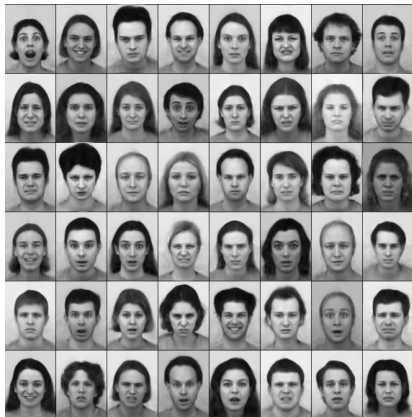
Note: Some form of *metamorphoses* approach may be better.

Richardson, Casey L., and Laurent Younes. "Metamorphosis of Images in Reproducing Kernel Hilbert Spaces." arXiv preprint arXiv:1409.6573 (2014).

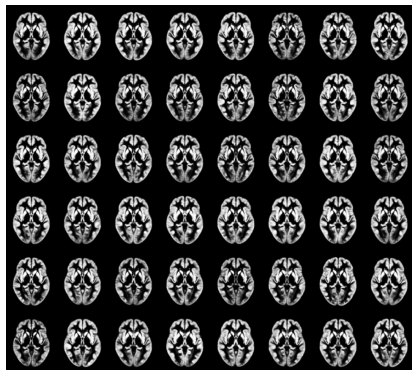
ORIGINAL IMAGES



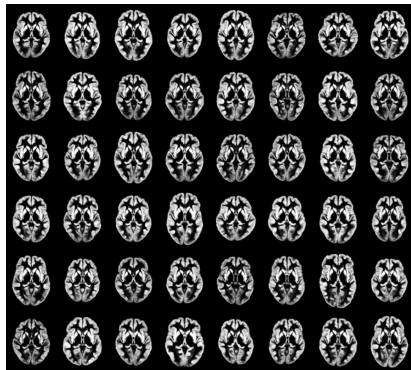
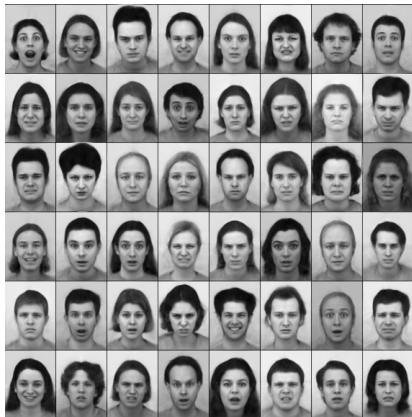
FULL MODEL FIT



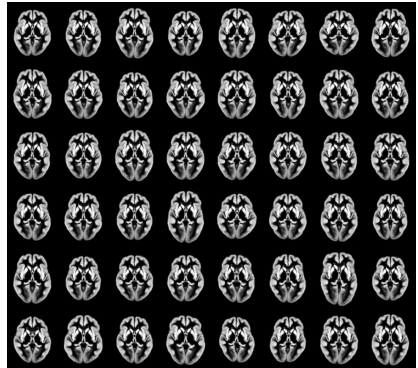
APPEARANCE FIT ONLY



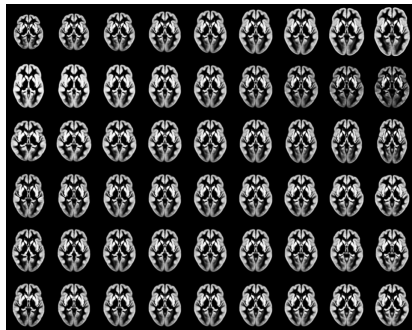
FULL MODEL FIT



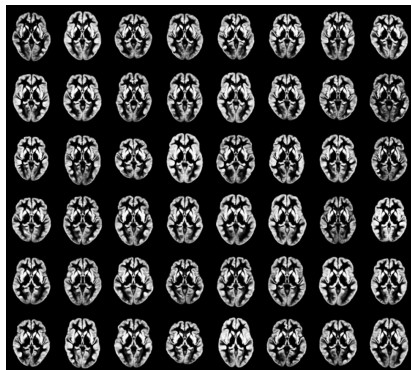
SHAPE FIT ONLY



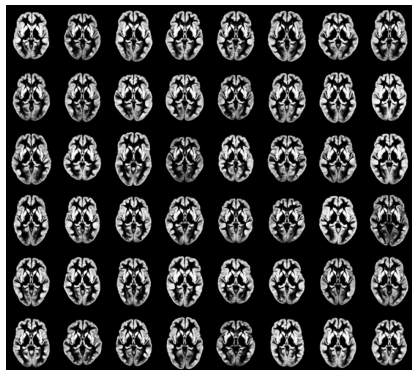
PRINCIPAL MODES



RANDOM SAMPLES (1)

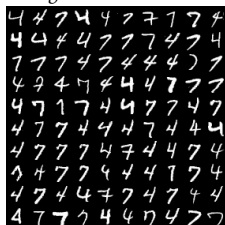


RANDOM SAMPLES (2)

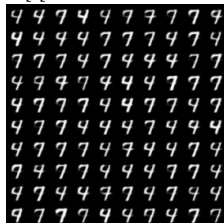


MNIST 4/7

Original



Appearance



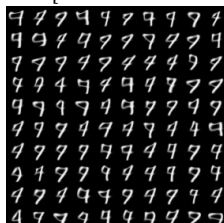
Principal modes



Full fit

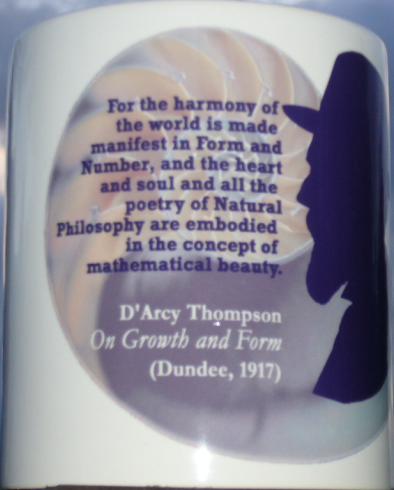


Shape



“To recognize shapes, first learn to generate images”

Geoffrey E Hinton (2007)



For the harmony of
the world is made
manifest in Form and
Number, and the heart
and soul and all the
poetry of Natural
Philosophy are embodied
in the concept of
mathematical beauty.

D'Arcy Thompson
On Growth and Form
(Dundee, 1917)