Modelling in the Context of Massively Missing Data

Neil D. Lawrence
Sheffield Institute of Translational Neuroscience and Department of Computer Science, University of Sheffield, U.K.

Max Planck Institute, Tübingen

18th March 2015

Outline

Box Quote

All models are wrong, but some are useful. *(Box, 1976)*

Box Quote

All models are wrong, but some are useful. *(Box, 1976)*

- Useful quote, but overused.

Box Quote

All models are wrong, but some are useful. *(Box, 1976)*

- Useful quote, but overused.
- Almost become an excuse, my model is wrong so it *might* be useful.

Box Quote

All models are wrong, but some are useful. (Box, 1976)

- Useful quote, but overused.
- Almost become an excuse, my model is wrong so it *might* be useful.

... the scientist must be alert to what is importantly wrong. It is inappropriate to worry about mice when there are tigers abroad. (Box, 1976)

An Incorrect Model

- Write down our data ...

$$\mathbf{Y} \in \Re^{n \times p}$$

An Incorrect Model

- Write down our data ...

$$\mathbf{Y} \in \mathfrak{R}^{n \times p}$$

... this is WRONG!

Is this Separation a Historical Anachronism?

- A presumption: there is something special and separate about indices over n and p.
- The subtle difference between features and data points.
- In practice both n and p could be uncountably large!
- Standard approach seems to assume that p is fixed.
- A historic anachronism from the days of collating statistical information?

There is nothing special about p ...

- Rather ... let's assume each data is indexed by the type of data, as well as location, time, etc.
- So $y_{17,234}$ is price of a hamburger from McDonald's in Leicester square on 13th April 1984 at 13:34 and $y_{239,201}$ is the price of a chicken wrap from Pret a Manger in Cambridge on 27th December 2001 at 14:34.
- Further $y_{734,124}$ might be the brand of car my mother currently drives.

Prediction

The answer to any prediction problem is a probability distribution. *(Peter McCullogh via Peter Diggle)*

- We assume that we are interested in predicting something about our variables (the likely cost of a burger given the cost of a chicken wrap).

Factorizations

- Often researchers write down the resulting factorization without a second thought:

$$p(\mathbf{Y}|\boldsymbol{\theta}) = \prod_{i=1}^{n} p(\mathbf{y}_{i,:}|\boldsymbol{\theta})$$

- This means that all our information about different data is stored in the parameters.
- If model is complex, and number of parameters is large, then they will be badly determined when data is few.
- For me: interesting *research* problems are defined by needing (more) complex models.

Data and Modelling

- "The Unreasonable Effectiveness of ...
  - ... Mathematics" (Wigner, 1960)
  - ...Data" (Halevy et al., 2009)
- This is a *false* dichotomy.
- Both are needed for challenging problems of the future.
  - The relative importance of each is dependent on application.
  - Norvig also accepts this (see Nando's question: `http://www.youtube.com/watch?v=yvDCzhbjYWs&t=54m40s`).
- Prediction requires model (mathematics) and data.
- Having better models is particularly important when there's *uncertainty*.

Open Data

- Automatic data curation: from curated data to curation of publicly available data.
- Open Data: `http://www.openstreetmap.org/?lat=53.38086&lon=-1.48545&zoom=17&layers=M`.

Open Data

- Automatic data curation: from curated data to curation of publicly available data.
- Open Data: `http://www.openstreetmap.org/?lat=53.38086&lon=-1.48545&zoom=17&layers=M`.

Open Data

- Automatic data curation: from curated data to curation of publicly available data.
- Open Data: `http://www.openstreetmap.org/?lat=53.38086&lon=-1.48545&zoom=17&layers=M`.

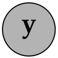
- Social network data, music information (Spotify), exercise.

Not Wrong ... Just Useless

- Here's a model that's not wrong ...

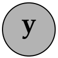
Not Wrong ... Just Useless

- Here's a (graphical) model that's not wrong ...



Not Wrong ... Just Useless

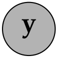
- Here's a model that's not wrong ...



... it's just useless.

Not Wrong ... Just Useless

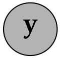
- Here's a model that's not wrong ...



  ... it's just useless.
- Does that imply all models that are not wrong are useless?

Not Wrong ... Just Useless

- Here's a model that's not wrong ...

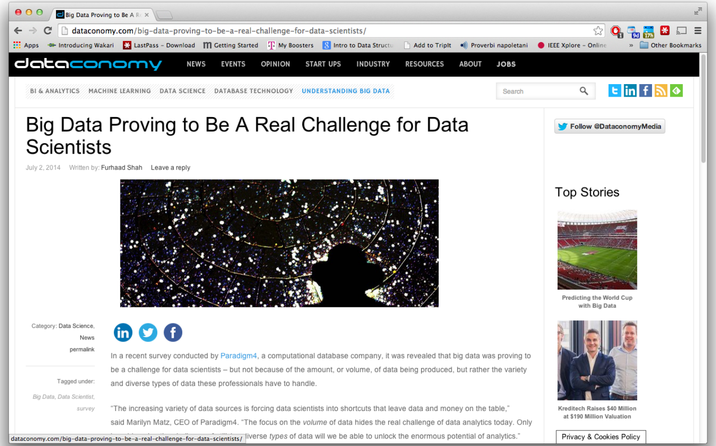


  ... it's just useless.

- Does that imply all models that are not wrong are useless?
- What is the minimum we can say about our data to get something useful?

Outline

Not the Scale it's the Diversity



Massive Missing Data

- If missing at random it can be marginalized.
- As data sets become very large (39 million in EMIS) data becomes extremely sparse.
- Imputation becomes impractical.

Missing Data

- If missing at random it can be marginalized.
- As data sets become very large (39 million in EMIS) data becomes extremely sparse.
- Imputation becomes impractical.

Imputation

- Expectation Maximization (EM) is gold standard imputation algorithm.
- Exact EM optimizes the log likelihood.
- Approximate EM optimizes a lower bound on log likelihood.
    - e.g. variational approximations (VIBES, Infer.net).
- Convergence is *guaranteed* to a local maxima in log likelihood.

Expectation Maximization

Require: An initial guess for missing data

Expectation Maximization

Require: An initial guess for missing data
repeat

Expectation Maximization

Require: An initial guess for missing data

repeat

Update model parameters (M-step)

Expectation Maximization

Require: An initial guess for missing data

repeat

Update model parameters (M-step)

Update guess of missing data (E-step)

Expectation Maximization

Require: An initial guess for missing data
  repeat
    Update model parameters (M-step)
    Update guess of missing data (E-step)
  until convergence

Imputation is Impractical

- In very sparse data imputation is impractical.
- EMIS: 39 million patients, thousands of tests.
- For most people, most tests are missing.
- M-step becomes confused by poor imputation.

Direct Marginalization is the Answer

- Perhaps we need joint distribution of two test outcomes,

$$p(y_1, y_2)$$

- Obtained through marginalizing over all missing data,

$$p(y_1, y_2) = \int p(y_1, y_2, y_3, \ldots, y_p) \mathrm{d}y_3, \ldots \mathrm{d}y_p$$

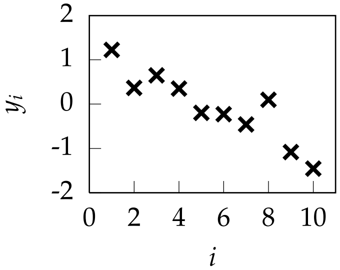
- Where $y_3, \ldots, y_p$ contains:
  1. all tests not applied to this patient
  2. all tests not yet invented!!

Magical Marginalization in Gaussians

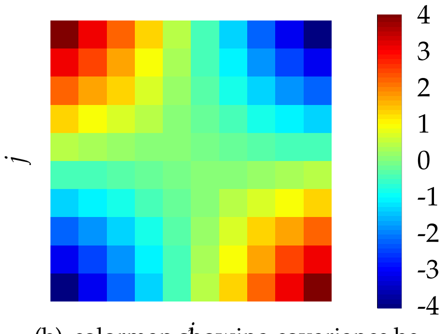
Multi-variate Gaussians

- Given 10 dimensional multivariate Gaussian, $\mathbf{y} \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$.
- Generate a single correlated sample $\mathbf{y} = [y_1, y_2 \ldots y_{10}]$.
- How do we find the marginal distribution of y_1, y_2?

Gaussian Marginalization Property



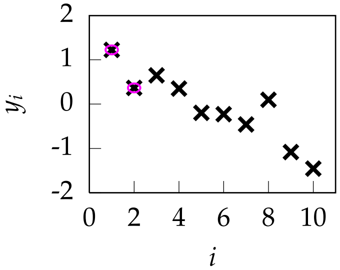
(a) A 10 dimensional sample



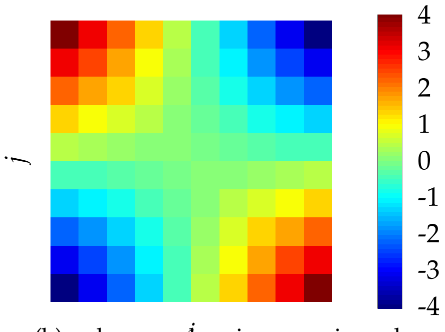
(b) colormap showing covariance between dimensions.

Figure : A sample from a 10 dimensional correlated Gaussian distribution.

Gaussian Marginalization Property



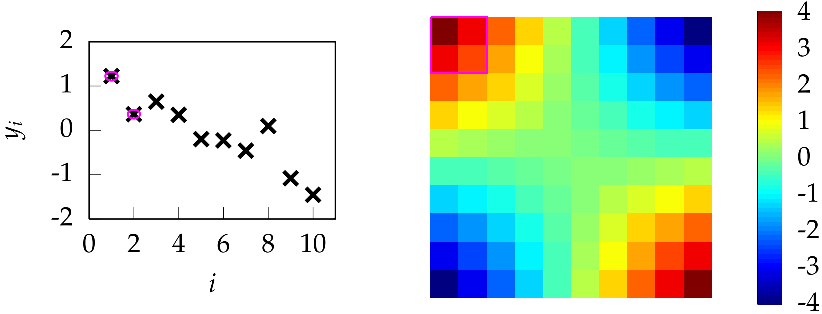
(a) A 10 dimensional sample



(b) colormap showing covariance between dimensions.

Figure : A sample from a 10 dimensional correlated Gaussian distribution.

Gaussian Marginalization Property

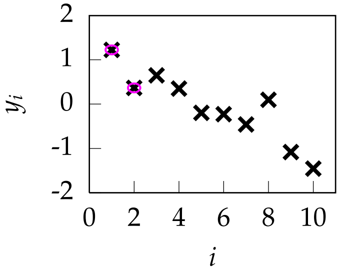


(a) A 10 dimensional sample

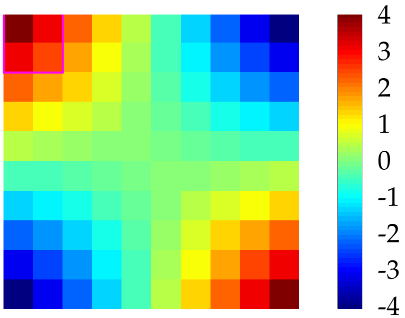
(b) colormap showing covariance between dimensions.

Figure : A sample from a 10 dimensional correlated Gaussian distribution.

Gaussian Marginalization Property



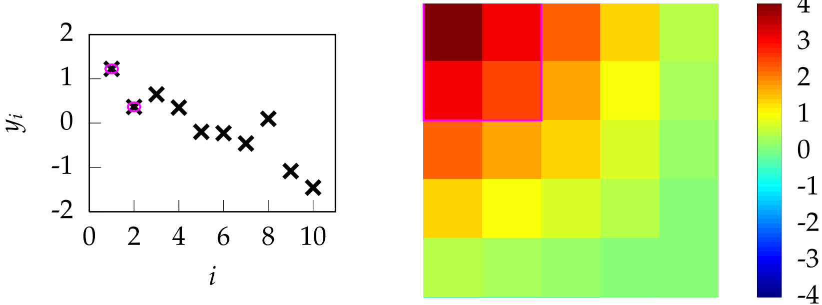
(a) A 10 dimensional sample



(b) colormap showing covariance between dimensions.

Figure : A sample from a 10 dimensional correlated Gaussian distribution.

Gaussian Marginalization Property

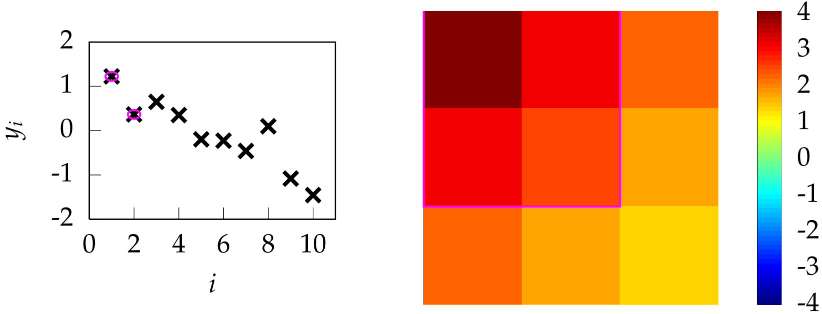


(a) A 10 dimensional sample

(b) colormap showing covariance between dimensions.

Figure : A sample from a 10 dimensional correlated Gaussian distribution.

Gaussian Marginalization Property

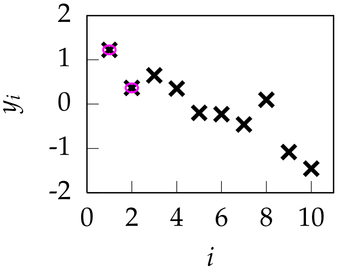


(a) A 10 dimensional sample

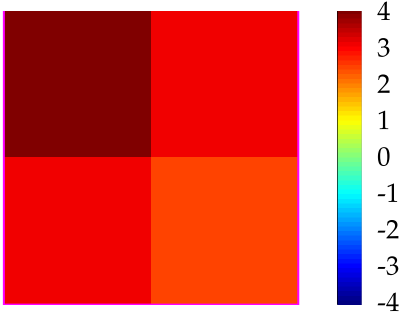
(b) colormap showing covariance between dimensions.

Figure : A sample from a 10 dimensional correlated Gaussian distribution.

Gaussian Marginalization Property



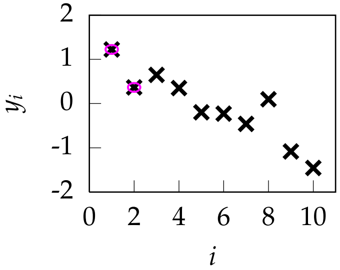
(a) A 10 dimensional sample



(b) colormap showing covariance between dimensions.

Figure : A sample from a 10 dimensional correlated Gaussian distribution.

Gaussian Marginalization Property



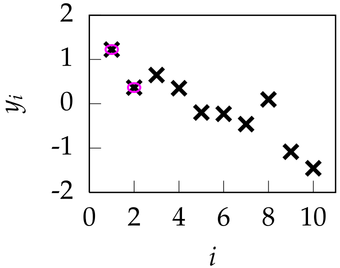
(a) A 10 dimensional sample

$$\begin{bmatrix} 4.1 & 3.1111 \\ 3.1111 & 2.5198 \end{bmatrix}$$

(b) covariance between y_1 and y_2.

Figure : A sample from a 10 dimensional correlated Gaussian distribution.

Gaussian Marginalization Property



(a) A 10 dimensional sample

$$\begin{bmatrix} 1 & 0.96793 \\ 0.96793 & 1 \end{bmatrix}$$

(b) correlation between y_1 and y_2.

Figure : A sample from a 10 dimensional correlated Gaussian distribution.

Avoid Imputation: Marginalize Directly

- Our approach: Avoid Imputation, Marginalize Directly.
- Explored in context of Collaborative Filtering.
- Similar challenges:
  - many users (patients),
  - many items (tests),
  - sparse data
- Implicitly marginalizes over all future tests too.

Work with Raquel Urtasun (Lawrence and Urtasun, 2009) and ongoing work with Max Zwießele and Nicoló Fusi.

Methods that Interrelate Covariates

- Need Class of models that interrelates data.
- Common assumption: high dimensional data lies on low dimensional manifold.
- Want to retain the marginalization property of Gaussians but deal with non-Gaussian data!

Linear Latent Variable Model

- Represent data, $\mathbf{Y}$, with a lower dimensional set of latent variables $\mathbf{X}$.
- Assume a linear relationship of the form

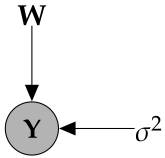
$$\mathbf{y}_{i,:} = \mathbf{W}\mathbf{x}_{i,:} + \boldsymbol{\epsilon}_{i,:},$$

where

$$\boldsymbol{\epsilon}_{i,:} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2\mathbf{I}\right).$$

Linear Latent Variable Model II

Probabilistic PCA Max. Likelihood Soln (Tipping and Bishop, 1999)



$$p\left(\mathbf{Y}|\mathbf{W}\right) = \prod_{i=1}^{n} \mathcal{N}\left(\mathbf{y}_{i,:}|\mathbf{0}, \mathbf{W}\mathbf{W}^{\top} + \sigma^2\mathbf{I}\right)$$

Linear Latent Variable Model II

Probabilistic PCA Max. Likelihood Soln (Tipping and Bishop, 1999)

$$p(\mathbf{Y}|\mathbf{W}) = \prod_{i=1}^{n} \mathcal{N}\left(\mathbf{y}_{i,:}|\mathbf{0}, \mathbf{C}\right), \quad \mathbf{C} = \mathbf{W}\mathbf{W}^{\top} + \sigma^2\mathbf{I}$$

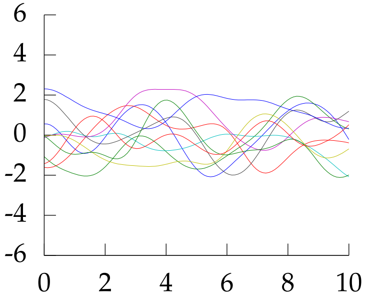
$$\log p(\mathbf{Y}|\mathbf{W}) = -\frac{n}{2}\log|\mathbf{C}| - \frac{1}{2}\mathrm{tr}\left(\mathbf{C}^{-1}\mathbf{Y}^{\top}\mathbf{Y}\right) + \text{const.}$$

If $\mathbf{U}_q$ are first q principal eigenvectors of $n^{-1}\mathbf{Y}^{\top}\mathbf{Y}$ and the corresponding eigenvalues are $\mathbf{\Lambda}_q$,

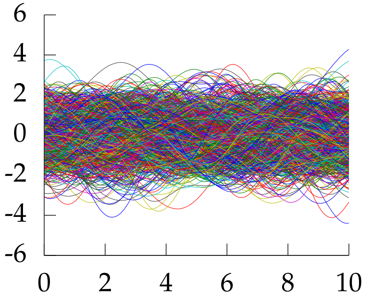
$$\mathbf{W} = \mathbf{U}_q\mathbf{L}\mathbf{R}^{\top}, \quad \mathbf{L} = \left(\mathbf{\Lambda}_q - \sigma^2\mathbf{I}\right)^{\frac{1}{2}}$$

where $\mathbf{R}$ is an arbitrary rotation matrix.

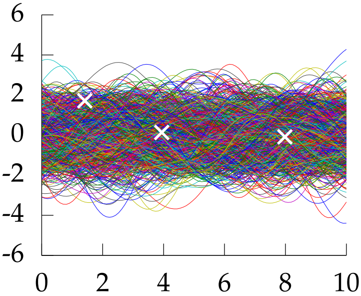
Gaussian Processes: Extremely Short Overview



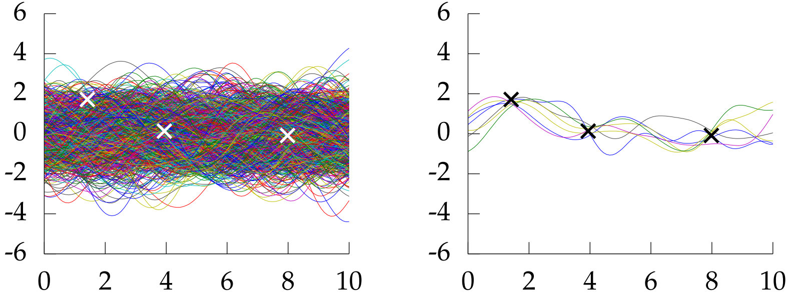
Gaussian Processes: Extremely Short Overview



Gaussian Processes: Extremely Short Overview



Gaussian Processes: Extremely Short Overview



Dealing with Non Gaussian Data

- Marginalization property of Gaussians very attractive.
- How to incorporate non-Gaussian data?
    - Data which isn't missing at random.
    - Binary data.
    - Ordinal categorical data.
    - Poisson counts.
    - Outliers.

Project Back into Gaussian

- Combine non-Gaussian likelihood with Gaussian prior.
- Either:
  - Project back to Gaussian posterior that is nearest in KL sense.
  - Expectation propagation.

- Or:
  - Fit a locally valid Gaussian approximation.
  - Laplace Approximation.

Ongoing work with Ricardo Andrade Pacheco (EP) and Alan Saul (Laplace) also James Hensman.

Gaussian Noise

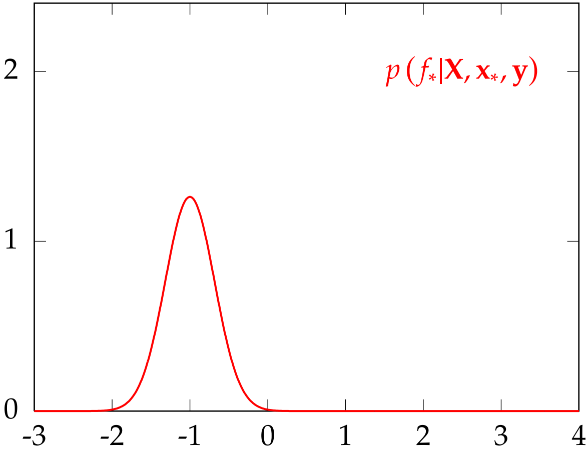


Figure : Inclusion of a data point with Gaussian noise.

Gaussian Noise

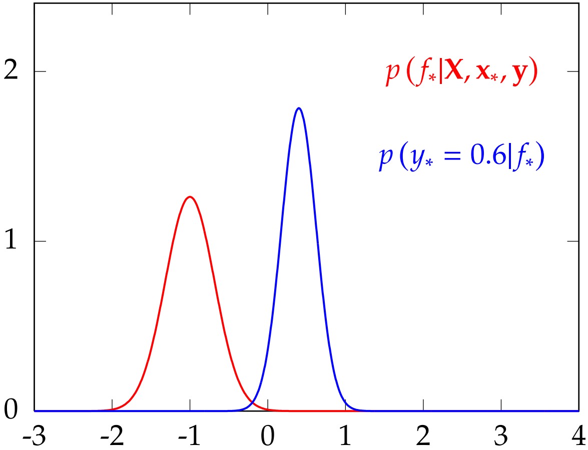


Figure : Inclusion of a data point with Gaussian noise.

Gaussian Noise

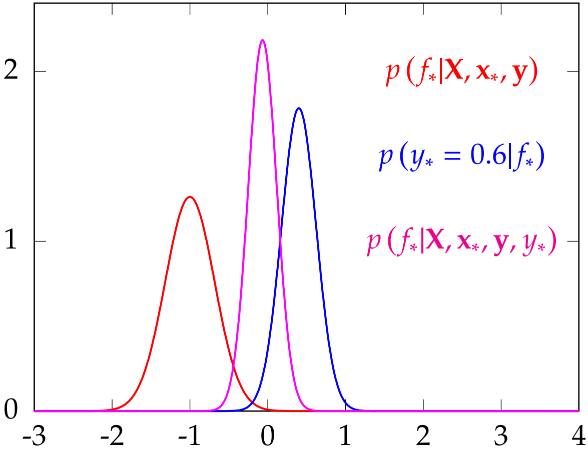


Figure : Inclusion of a data point with Gaussian noise.

Classification Noise Model

Probit Noise Model

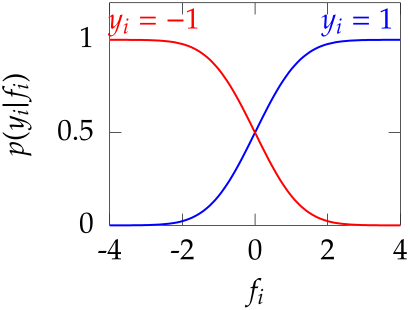


Figure : The probit model (classification). The plot shows $p(y_i|f_i)$ for different values of y_i. For $y_i = 1$ we have
$p(y_i|f_i) = \phi(f_i) = \int_{-\infty}^{f_i} \mathcal{N}(z|0, 1)\, \mathrm{d}z.$

Classification

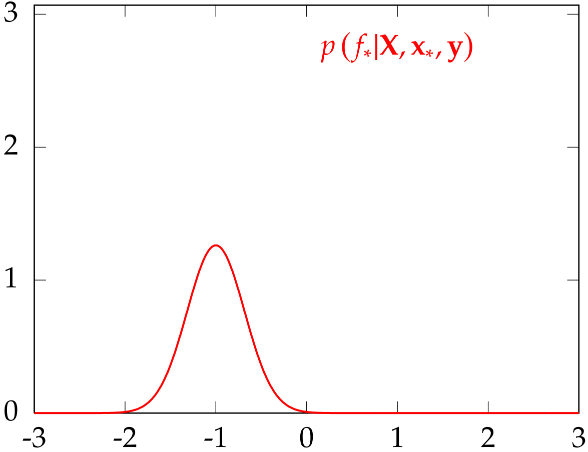


Figure : An EP style update with a classification noise model.

Classification

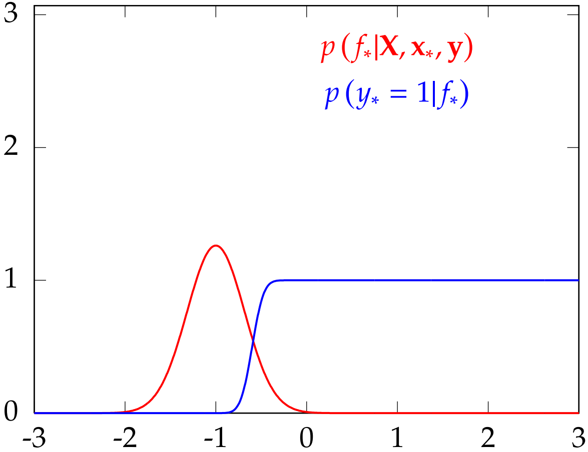


Figure : An EP style update with a classification noise model.

Classification

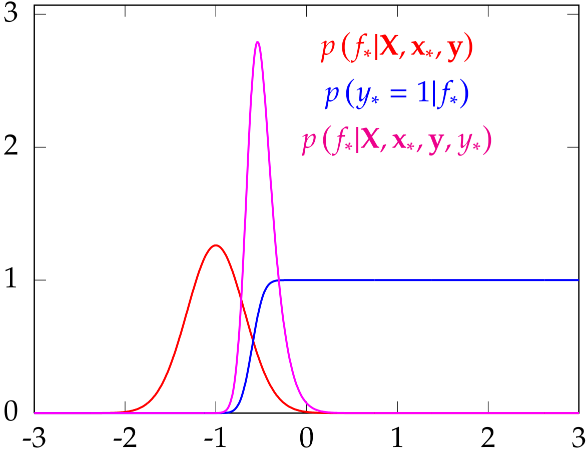


Figure : An EP style update with a classification noise model.

Classification

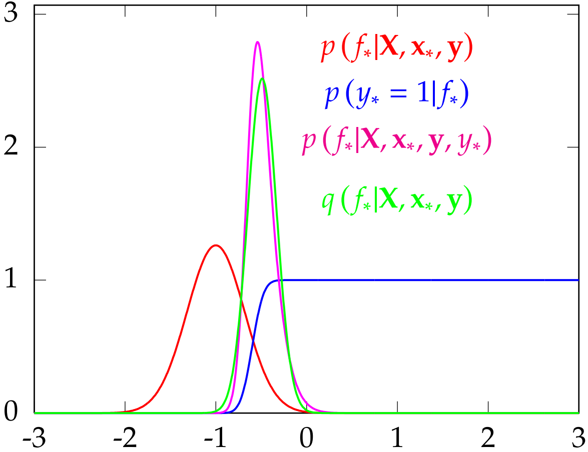


Figure : An EP style update with a classification noise model.

Ordinal Noise Model

Ordered Categories

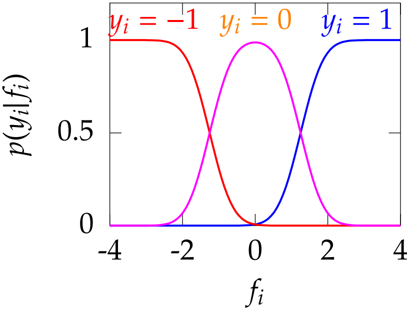


Figure : The ordered categorical noise model (ordinal regression). The plot shows $p\left(y_i|f_i\right)$ for different values of y_i. Here we have assumed three categories.

Other Challenges

- Spatial Data (workshops in November 2013 and January 2014with Peter Diggle, work with Ricardo Andrade Pacheco and John Quinn's group).

Survival Data

- Survival Data (work with Alan Saul and Aki Vehtari's group and HeRC).

Other Data

- Image Data (work with Teo de Campos, Fariba Yousefi, Zhenwen Dai, GaussianFace)
- Text Data (long time planned collaboration with Trevor Cohn)

Outline

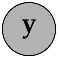
Inducing Variable Approximations

- Date back to (Williams and Seeger, 2001; Smola and Bartlett, 2001; Csató and Opper, 2002; Seeger et al., 2003; Snelson and Ghahramani, 2006). See Quiñonero Candela and Rasmussen (2005) for a review.
- We follow variational perspective of (Titsias, 2009).
- This is an augmented variable method, followed by a collapsed variational approximation (King and Lawrence, 2006; Hensman et al., 2012).

Augmented Variable Model: Not Wrong but Useful?

Augment standard model with a set of m new inducing variables, $\mathbf{u}$.

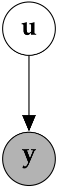
$$p(\mathbf{y}) = \int p(\mathbf{y}, \mathbf{u}) \mathrm{d}\mathbf{u}$$



Augmented Variable Model: Not Wrong but Useful?

Augment standard model with a set of m new inducing variables, $\mathbf{u}$.

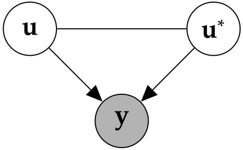
$$p(\mathbf{y}) = \int p(\mathbf{y}|\mathbf{u})p(\mathbf{u})\mathrm{d}\mathbf{u}$$



Augmented Variable Model: Not Wrong but Useful?

Important: Ensure inducing variables are *also* Kolmogorov consistent (we have m^* other inducing variables we are not *yet* using.)

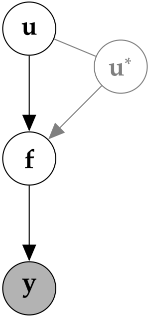
$$p(\mathbf{u}) = \int p(\mathbf{u}, \mathbf{u}^*) \mathrm{d}\mathbf{u}^*$$



Augmented Variable Model: Not Wrong but Useful?

Assume that relationship is through **f** (represents 'fundamentals'—push Kolmogorov consistency up to here).

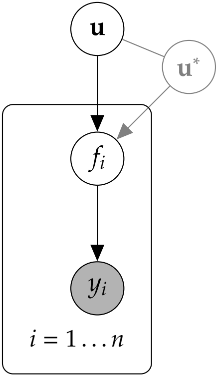
$$p(\mathbf{y}) = \int p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{u})p(\mathbf{u})\mathrm{d}\mathbf{f}\mathrm{d}\mathbf{u}$$



Augmented Variable Model: Not Wrong but Useful?

Convenient to assume factorization (*doesn't* invalidate model—think delta function as worst case).

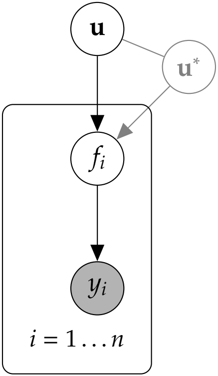
$$p(\mathbf{y}) = \int \prod_{i=1}^{n} p(y_i|f_i)p(\mathbf{f}|\mathbf{u})p(\mathbf{u})\mathrm{d}\mathbf{f}\mathrm{d}\mathbf{u}$$



Augmented Variable Model: Not Wrong but Useful?

Focus on integral over $\mathbf{f}$.

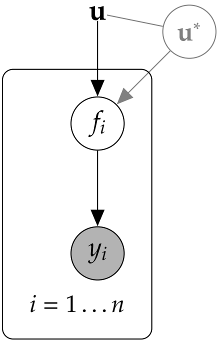
$$p(\mathbf{y}) = \int \int \prod_{i=1}^{n} p(y_i|f_i)p(\mathbf{f}|\mathbf{u})\mathrm{d}\mathbf{f}p(\mathbf{u})\mathrm{d}\mathbf{u}$$



Augmented Variable Model: Not Wrong but Useful?

Focus on integral over $\mathbf{f}$.

$$p(\mathbf{y}|\mathbf{u}) = \int \prod_{i=1}^{n} p(y_i|f_i) p(\mathbf{f}|\mathbf{u}) \mathrm{d}\mathbf{f}$$



Leads to Other Approximations ...

- Let's be explicity about storing approximate posterior of $\mathbf{u}$, $q(\mathbf{u})$.
- Now we have

$$p(\mathbf{y}^*|\mathbf{y}) = \int p(\mathbf{y}^*|\mathbf{u})q(\mathbf{u}|\mathbf{y})\mathbf{u}$$

- Inducing variables look a lot like regular parameters.
- *But*: their dimensionality does not need to be set at design time.
- They can be modified arbitrarily at run time without effecting the model likelihood.
- They only effect the quality of compression and the lower bound.

In GPs for Big Data

- Exploit the resulting factorization ...

$$p(\mathbf{y}^*|\mathbf{y}) = \int p(\mathbf{y}^*|\mathbf{u})q(\mathbf{u}|\mathbf{y})\mathbf{u}$$

In GPs for Big Data

- Exploit the resulting factorization ...

$$p(\mathbf{y}^*|\mathbf{y}) = \int p(\mathbf{y}^*|\mathbf{u})q(\mathbf{u}|\mathbf{y})\mathbf{u}$$

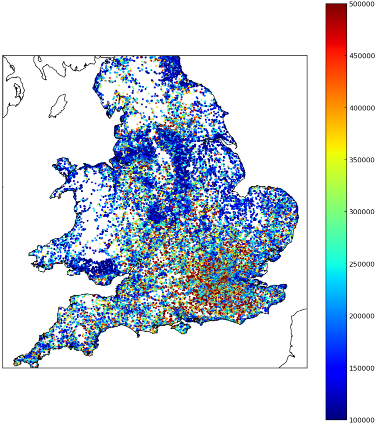
- The distribution now *factorizes*:

$$p(\mathbf{y}^*|\mathbf{y}) = \int \prod_{i=1}^{n^*} p(y_i^*|\mathbf{u})q(\mathbf{u}|\mathbf{y})\mathbf{u}$$

- This factorization can be exploited for stochastic variational inference (Hoffman et al., 2012).

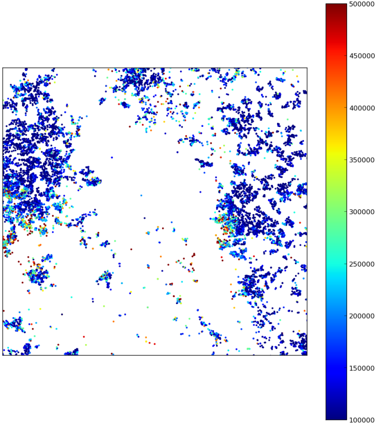
Nonparametrics for Very Large Data Sets

Modern data availability



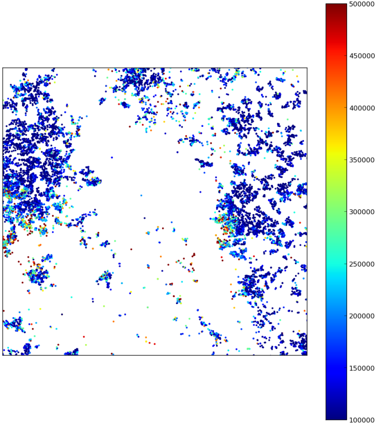
Nonparametrics for Very Large Data Sets

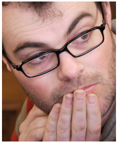
Proxy for index of deprivation?



Nonparametrics for Very Large Data Sets

Actually index of deprivation is a proxy for this ...





Hensman et al. (2013)

Gaussian Processes for Big Data

James Hensman* Dept. Computer Science The University of Sheffield Sheffield, UK

Nicolò Fusi* Dept. Computer Science The University of Sheffield Sheffield, UK

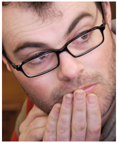
Neil D. Lawrence* Dept. Computer Science The University of Sheffield Sheffield, UK

Abstract

We introduce stochastic variational inference for Gaussian process models. This enables the application of Gaussian process (GP) models to data sets containing millions of data points. We show how GPs can be variationally decomposed to depend on a set

Even to accommodate these data sets, various approximate techniques are required. One approach is to partition the data set into separate groups [e.g. Snelson and Ghahramani, 2007, Urtasun and Darrell, 2008]. An alternative is to build a low rank approximation to the covariance matrix based around 'inducing variables' [see e.g. Csató and Opper, 2002, Seeger et al., 2003, Quiñonero Candela and Rasmussen, 2005, Tit-

Hensman et al. (2013)



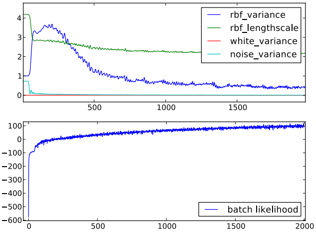


Figure 4: Convergence of the SVIGP algorithm on the two dimensional toy data

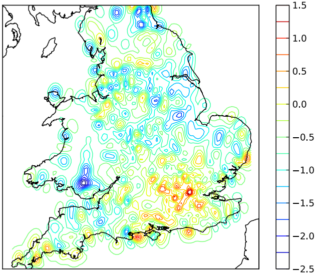


Figure 5: Variability of apartment price (logarithmically!) throughout England and Wales.

`land-registry-monthly-price-paid-data/`, which covers England and Wales, and filtered for apartments. This resulted in a data set with 75,000 entries,

ted a GP with the same covariance function as our

Prior and Likelihood Choice

- Choose a Gaussian process prior for $\mathbf{f}$.
  - This is not always correct, have a need for more flexible priors ... see Deep GPs (Damianou and Lawrence, 2013).
- Choose a factorized Gaussian likelihood for $\mathbf{y}|\mathbf{f}$.
  - Gaussian assumption can also be relaxed (Hensman et al., 2014).

Outline

Introduction

Data Heterogenity

Variational Compression

Process Composition

Conclusions

Mathematically

- Composite *multivariate* function

$$\mathbf{g}(\mathbf{x}) = \mathbf{f}_5(\mathbf{f}_4(\mathbf{f}_3(\mathbf{f}_2(\mathbf{f}_1(\mathbf{x})))))$$

Why Deep?

- Gaussian processes give priors over functions.
- Elegant properties:
  - e.g. *Derivatives* of process are also Gaussian distributed (if they exist).
- For particular covariance functions they are 'universal approximators', i.e. all functions can have support under the prior.
- Gaussian derivatives might ring alarm bells.
- E.g. a priori they don't believe in function 'jumps'.

Process Composition

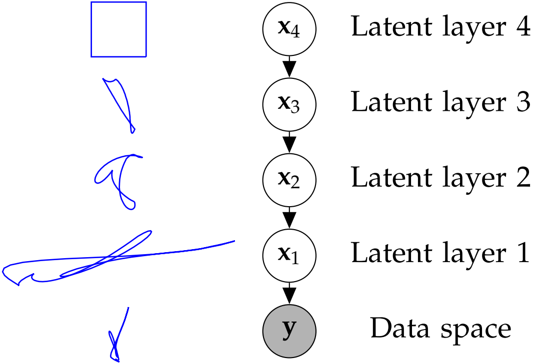
- From a process perspective: *process composition*.
- A (new?) way of constructing more complex *processes* based on simpler components.

Note: To retain *Kolmogorov consistency* introduce IBP priors over latent variables in each layer (Zhenwen Dai).

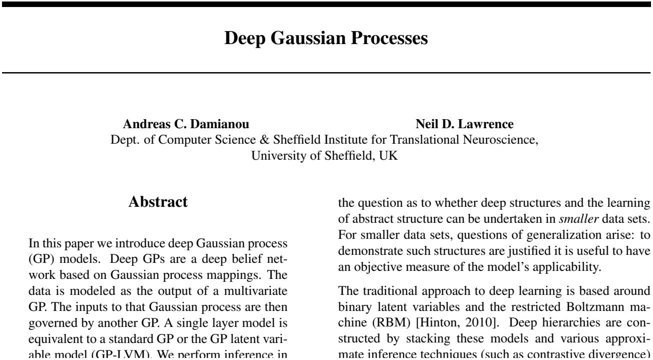
Analysis of Deep GPs

- Duvenaud et al. (2014) Duvenaud et al show that the derivative distribution of the process becomes more *heavy tailed* as number of layers increase.

Structures for Extracting Information from Data



Damianou and Lawrence (2013)



Deep Gaussian Processes

Andreas C. Damianou **Neil D. Lawrence**
Dept. of Computer Science & Sheffield Institute for Translational Neuroscience,
University of Sheffield, UK

Abstract

In this paper we introduce deep Gaussian process (GP) models. Deep GPs are a deep belief network based on Gaussian process mappings. The data is modeled as the output of a multivariate GP. The inputs to that Gaussian process are then governed by another GP. A single layer model is equivalent to a standard GP or the GP latent variable model (GP-LVM). We perform inference in

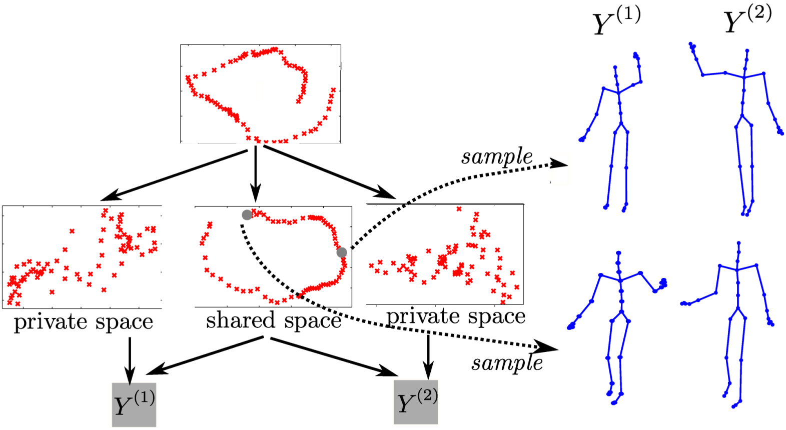
the question as to whether deep structures and the learning of abstract structure can be undertaken in *smaller* data sets. For smaller data sets, questions of generalization arise: to demonstrate such structures are justified it is useful to have an objective measure of the model's applicability.

The traditional approach to deep learning is based around binary latent variables and the restricted Boltzmann machine (RBM) [Hinton, 2010]. Deep hierarchies are constructed by stacking these models and various approximate inference techniques (such as contrastive divergence)

Motion Capture

- 'High five' data.
- Model learns structure between two interacting subjects.

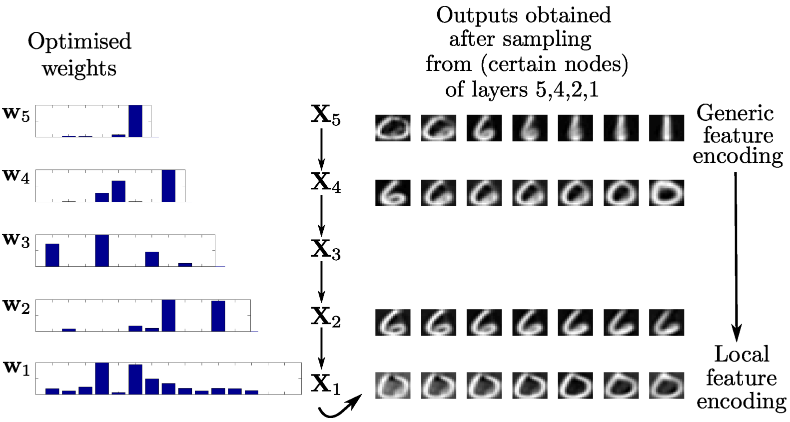
Deep hierarchies – motion capture



Digits Data Set

- Are deep hierarchies justified for small data sets?
- We can lower bound the evidence for different depths.
- For 150 6s, 0s and 1s from MNIST we found at least 5 layers are required.

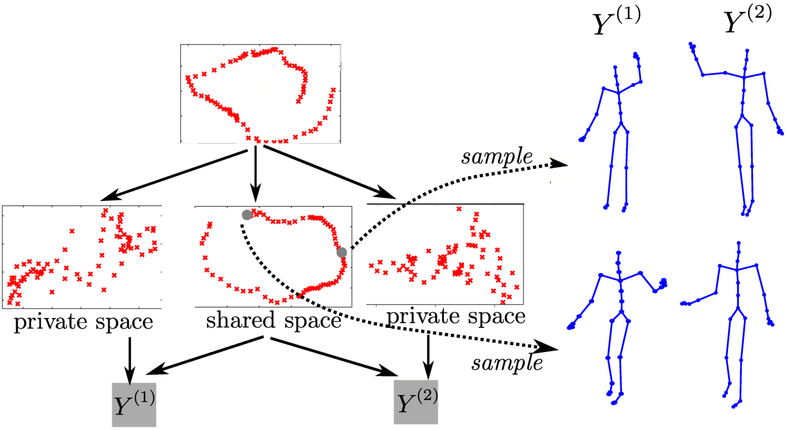
Deep hierarchies – MNIST



Motion Capture

- 'High five' data.
- Model learns structure between two interacting subjects.

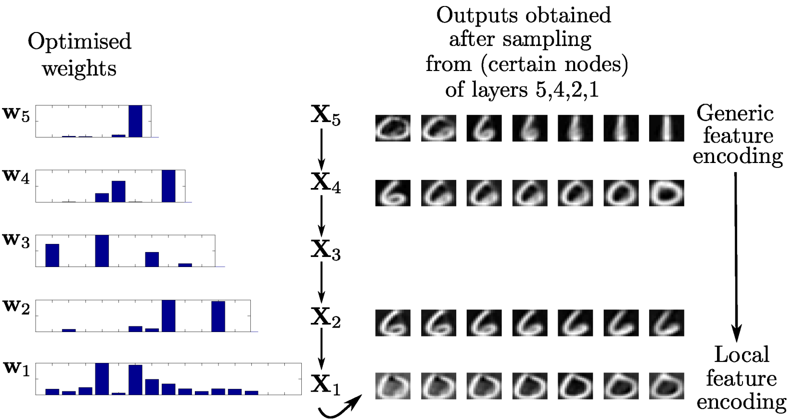
Deep hierarchies – motion capture



Digits Data Set

- Are deep hierarchies justified for small data sets?
- We can lower bound the evidence for different depths.
- For 150 6s, 0s and 1s from MNIST we found at least 5 layers are required.

Deep hierarchies – MNIST



What Can We Do that Internet Giants Can't?

- Google's resources give them access to volumes of data (or Facebook, or Microsoft, or Amazon).
- Is there anything for Universities to contribute?
- Assimilation of multiple views of the patient: each perhaps from a different patient.
- This may be done by small companies (with support of Universities).
- A Facebook app for your personalised health.
- These methodologies are part of that picture.

Challenges for Companies

- Trying to dominate the modern interconnected data market (e.g. Amazon, Google, Facebook) — buying up talent and competitors.
- or trying to exploit current 'data silos' (e.g. Tescos clubcard, Experian) — monetising our data today (limited shelf life?)
- or trying to understand their own systems (the internal google search)
- or new companies with new ideas that will generate data.

Challenges for Companies

- How do they break the natural data monopoly?
- How do they access the necessary expertise?

Challenges in Science

Data sharing is more widely accepted but:

- Most analysis is simple statistical tests or explorative modelling with PCA or clustering.
- Few scientists understand these methodologies, apply them as black box.
- There is an understanding gap between the data & scientist and the data scientist.

Challenges in Health

- Ensure the privacy of patients is respected.
- Leverage the wide range of data available for wider societal benefit.

International Development

- Exploit new telecommunications infrastructure to develop a leap-frog developed countries.
- Needs mechanisms for data sharing that retain the individual's control.
- Widespread education of *local* talent in code and model development.

Common Strands

- Improving access to data whilst balancing against individual's right to privacy against societal needs to advance.
- Advancing methodologies: development of methodologies needed to characterize large interconnected complex data sets.
- Analysis empowerment: giving scientists, clinicians, students, commercial and academic partners ability to analyze their own data with latest methodologies.

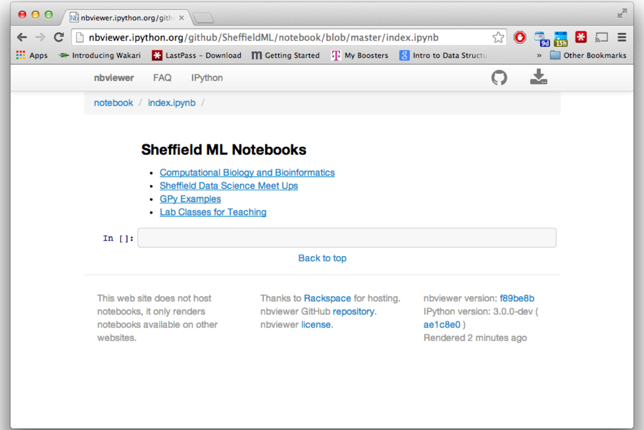
Open Data Science: A Magic Bullet?

- Make new methodologies available as widely and rapidly as possible with as few conditions on their use as possible.
- Educate commercial, scientific and medical partners in use of these methodologies.
- Act to achieve a balance between data sharing for societal benefit and right of an individual to own their own data.

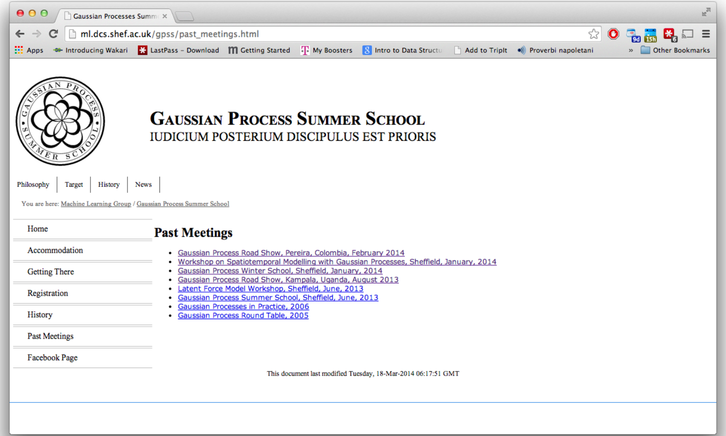
Achieving This

- Use BSD-like licenses on software.
- Educate our partners (summer schools, courses etc).
- Act to achieve a balance between data sharing for societal benefit and rights of the individual.

Make Analysis Available

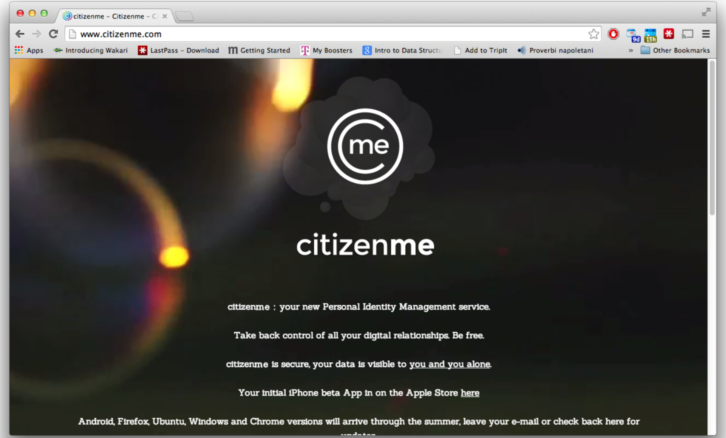


Educating



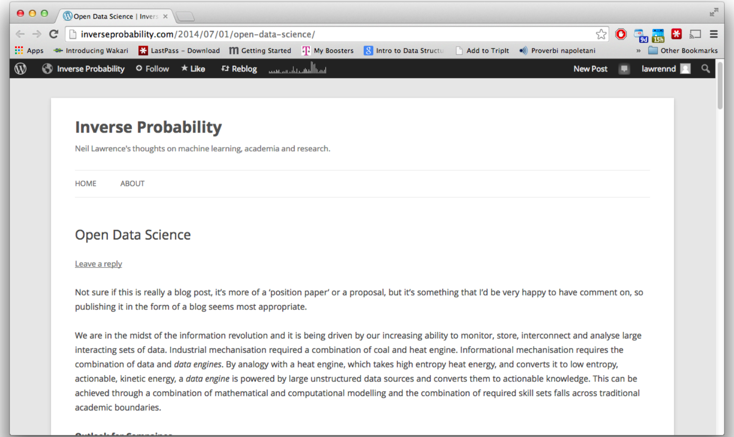
But we need to do much more!

Digital Identity and Data Ownership

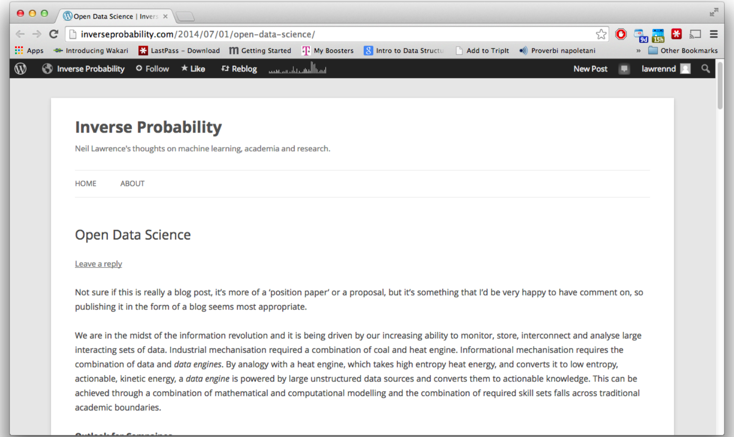


Data Warehousing

Blog Post


Open Data Science | Invers
inverseprobability.com/2014/07/01/open-data-science/
Apps
Introducing Wakari
LastPass - Download
Getting Started
My Boosters
Intro to Data Structu
Add to Triplt
Proverbi napoletani
Other Bookmarks
Inverse Probability
Follow
Like
Reblog
New Post
lawrennd
Inverse Probability
Neil Lawrence's thoughts on machine learning, academia and research.
HOME
ABOUT
Open Data Science
Leave a reply
Not sure if this is really a blog post, it's more of a 'position paper' or a proposal, but it's something that I'd be very happy to have comment on, so publishing it in the form of a blog seems most appropriate.
We are in the midst of the information revolution and it is being driven by our increasing ability to monitor, store, interconnect and analyse large interacting sets of data. Industrial mechanisation required a combination of coal and heat engine. Informational mechanisation requires the combination of data and data engines. By analogy with a heat engine, which takes high entropy heat energy, and converts it to low entropy, actionable, kinetic energy, a data engine is powered by large unstructured data sources and converts them to actionable knowledge. This can be achieved through a combination of mathematical and computational modelling and the combination of required skill sets falls across traditional academic boundaries.

Blog Post



Open Data Science | Invers
inverseprobability.com/2014/07/01/open-data-science/
Apps
Introducing Wakari
LastPass - Download
Getting Started
My Boosters
Intro to Data Structu
Add to Triplt
Proverbi napoletani
Other Bookmarks
Inverse Probability
Follow
Like
Reblog
New Post
lawrennd
Inverse Probability
Neil Lawrence's thoughts on machine learning, academia and research.
HOME
ABOUT
Open Data Science
Leave a reply
Not sure if this is really a blog post, it's more of a 'position paper' or a proposal, but it's something that I'd be very happy to have comment on, so publishing it in the form of a blog seems most appropriate.
We are in the midst of the information revolution and it is being driven by our increasing ability to monitor, store, interconnect and analyse large interacting sets of data. Industrial mechanisation required a combination of coal and heat engine. Informational mechanisation requires the combination of data and data engines. By analogy with a heat engine, which takes high entropy heat energy, and converts it to low entropy, actionable, kinetic energy, a data engine is powered by large unstructured data sources and converts them to actionable knowledge. This can be achieved through a combination of mathematical and computational modelling and the combination of required skill sets falls across traditional academic boundaries.


sign in subscribe search jobs dating more UK edition

theguardian
Winner of the Pulitzer prize

UK world sport football opinion culture economy lifestyle fashion environment tech money travel all

home

media network

Beware the rise of the digital oligarchy
Neil Lawrence

Powerful algorithms and the concentration of data in the hands means we need better models of data-ownership

0 comments

Eight lessons political parties need to learn to woo young voters
Matthew Hook

2 comments

Mobile World Congress 2015: what it means for marketing pros
James Hilton

How we made MailMen for Royal Mail

DISCOVER MAIL IN THE DIGITAL AGE AT MAILMEN .co.uk

PocketHighStreet: linking bricks and clicks at a local level

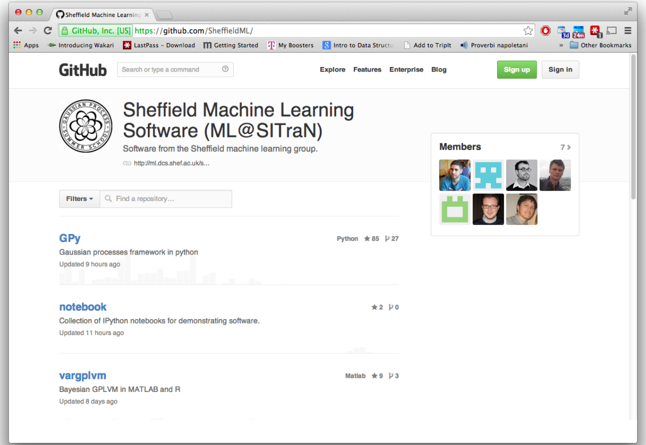
…ments

The return of the full-service agency approach
Olly Markeson

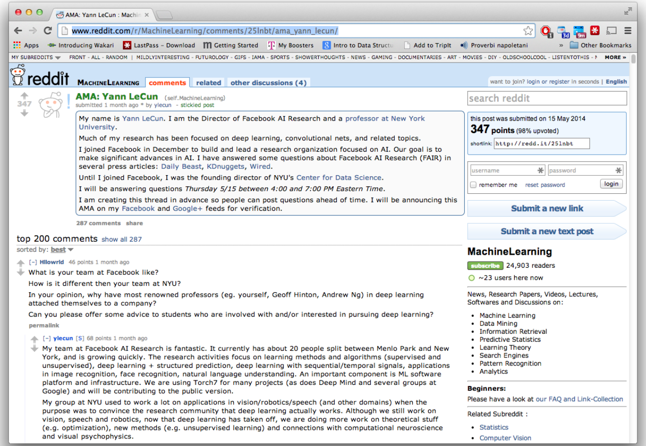
0 comments

.theguardian.com/media-network/2015/mar/05/political-parties-woo-young-voters-general-election

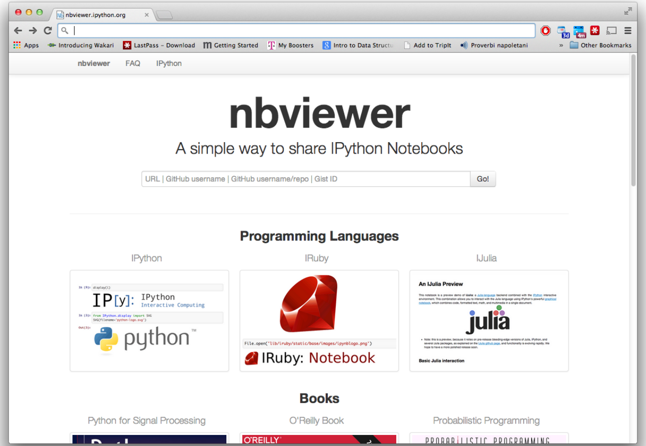
Modern Tools: Github



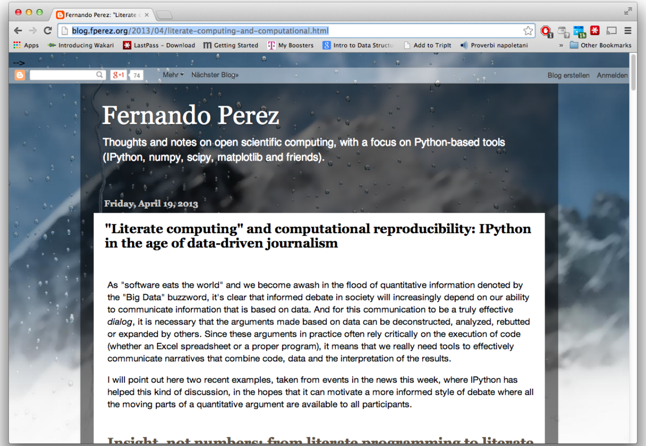
Modern Tools: Reddit



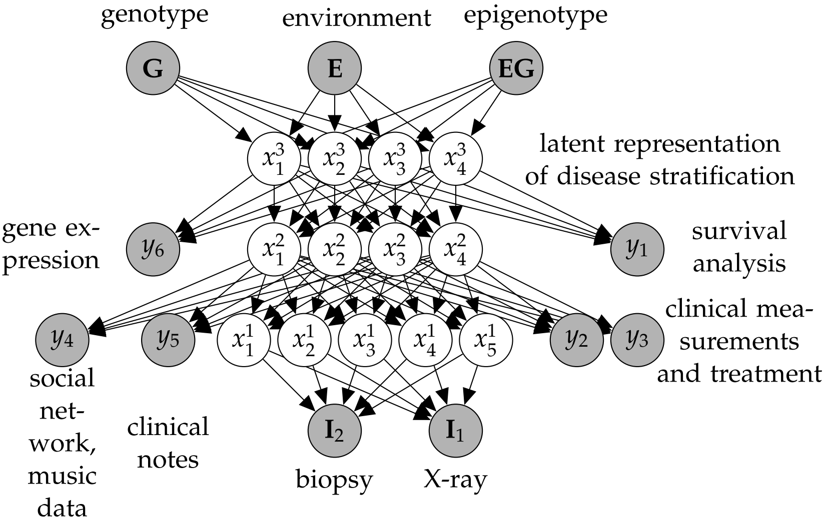
Modern Tools: IPython Notebook



Literate Computing



Deep Health



genotype
environment
epigenotype
G
E
EG
x_1^3
x_2^3
x_3^3
x_4^3
latent representation of disease stratification
gene ex- pression
y_6
x_1^2
x_2^2
x_3^2
x_4^2
y_1
survival analysis
clinical mea- surements and treatment
y_4
y_5
x_1^1
x_2^1
x_3^1
x_4^1
x_5^1
y_2
y_3
social net- work, music data
clinical notes
$\mathbf{I}_2$
$\mathbf{I}_1$
biopsy
X-ray

Summary

- ‘Big Data’ and simple models only takes us so far.
- Key question: what do we do when ‘Big Data’ is *small*.
- Examples include computational biology and personalised health.
- Our approach is *process composition* (e.g. (Damianou and Lawrence, 2013)).
- Developing approximate inference algorithms that scale for these models (e.g. (Hensman et al., 2013)).
- Intention is to deploy these models for assimilating a wide range of data types in personalized health (text, survival times, images, genotype, phenotype).
- Requires population scale models with millions of features.

References I

G. E. P. Box. Science and statistics. *Journal of the American Statistical Association*, 71(365), 1976.

L. Csató and M. Opper. Sparse on-line Gaussian processes. *Neural Computation*, 14(3):641–668, 2002.

A. Damianou and N. D. Lawrence. Deep Gaussian processes. In C. Carvalho and P. Ravikumar, editors, *Proceedings of the Sixteenth International Workshop on Artificial Intelligence and Statistics*, volume 31, AZ, USA, 2013. JMLR W&CP 31. [PDF].

D. Duvenaud, O. Rippel, R. Adams, and Z. Ghahramani. Avoiding pathologies in very deep networks. In Kaski and Corander (2014).

A. Y. Halevy, P. Norvig, and F. Pereira. The unreasonable effectiveness of data. *IEEE Intelligent Systems*, 24(2):8–12, 2009. [DOI].

J. Hensman, N. Fusi, and N. D. Lawrence. Gaussian processes for big data. In A. Nicholson and P. Smyth, editors, *Uncertainty in Artificial Intelligence*, volume 29. AUAI Press, 2013. [PDF].

J. Hensman, M. Rattray, and N. D. Lawrence. Fast variational inference in the conjugate exponential family. In P. L. Bartlett, F. C. N. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25, Cambridge, MA, 2012. [PDF].

J. Hensman, M. Zwiessele, and N. D. Lawrence. Tilted variational Bayes. In Kaski and Corander (2014).

M. Hoffman, D. M. Blei, C. Wang, and J. Paisley. Stochastic variational inference. Technical report,

S. Kaski and J. Corander, editors. *Artificial Intelligence and Statistics*, volume 33, Iceland, 2014. JMLR W&CP 33.

N. J. King and N. D. Lawrence. Fast variational inference for Gaussian Process models through KL-correction. In *ECML, Berlin, 2006*, Lecture Notes in Computer Science, pages 270–281, Berlin, 2006. Springer-Verlag. [PDF].

N. D. Lawrence and R. Urtasun. Non-linear matrix factorization with Gaussian processes. In L. Bottou and M. Littman, editors, *Proceedings of the International Conference in Machine Learning*, volume 26, San Francisco, CA, 2009. Morgan Kauffman. [PDF].

T. K. Leen, T. G. Dietterich, and V. Tresp, editors. *Advances in Neural Information Processing Systems*, volume 13, Cambridge, MA, 2001. MIT Press.

J. Quiñonero Candela and C. E. Rasmussen. A unifying view of sparse approximate Gaussian process regression. *Journal of Machine Learning Research*, 6:1939–1959, 2005.

References II

M. Seeger, C. K. I. Williams, and N. D. Lawrence. Fast forward selection to speed up sparse Gaussian process regression. In C. M. Bishop and B. J. Frey, editors, *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*, Key West, FL, 3–6 Jan 2003.

A. J. Smola and P. L. Bartlett. Sparse greedy Gaussian process regression. In Leen et al. (2001), pages 619–625.

E. Snelson and Z. Ghahramani. Sparse Gaussian processes using pseudo-inputs. In Y. Weiss, B. Schölkopf, and J. C. Platt, editors, *Advances in Neural Information Processing Systems*, volume 18, Cambridge, MA, 2006. MIT Press.

M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society, B*, 6 (3):611–622, 1999. [PDF]. [DOI].

M. K. Titsias. Variational learning of inducing variables in sparse Gaussian processes. In D. van Dyk and M. Welling, editors, *Proceedings of the Twelfth International Workshop on Artificial Intelligence and Statistics*, volume 5, pages 567–574, Clearwater Beach, FL, 16-18 April 2009. JMLR W&CP 5.

E. P. Wigner. The unreasonable effectiveness of mathematics in the natural sciences. *Communications on Pure and Applied Mathematics*, 13(1):1–14, 1960. [DOI].

C. K. I. Williams and M. Seeger. Using the Nyström method to speed up kernel machines. In Leen et al. (2001), pages 682–688.