Deep Health

Neil D. Lawrence

Manchester and Sheffield Learning Meetings

17th June 2013

Outline

Health

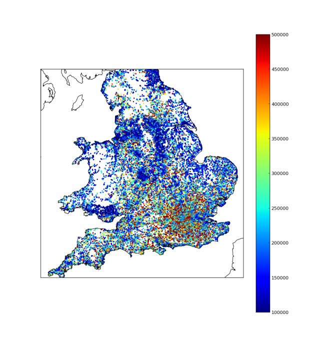
Data Heterogenity

Deep Learning

Conclusions

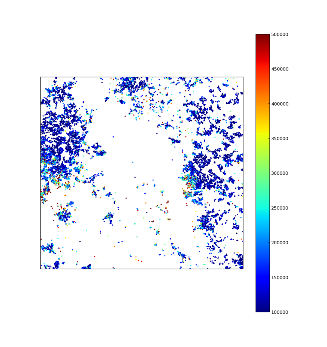
What's Changed (Changing) for Medicine?

- Modern data availability.



What's Changed (Changing) for Medicine?

- Modern data availability.



What's Changed (Changing) for Medicine?

- Google: patient data ...

What's Changed (Changing) for Medicine?

- The Red Flag Analogy.
- ... but why I work in Medical data ...

What's Changed (Changing) for Medicine?

- Genotyping.
- Epigenotyping.
- Transcriptome: detailed characterization of phenotype.
    - Self-organizing-stratifications of data.
- Automatic data curation: from curated data to curation of publicly available data.
- Patient Access: `http://www.patient.co.uk/patient-access.asp`
- Open Data: `http://www.openstreetmap.org/?lat=53.38086&lon=-1.48545&zoom=17&layers=M`.
- Tescos and Facebook.

Outline

Health

Data Heterogenity

Deep Learning

Conclusions

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

repeat

Update model parameters (M-step)

Update guess of missing data (E-step)

until convergence

Expectation Maximization

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

Expectation Maximization

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

Expectation Maximization

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

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,

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

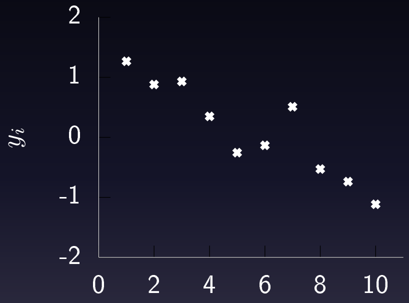
- Where $y_3, \ldots, y_p$ contains:
  - all tests not applied to this patient
  - all tests not yet invented!!

Magical Marginalization in Gaussians

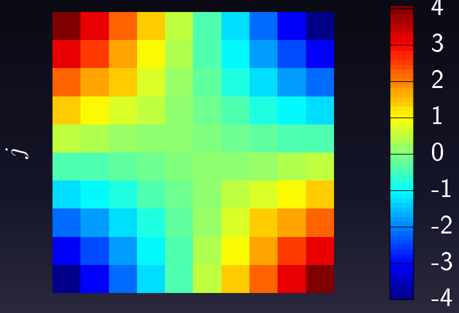
Multi-variate Gaussians

- 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



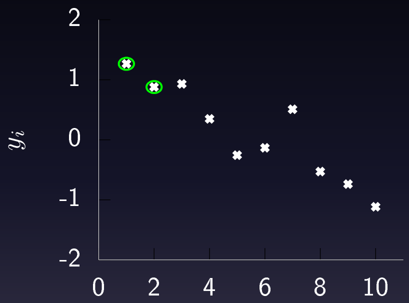
(a) A 10 dimensional sample



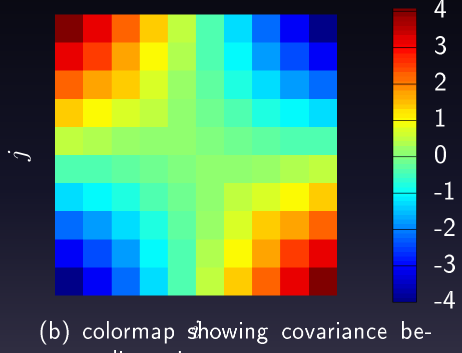
(b) colormap showing covariance between dimensions.

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

Gaussian Marginalization Property



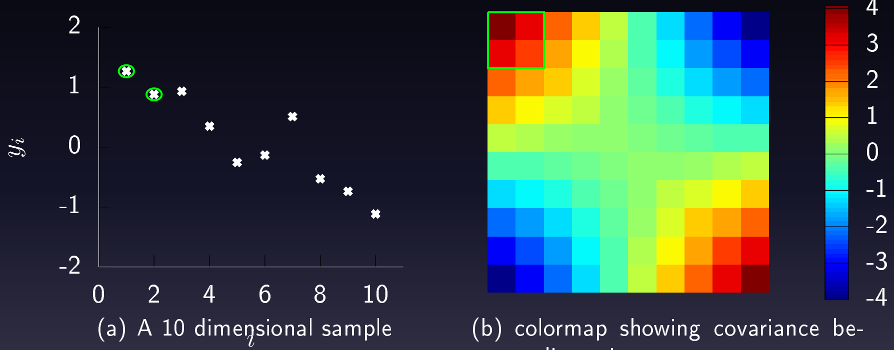
(a) A 10 dimensional sample



(b) colormap showing covariance between dimensions.

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

Gaussian Marginalization Property

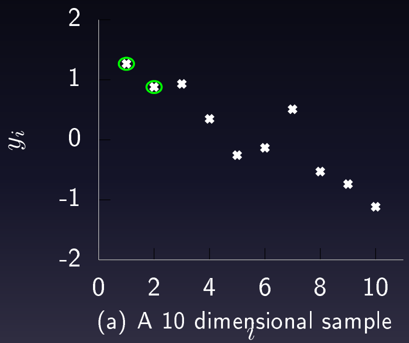


(a) A 10 dimensional sample

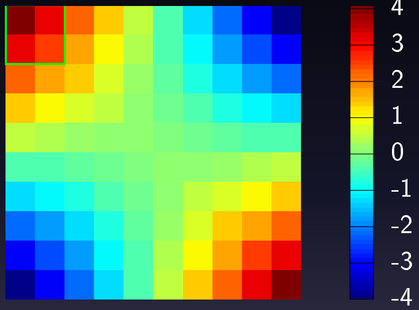
(b) colormap showing covariance between dimensions.

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

Gaussian Marginalization Property



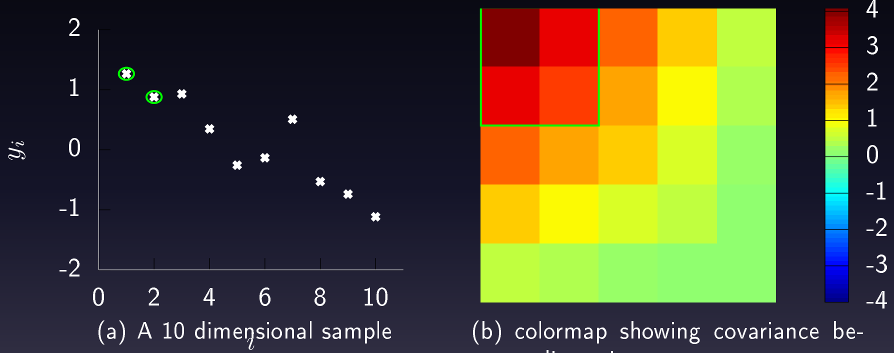
(a) A 10 dimensional sample



(b) colormap showing covariance between dimensions.

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

Gaussian Marginalization Property

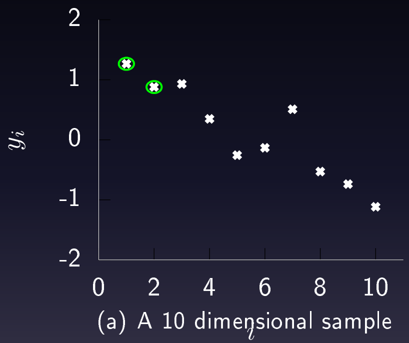


(a) A 10 dimensional sample

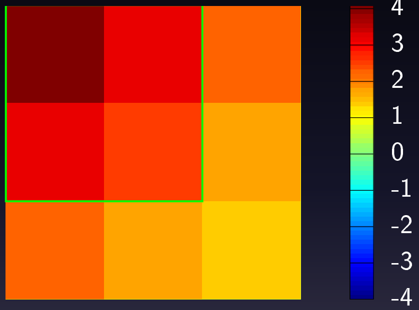
(b) colormap showing covariance between dimensions.

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

Gaussian Marginalization Property



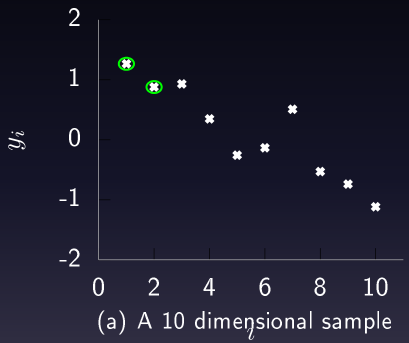
(a) A 10 dimensional sample



(b) colormap showing covariance between dimensions.

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

Gaussian Marginalization Property



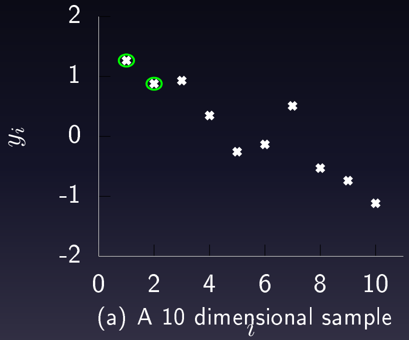
(a) A 10 dimensional sample

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

(b) colormap showing covariance between dimensions.

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

Gaussian Marginalization Property



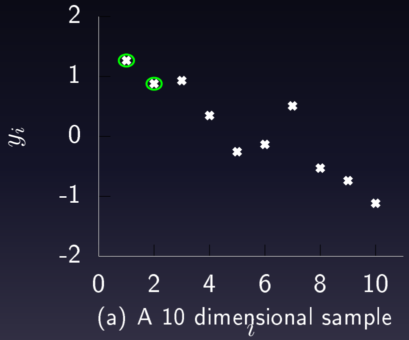
(a) A 10 dimensional sample

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

(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 recent submission with 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.

Linear Dimensionality Reduction

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

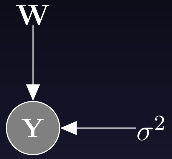
$$\mathbf{y}_{i,:} = \mathbf{W}\mathbf{x}_{i,:} + \boldsymbol{\epsilon}_{i,:},$$

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)

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

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

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.

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

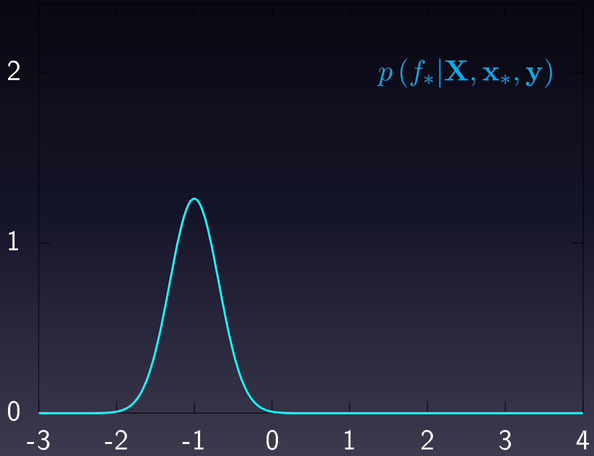


Figure: Inclusion of a data point with Gaussian noise.

Gaussian Noise

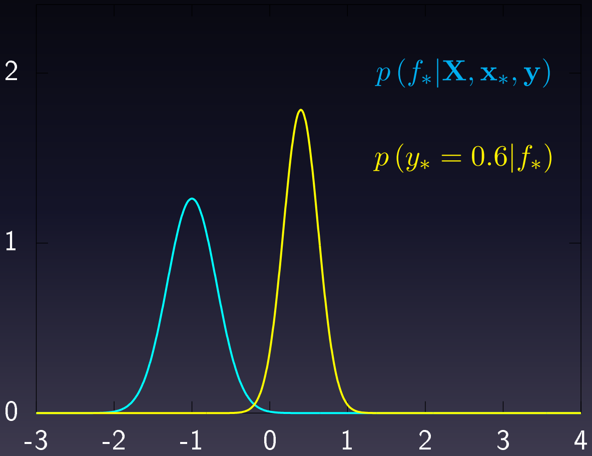


Figure: Inclusion of a data point with Gaussian noise.

Gaussian Noise

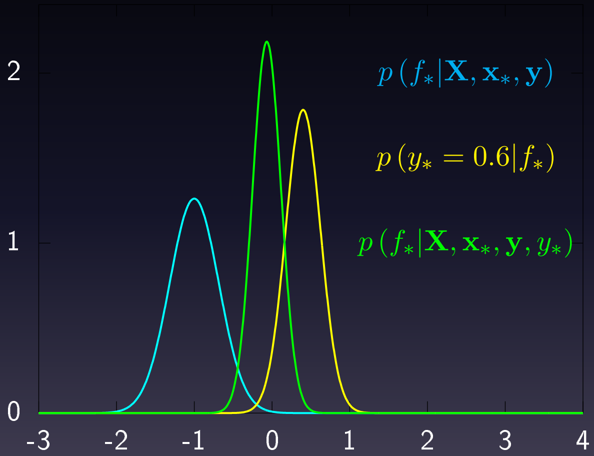


Figure: Inclusion of a data point with Gaussian noise.

Classification Noise Model

Probit Noise Model

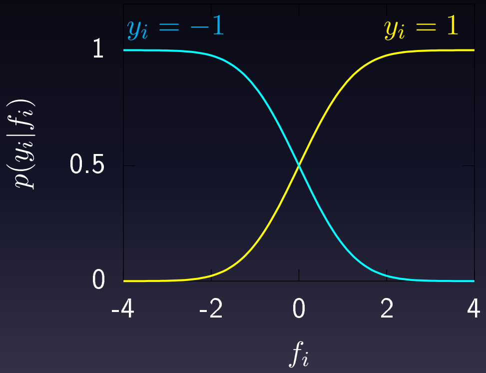


Figure: The probit model (classification). The plot shows $p\left(y_i|f_i\right)$ for different values of y_i. For $y_i = 1$ we have $p\left(y_i|f_i\right) = \phi\left(f_i\right) = \int_{-\infty}^{f_i} \mathcal{N}\left(z|0,1\right) \mathrm{d}z$.

Classification

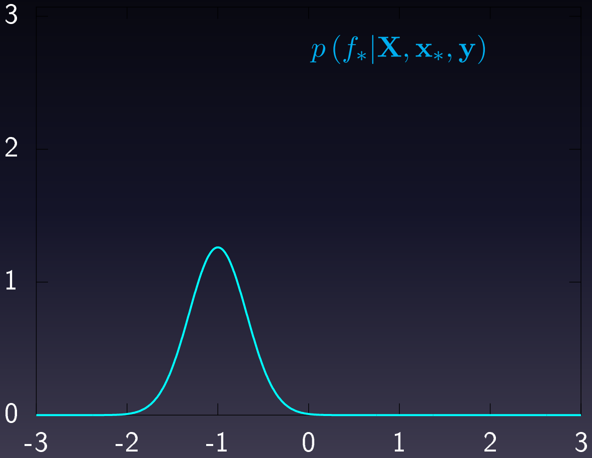


Figure: An EP style update with a classification noise model.

Classification

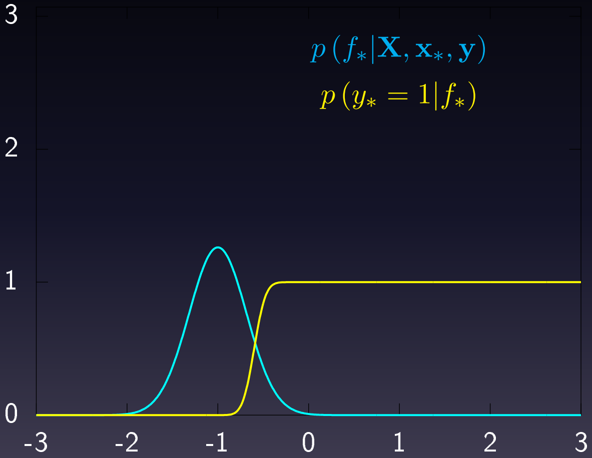


Figure: An EP style update with a classification noise model.

Classification

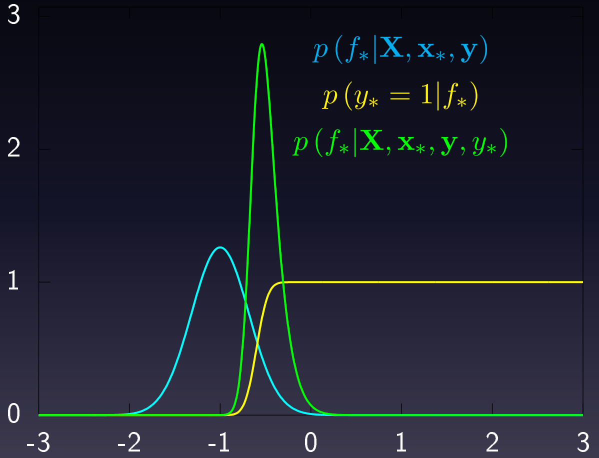


Figure: An EP style update with a classification noise model.

Classification

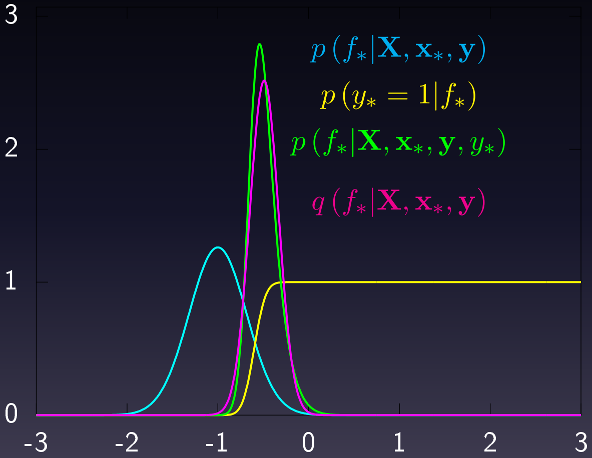


Figure: An EP style update with a classification noise model.

Ordinal Noise Model

Ordered Categories

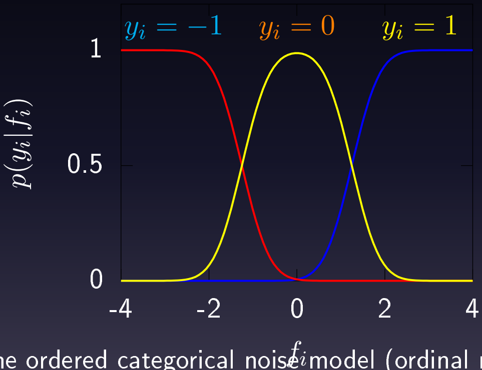


Figure: The ordered categorical noise model (ordinal regression). The plot shows $p(y_i|f_i)$ for different values of y_i. Here we have assumed three categories.

Ordinal Regression

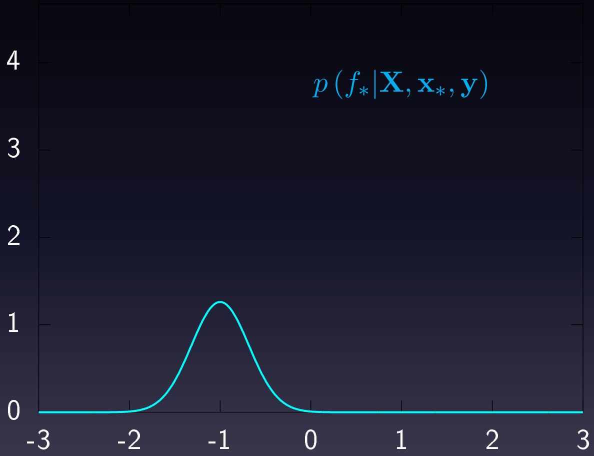


Figure: An EP style update with an ordered category noise model.

Return

Ordinal Regression

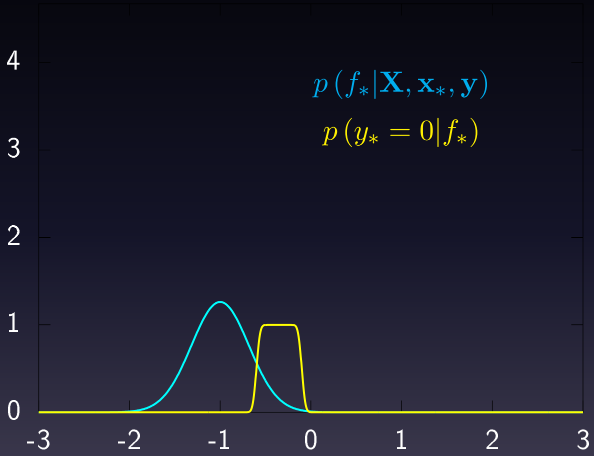


Figure: An EP style update with an ordered category noise model.

Return

Ordinal Regression

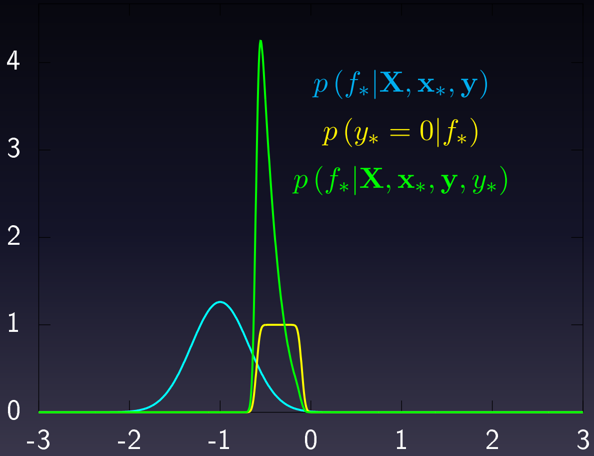


Figure: An EP style update with an ordered category noise model.

Return

Ordinal Regression

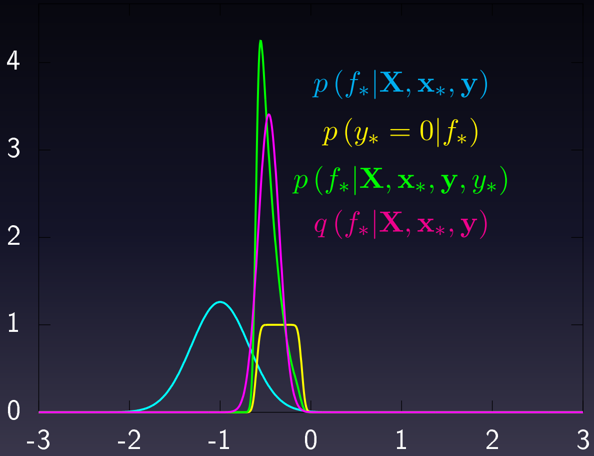


Figure: An EP style update with an ordered category noise model.

Return

Other Challenges

- Spatial Data (workshop in November with Peter Diggle, work with Ricardo Andrade Pacheco and John Quinn's group).
- Survival Data (work with Alan Saul and Aki Vehtari's group and HeRC).
- Image Data (work with Teo de Campos, Violet Snell and imminent arrival of Zhenwen Dai)
- Text Data (planned collaboration with Trevor Cohn)

Outline

Health

Data Heterogenity

Deep Learning

Conclusions

Scientists See Promise in Deep-Learning Programs

Hao Zhang/The New York Times

A voice recognition program translated a speech given by Richard F. Rashid, Microsoft's top scientist, into Mandarin Chinese.

By JOHN MARKOFF
Published: November 23, 2012

FACEBOOK
TWITTER
GOOGLE+
SAVE
E-MAIL
SHARE
PRINT

Using an artificial intelligence technique inspired by theories about how the brain recognizes patterns, technology companies are reporting startling gains in fields as diverse as computer vision, speech recognition and the identification of promising new molecules for designing drugs.

点击查看本文中文版。

The advances have led to widespread enthusiasm among researchers who design software to perform human

Connect With

What's Popular Now

King Abdullah of Jordan Has Criticism for All Concerned


7 Marines Killed in Nevada Training Exercise


MOST E-MAILED | MOST VIEWED


1. WELL
Lost Sleep Can Lead to Weight Gain


2. THIS LIFE
The Stories That Bind Us


3. WELL
A New Approach to Hip Surgery


4. Unwanted Electronic Gear Rising in Toxic Piles


5. DAVID BROOKS
The Progressive Shift


6. CONTINUING EDUCATION SPECIAL SECTION
A Gray Jobs Market for All Ages


7. Vatican's Bureaucracy Tests Even the Infallible

THE NEW YORKER

NEWS DESK

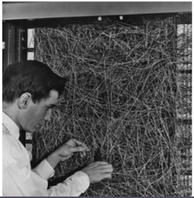
Reporting the latest on Washington and the world.

« How Susan Rice Sees the World | Main | Moral Machines »

NOVEMBER 25, 2012

IS "DEEP LEARNING" A REVOLUTION IN ARTIFICIAL INTELLIGENCE?

POSTED BY GARY MARCUS



Can a new technique known as deep learning revolutionize artificial intelligence, as yesterday's front-page article at the New York *Times* suggests? There is good reason to be excited about deep learning, a sophisticated "machine learning" algorithm that far exceeds many of its predecessors in its abilities to recognize syllables and images. But there's also good reason to be skeptical. While the *Times* reports that "advances in an artificial intelligence technology that can recognize patterns

1. Andy Borowitz: Cheney Marks Tenth Anniversary of Pretending There Was Reason to Invade Iraq
2. Amy Davidson: Life After Steubenville
3. William Finnegan: Gina Rinehart, Australia's Mining Billionaire
4. Maria Bustillos: On Video Games and Storytelling: An Interview with Tom Bissell
5. Lena Dunham: Lifelong Canine Cravings

Google Hires Geoffrey Hin × | Google Hires Brains that × | Geoffrey Hinton - Google ×

www.seroundtable.com/google-hires-geoffrey-hinton-16499.html

My Boosters | LastPass - Dow... | Add to TripIt | Proverbi napol... | IEEE Xplore - On... | Google Maps | Other Bookmarks

Google To Expand Knowledge Graph Through Hire Of Geoffrey Hinton

‹ PREV STORY | NEXT STORY ›

49 Tweet | 10 +1 | 16 Like | SHARE

Mar 14, 2013 • 8:23 am | (10)
by Barry Schwartz | Filed Under Google Search Engine

If I had to place one search priority above all else, I'd say right now, Google's most ambitious project is the knowledge graph. Yea, they are pushing Google+ big time, but the knowledge graph is a level above all of that technically.

Of course, Google has an outstanding team working on this project lead by one of the smartest people I've ever met Amit Singhal.

To take the knowledge graph to the next level, Google has hired/acquired Geoffrey Hinton and his team at DNNresearch. Geoffrey posted a note on his Google+ page about it:

> Last summer, I spent several months working with Google's Knowledge team in Mountain View, working with Jeff Dean and an incredible group of scientists and engineers who have a real shot at making spectacular progress in machine learning. Together with two of my recent graduate students, Ilya Sutskever and Alex Krizhevsky (who won the 2012 ImageNet competition), I am betting on Google's team to be the epicenter of future breakthroughs. That means we'll soon be joining Google to work with some of the smartest engineering minds to tackle some of the biggest challenges in computer science. I'll remain part-time at the University of Toronto, where I still have a lot of excellent graduate students, but at Google I will get to see what we can do with very large-scale computation.

I know we just scratched the surface of the knowledge graph and I am excited to see where it takes us in the future.

I am just glad I don't have to figure out how to get us there. I get to just sit and enjoy the ride.

SEARCH BUZZ VIDEO | Subscribe

SUBSCRIBE › | MORE VIDEOS › | VIDEO DETAILS ›

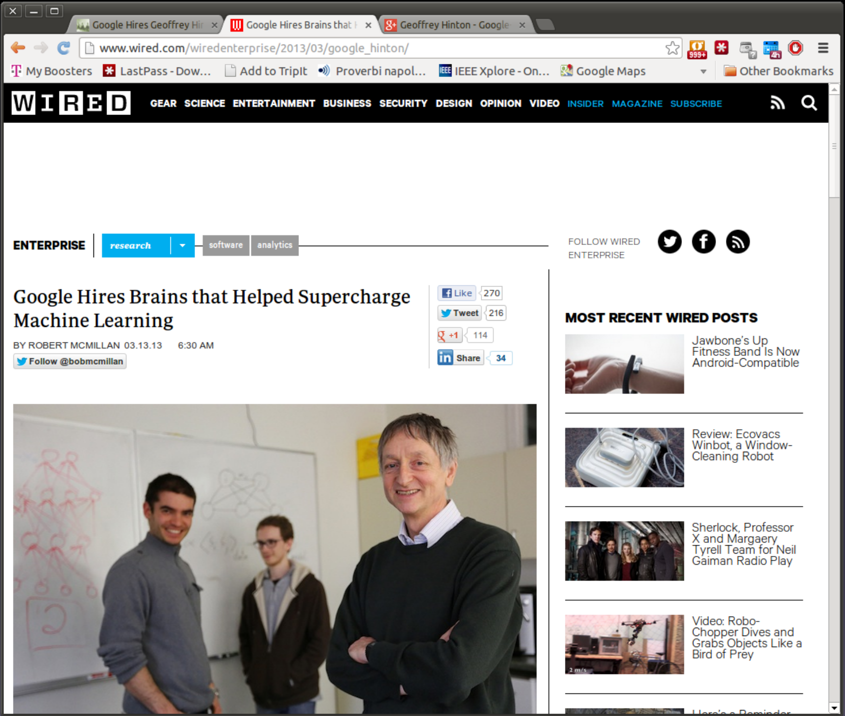
ROUNDTABLE SPONSORS

BROWSE BY:

- Browse by Date
- Find by Category
- Discover by Author
- Scan Most Recent
- See Comments
- View Tag Cloud

SEM FORUM THREADS

WebmasterWorld Forums



Google Hires Geoffrey Hin × Google Hires Brains that × Geoffrey Hinton - Google ×

www.wired.com/wiredenterprise/2013/03/google_hinton/

My Boosters LastPass - Dow... Add to TripIt Proverbi napol... IEEE Xplore - On... Google Maps Other Bookmarks

WIRED GEAR SCIENCE ENTERTAINMENT BUSINESS SECURITY DESIGN OPINION VIDEO INSIDER MAGAZINE SUBSCRIBE

ENTERPRISE | research | software | analytics

FOLLOW WIRED ENTERPRISE

Google Hires Brains that Helped Supercharge Machine Learning

BY ROBERT MCMILLAN 03.13.13 6:30 AM

Follow @bobmcmillan

Like 270 · Tweet 216 · +1 114 · Share 34

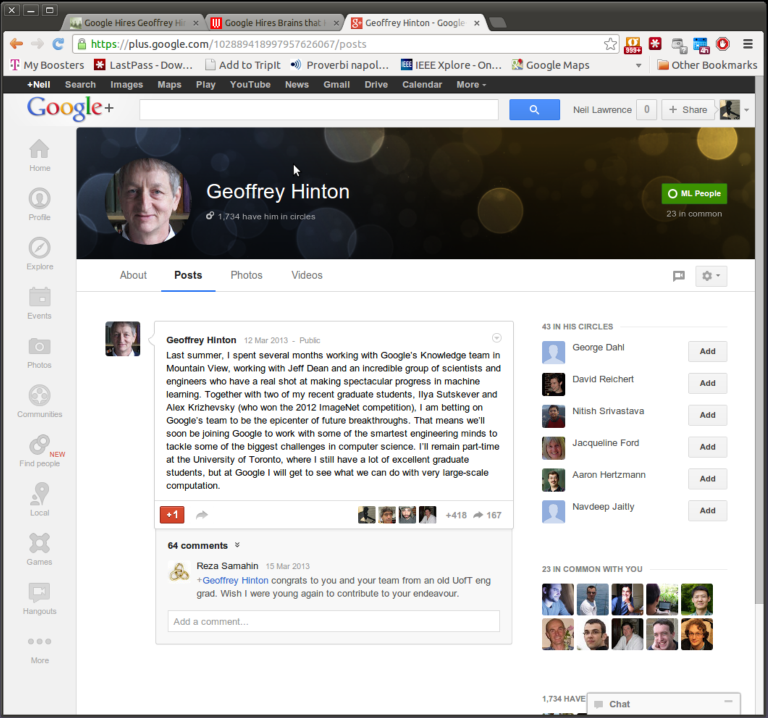
MOST RECENT WIRED POSTS

Jawbone's Up Fitness Band Is Now Android-Compatible

Review: Ecovacs Winbot, a Window-Cleaning Robot

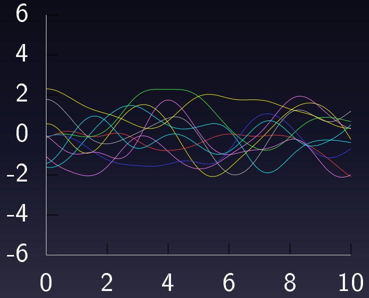
Sherlock, Professor X and Margaery Tyrell Team for Neil Gaiman Radio Play

Video: Robo-Chopper Dives and Grabs Objects Like a Bird of Prey

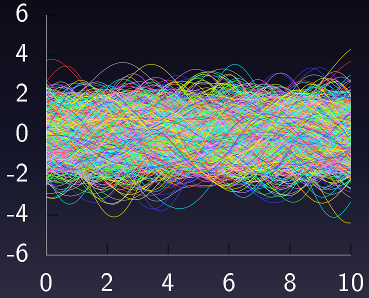


Google Hires Geoffrey Hi
Google Hires Brains that
Geoffrey Hinton - Google
https://plus.google.com/102889418997957626067/posts
My Boosters
LastPass - Dow...
Add to TripIt
Proverbi napol...
IEEE Xplore - On...
Google Maps
Other Bookmarks
+Neil Search Images Maps Play YouTube News Gmail Drive Calendar More
Google+
Neil Lawrence
0
Share
Home
Profile
Explore
Events
Photos
Communities
NEW
Find people
Local
Games
Hangouts
More
Geoffrey Hinton
1,734 have him in circles
ML People
23 in common
About Posts Photos Videos
Geoffrey Hinton 12 Mar 2013 - Public
Last summer, I spent several months working with Google's Knowledge team in Mountain View, working with Jeff Dean and an incredible group of scientists and engineers who have a real shot at making spectacular progress in machine learning. Together with two of my recent graduate students, Ilya Sutskever and Alex Krizhevsky (who won the 2012 ImageNet competition), I am betting on Google's team to be the epicenter of future breakthroughs. That means we'll soon be joining Google to work with some of the smartest engineering minds to tackle some of the biggest challenges in computer science. I'll remain part-time at the University of Toronto, where I still have a lot of excellent graduate students, but at Google I will get to see what we can do with very large-scale computation.
+1
+418 167
64 comments
Reza Samahin 15 Mar 2013
+Geoffrey Hinton congrats to you and your team from an old UofT eng grad. Wish I were young again to contribute to your endeavour.
Add a comment...
43 IN HIS CIRCLES
George Dahl Add
David Reichert Add
Nitish Srivastava Add
Jacqueline Ford Add
Aaron Hertzmann Add
Navdeep Jaitly Add
23 IN COMMON WITH YOU
1,734 HAVE
Chat

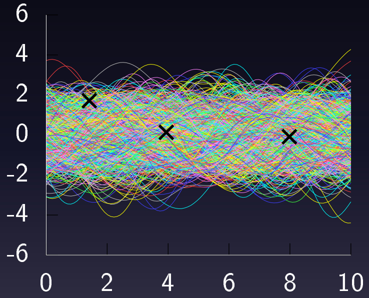
Gaussian Processes



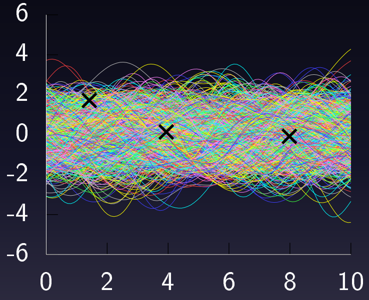
Gaussian Processes



Gaussian Processes



Gaussian Processes



direction for further research.

11.1. HAVE WE THROWN THE BABY OUT WITH THE BATH WATER?

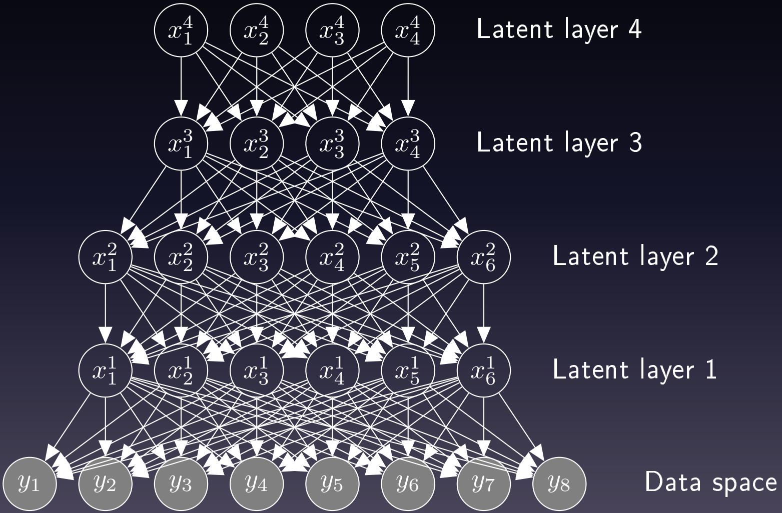
According to the hype of 1987, neural networks were meant to be intelligent models which discovered features and patterns in data. Gaussian processes in contrast are simply smoothing devices. How can Gaussian processes possibly replace neural networks? What is going on?

I think what the work of Williams and Rasmussen (1996) shows is that many real–world data modelling problems are perfectly well solved by sensible smoothing methods. The most interesting problems, the task of feature discovery for example, are not ones which Gaussian processes will solve. But maybe multilayer perceptrons can't solve them either. On the other hand, it may be that the limit of an infinite number of hidden units, to which Gaussian processes correspond, was a bad limit to take; maybe we should backtrack, or modify the prior on neural network parameters, so as to create new models more interesting than Gaussian processes. Evidence that this infinite limit has lost something compared with finite neural networks comes from the observation that in a finite neural network with more than one output, there are non–trivial correlations between the outputs (since they share inputs from common hidden units); but in the limit of an infinite number of hidden units, these correlations vanish. Radford Neal has suggested the use of non–Gaussian priors in networks with multiple hidden layers. Or perhaps a completely fresh start is needed, approaching the problem of machine learning from a paradigm different from the supervised feedforward mapping.

Structure of Priors

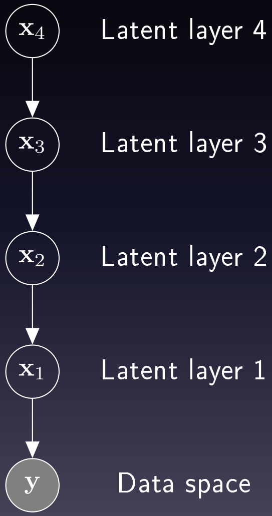
MacKay: NIPS Tutorial 1997 "Have we thrown out the baby with the bathwater?" (Published as MacKay, 1998) Also noted by (Wilson et al., 2012)

Deep Models

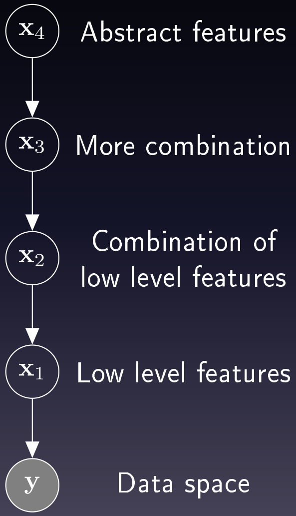


x_1^4 x_2^4 x_3^4 x_4^4
Latent layer 4
x_1^3 x_2^3 x_3^3 x_4^3
Latent layer 3
x_1^2 x_2^2 x_3^2 x_4^2 x_5^2 x_6^2
Latent layer 2
x_1^1 x_2^1 x_3^1 x_4^1 x_5^1 x_6^1
Latent layer 1
y_1 y_2 y_3 y_4 y_5 y_6 y_7 y_8
Data space

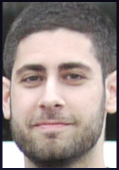
Deep Models


$\mathbf{x}_4$
Latent layer 4
$\mathbf{x}_3$
Latent layer 3
$\mathbf{x}_2$
Latent layer 2
$\mathbf{x}_1$
Latent layer 1
$\mathbf{y}$
Data space

Deep Models


$\mathbf{x}_4$
Abstract features
$\mathbf{x}_3$
More combination
$\mathbf{x}_2$
Combination of low level features
$\mathbf{x}_1$
Low level features
$\mathbf{y}$
Data space

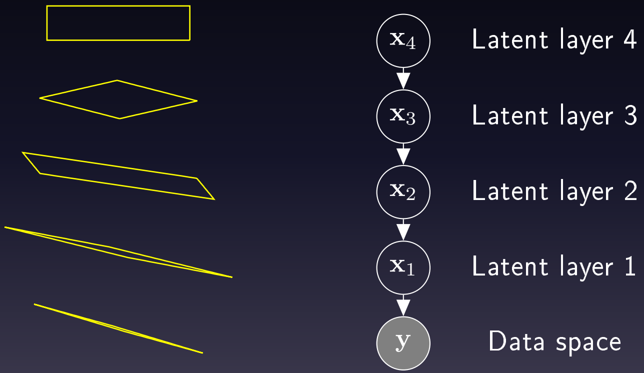
Deep Gaussian Processes



Damianou and Lawrence (2013)

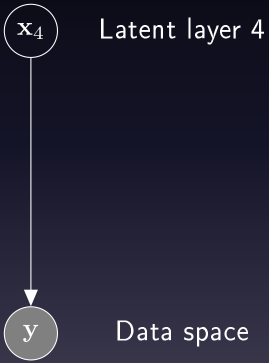
- Deep architectures allow abstraction of features (Bengio, 2009; Hinton and Osindero, 2006; Salakhutdinov and Murray, 2008).
- We use variational approach to stack GP models.

Stacked PCA



$\mathbf{x}_4$
Latent layer 4
$\mathbf{x}_3$
Latent layer 3
$\mathbf{x}_2$
Latent layer 2
$\mathbf{x}_1$
Latent layer 1
$\mathbf{y}$
Data space

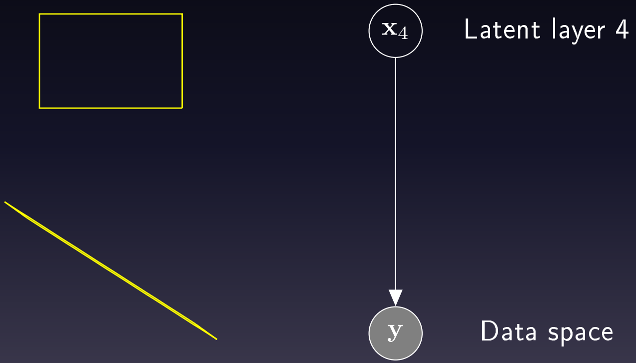
Stacked PCA



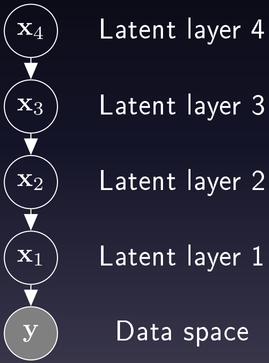
Stacked PCA

$\mathbf{x}_4$
Latent layer 4
$\mathbf{x}_3$
Latent layer 3
$\mathbf{x}_2$
Latent layer 2
$\mathbf{x}_1$
Latent layer 1
$\mathbf{y}$
Data space

Stacked PCA

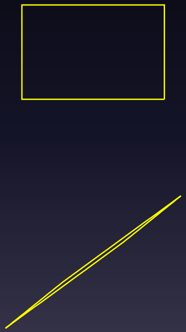


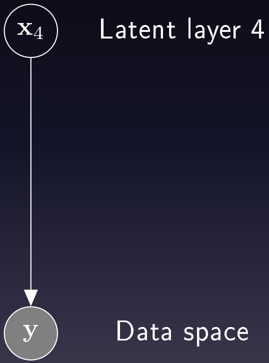
Stacked PCA



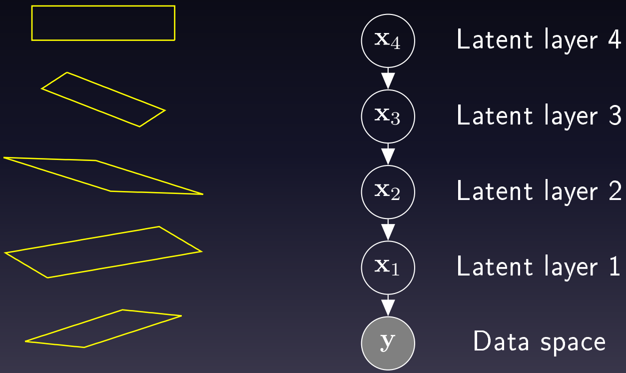
$\mathbf{x}_4$
Latent layer 4
$\mathbf{x}_3$
Latent layer 3
$\mathbf{x}_2$
Latent layer 2
$\mathbf{x}_1$
Latent layer 1
$\mathbf{y}$
Data space

Stacked PCA



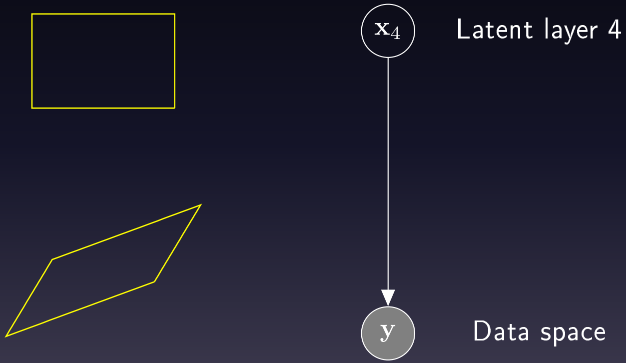


Stacked PCA

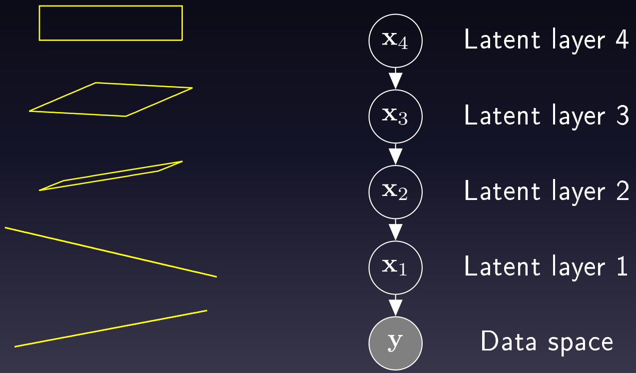


$\mathbf{x}_4$
Latent layer 4
$\mathbf{x}_3$
Latent layer 3
$\mathbf{x}_2$
Latent layer 2
$\mathbf{x}_1$
Latent layer 1
$\mathbf{y}$
Data space

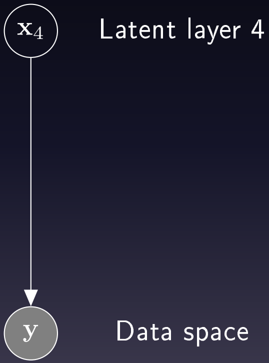
Stacked PCA



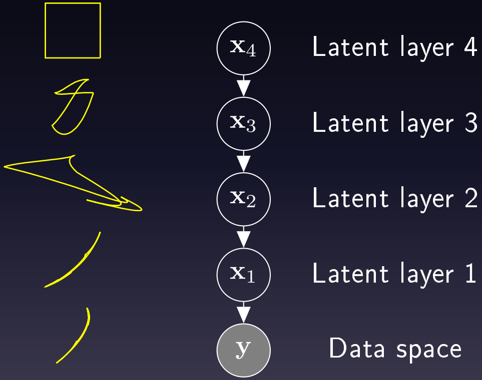
Stacked PCA


$\mathbf{x}_4$
Latent layer 4
$\mathbf{x}_3$
Latent layer 3
$\mathbf{x}_2$
Latent layer 2
$\mathbf{x}_1$
Latent layer 1
$\mathbf{y}$
Data space

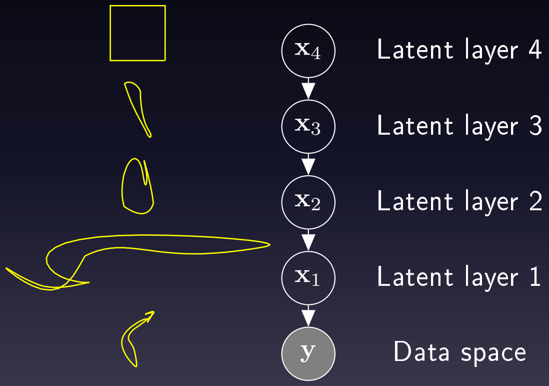
Stacked PCA



Stacked GPs

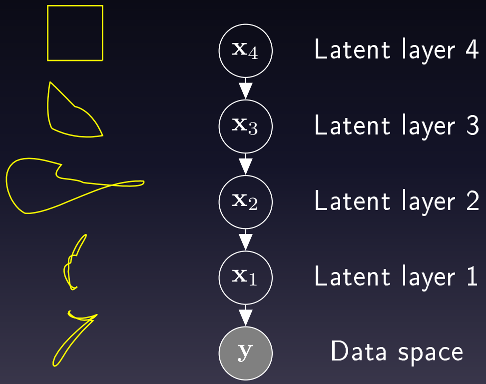

$\mathbf{x}_4$
Latent layer 4
$\mathbf{x}_3$
Latent layer 3
$\mathbf{x}_2$
Latent layer 2
$\mathbf{x}_1$
Latent layer 1
$\mathbf{y}$
Data space

Stacked GPs

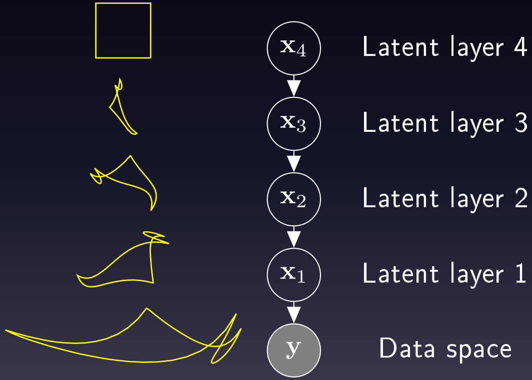


$\mathbf{x}_4$
Latent layer 4
$\mathbf{x}_3$
Latent layer 3
$\mathbf{x}_2$
Latent layer 2
$\mathbf{x}_1$
Latent layer 1
$\mathbf{y}$
Data space

Stacked GPs

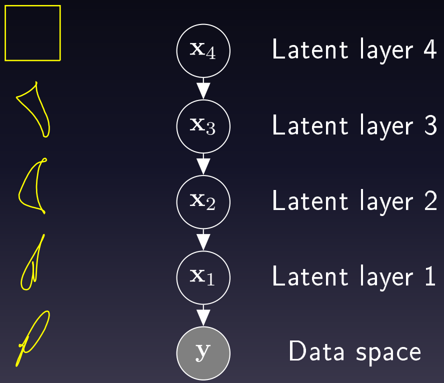

$\mathbf{x}_4$
Latent layer 4
$\mathbf{x}_3$
Latent layer 3
$\mathbf{x}_2$
Latent layer 2
$\mathbf{x}_1$
Latent layer 1
$\mathbf{y}$
Data space

Stacked GPs



$\mathbf{x}_4$ Latent layer 4
$\mathbf{x}_3$ Latent layer 3
$\mathbf{x}_2$ Latent layer 2
$\mathbf{x}_1$ Latent layer 1
$\mathbf{y}$ Data space

Stacked GPs

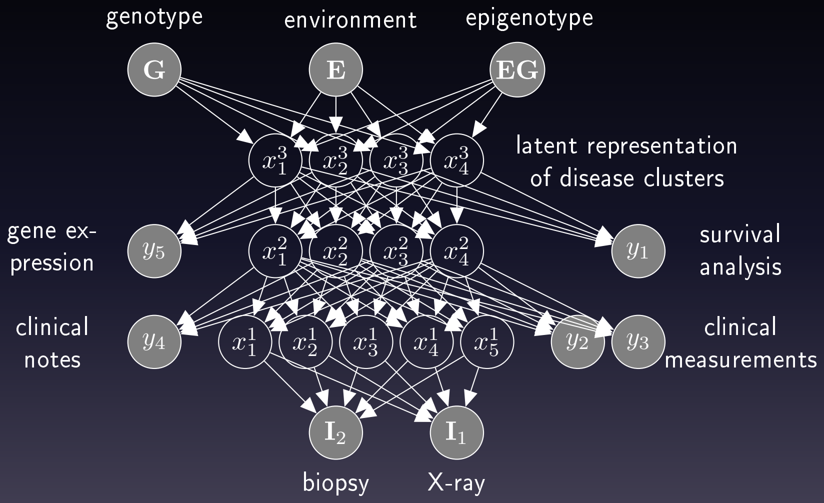


$\mathbf{x}_4$
Latent layer 4
$\mathbf{x}_3$
Latent layer 3
$\mathbf{x}_2$
Latent layer 2
$\mathbf{x}_1$
Latent layer 1
$\mathbf{y}$
Data space

What Can We Do that Google 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?
- Universities are the right place to deal with sensitive data for personalized health.
- These methodologies are part of that picture.

Deep Health

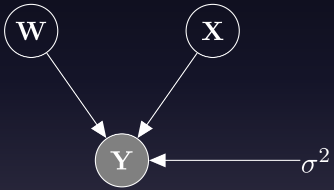

genotype
environment
epigenotype
G
E
EG
x_1^3 x_2^3 x_3^3 x_4^3
latent representation
of disease clusters
gene ex-
pression
y_5
x_1^2 x_2^2 x_3^2 x_4^2
y_1
survival
analysis
clinical
notes
y_4
x_1^1 x_2^1 x_3^1 x_4^1 x_5^1
y_2 y_3
clinical
measurements
$\mathbf{I}_2$ $\mathbf{I}_1$
biopsy
X-ray

Deep GPs

- Stacking PPCA still leads to a linear latent variable model.
- To stack latent variable models, need a non-linear model.
- The GP-LVM is a non-linear latent variable model.
- Stacking GP-LVM leads to hierarchical GP-LVM.

Bayesian GP-LVM

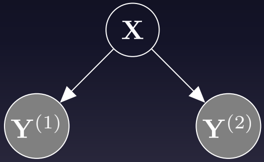
- Bayesian GP-LVM allows variational marginalization of $\mathbf{X}$ and $\mathbf{W}$.



- This leads to a Bayesian model where latent dimensionality can be learnt.

Modeling Multiple 'Views'

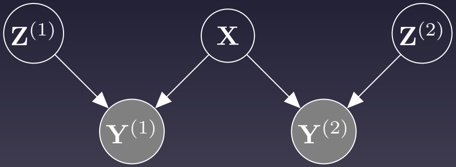
- Single space to model correlations between two different data sources, e.g., images & text, image & pose.
- Shared latent spaces: (Shon et al., 2006; Navaratnam et al., 2007; Ek et al., 2008b)



- Effective when the 'views' are correlated.
- But not all information is shared between both 'views'.
- PCA applied to concatenated data vs CCA applied to data.

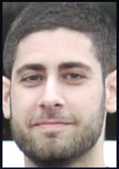
Shared-Private Factorization

- In real scenarios, the 'views' are neither fully independent, nor fully correlated.
- Shared models
  - either allow information relevant to a single view to be mixed in the shared signal,
  - or are unable to model such private information.
- Solution: Model shared and private information (Virtanen et al., 2011; Ek et al., 2008a; Leen and Fyfe, 2006; Klami and Kaski, 2007, 2008; Tucker, 1958)

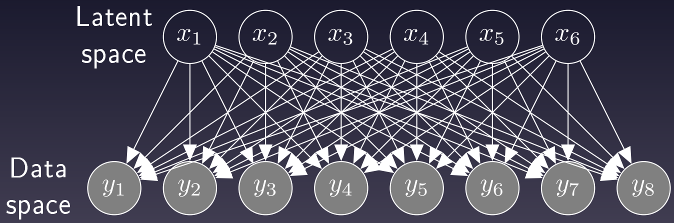


- Probabilistic CCA is case when dimensionality of $\mathbf{Z}$ matches $\mathbf{Y}^{(i)}$ (cf Inter Battery Factor Analysis (Tucker, 1958)).

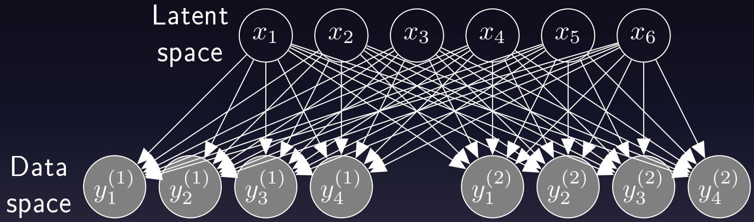
Manifold Relevance Determination



Damianou et al. (2012)



Shared GP-LVM

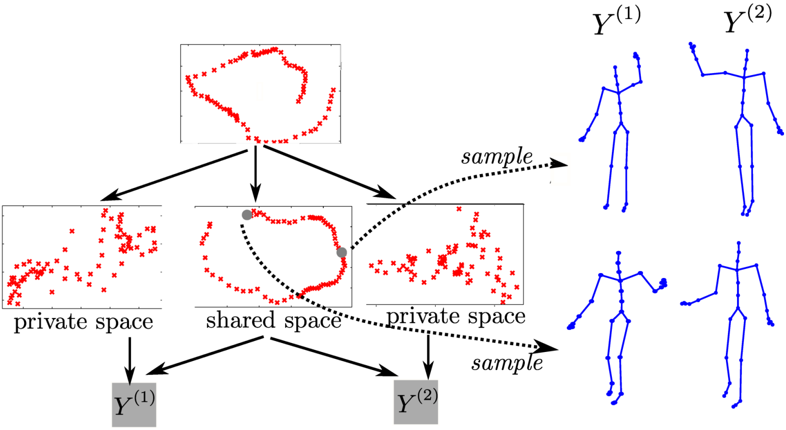


Separate ARD parameters for mappings to $\mathbf{Y}^{(1)}$ and $\mathbf{Y}^{(2)}$.

Motion Capture

- Revisit 'high five' data.
- This time allow model to learn structure, rather than imposing it.

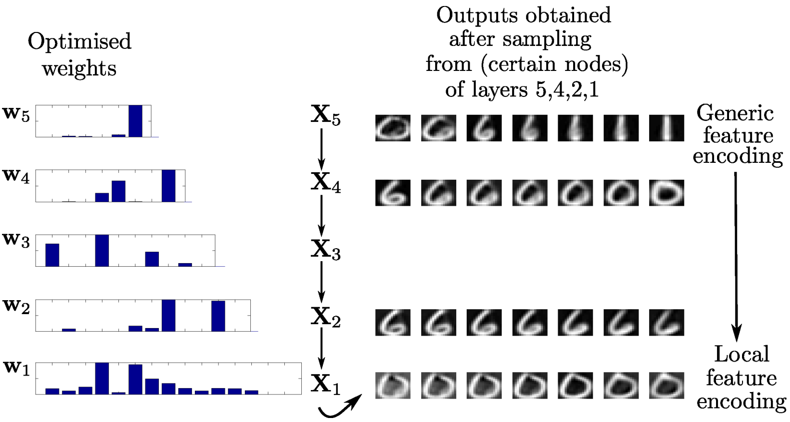
Deep hierarchies – motion capture



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



Summary

- Gaussian models good for missing data.
- Disparate data types handled with EP and Laplace.
- Deep models allow complex abstract representation of data sets at higher levels.
- Current limitation is on data set size.
- Addressing this through work by James Hensman on Stochastic Variational Inference for GPs (recent UAI paper).
- 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

Y. Bengio. Learning Deep Architectures for AI. *Found. Trends Mach. Learn.*, 2(1):1–127, Jan. 2009. ISSN 1935-8237. [DOI].

A. Damianou, C. H. Ek, M. K. Titsias, and N. D. Lawrence. Manifold relevance determination. In Langford and Pineau (2012). [PDF].

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].

C. H. Ek, J. Rihan, P. Torr, G. Rogez, and N. D. Lawrence. Ambiguity modeling in latent spaces. In A. Popescu-Belis and R. Stiefelhagen, editors, *Machine Learning for Multimodal Interaction (MLMI 2008)*, LNCS, pages 62–73. Springer-Verlag, 28–30 June 2008a. [PDF].

C. H. Ek, P. H. Torr, and N. D. Lawrence. Gaussian process latent variable models for human pose estimation. In A. Popescu-Belis, S. Renals, and H. Bourlard, editors, *Machine Learning for Multimodal Interaction (MLMI 2007)*, volume 4892 of *LNCS*, pages 132–143, Brno, Czech Republic, 2008b. Springer-Verlag. [PDF].

Z. Ghahramani, editor. *Proceedings of the International Conference in Machine Learning*, volume 24, 2007. Omnipress. [Google Books] .

G. E. Hinton and S. Osindero. A fast learning algorithm for deep belief nets. *Neural Computation*, 18: 2006, 2006.

A. Klami and S. Kaski. Local dependent components analysis. In Ghahramani (2007). [Google Books] .

A. Klami and S. Kaski. Probabilistic approach to detecting dependencies between data sets. *Neurocomputing*, 72:39–46, 2008.

J. Langford and J. Pineau, editors. *Proceedings of the International Conference in Machine Learning*, volume 29, San Francisco, CA, 2012. Morgan Kauffman.

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].

References II

G. Leen and C. Fyfe. A Gaussian process latent variable model formulation of canonical correlation analysis. Bruges (Belgium), 26-28 April 2006 2006.

D. J. C. MacKay. Introduction to Gaussian Processes. In C. M. Bishop, editor, *Neural Networks and Machine Learning*, volume 168 of *Series F: Computer and Systems Sciences*, pages 133–166. Springer-Verlag, Berlin, 1998.

R. Navaratnam, A. Fitzgibbon, and R. Cipolla. The joint manifold model for semi-supervised multi-valued regression. In *IEEE International Conference on Computer Vision (ICCV)*. IEEE Computer Society Press, 2007.

R. Salakhutdinov and I. Murray. On the quantitative analysis of deep belief networks. In S. Roweis and A. McCallum, editors, *Proceedings of the International Conference in Machine Learning*, volume 25, pages 872–879. Omnipress, 2008.

A. P. Shon, K. Grochow, A. Hertzmann, and R. P. N. Rao. Learning shared latent structure for image synthesis and robotic imitation. 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].

L. R. Tucker. An inter-battery method of factor analysis. *Psychometrika*, 23(2):111–136, 1958.

S. Virtanen, A. Klami, and S. Kaski. Bayesian CCA via group sparsity. In L. Getoor and T. Scheffer, editors, *Proceedings of the International Conference in Machine Learning*, volume 28, 2011.

A. G. Wilson, D. A. Knowles, and Z. Ghahramani. Gaussian process regression networks. In Langford and Pineau (2012).