

Big Data and Open Data Science

Neil D. Lawrence

UCLID Workshop

2nd July 2014

What is Machine Learning?

data

- ▶ **data**: observations, could be actively or passively acquired (meta-data).

What is Machine Learning?

data +

- ▶ **data**: observations, could be actively or passively acquired (meta-data).

What is Machine Learning?

data + **model**

- ▶ **data**: observations, could be actively or passively acquired (meta-data).
- ▶ **model**: assumptions, based on previous experience (other data! transfer learning etc), or beliefs about the regularities of the universe. Inductive bias.

What is Machine Learning?

data + **model** =

- ▶ **data**: observations, could be actively or passively acquired (meta-data).
- ▶ **model**: assumptions, based on previous experience (other data! transfer learning etc), or beliefs about the regularities of the universe. Inductive bias.

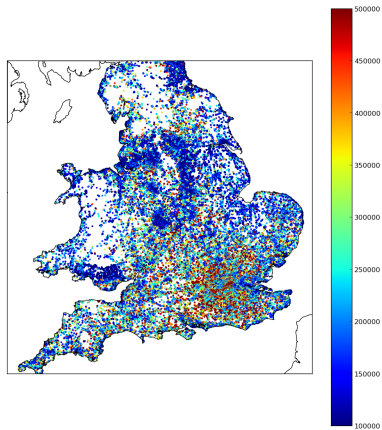
What is Machine Learning?

$$\text{data} + \text{model} = \text{prediction}$$

- ▶ **data**: observations, could be actively or passively acquired (meta-data).
- ▶ **model**: assumptions, based on previous experience (other data! transfer learning etc), or beliefs about the regularities of the universe. Inductive bias.
- ▶ **prediction**: an action to be taken or a categorization or a quality score.

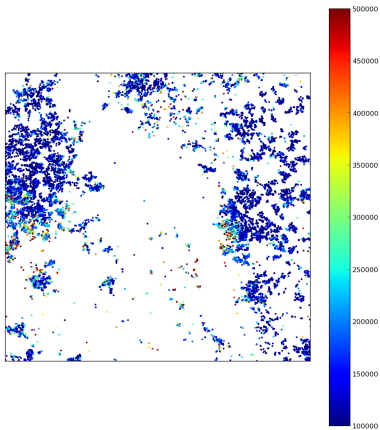
Nonparametrics for Very Large Data Sets

Modern data availability



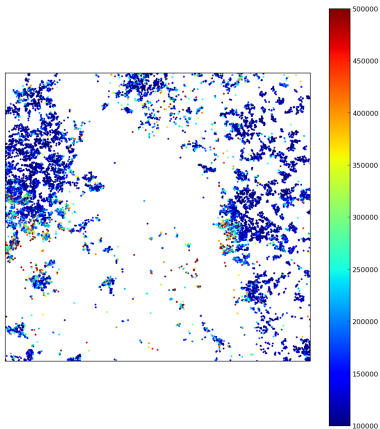
Nonparametrics for Very Large Data Sets

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, Quiñero Candela and Rasmussen, 2005, Tits-



Hensman et al. (2013)

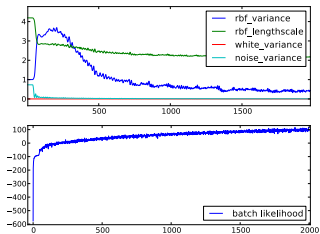


Figure 4: Convergence of the SVIGP algorithm on the two dimensional toy data

`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,

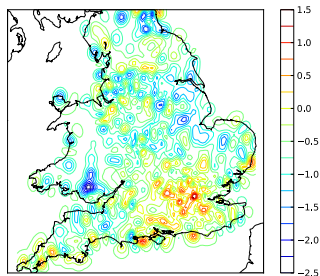


Figure 5: Variability of apartment price (logarithmically!) throughout England and Wales.

ted a GP with the same covariance function as our

What's Changed (Changing) for Medical Data?

- ▶ Try Googling for: “patient data ”...



Image from [Wikimedia Commons](#)



Image from [Wikimedia Commons](#)



INF57

A brief history *of Registration*

For more information go to: www.direct.gov.uk/motoring

A brief history of registration

The early days

Prior to the appearance of the first railways in Britain, there was a brief development and interest in steam powered road going vehicles. In 1834, a Mr Hancock started a steam coach called the “Era”, carrying up to 14 passengers from Paddington to Regents Park and the City at 6d a head. And in the following year, a Mr Church built an omnibus capable of carrying 40 passengers for the London and Birmingham Steam Carriage Company.

However, the success of the railway movement drove all such traffic off the roads.

A **Parliamentary Commission of Enquiry in 1836** reported “strongly in favour of steam carriages on roads”, but subsequent Acts of Parliament tended to have a discouraging and restrictive effect. **The Locomotive Act 1861** limited the weight of steam engines to 12 tons and imposed a speed limit of 10 mph.

The Locomotive Act 1865 set a speed limit of 4 mph in the country and 2 mph in towns. The 1865 Act also provided for the famous “man with a red flag”. Walking 60 yards ahead of each vehicle, a man with a red flag or lantern enforced a walking pace, and warned horse riders and horse drawn traffic of the approach of a self propelled machine.

The Locomotive Amendment Act 1878 made the red flag optional under local regulations, and

[Crown Copyright Reserved.]



Ministry of Transport.

THE
HIGHWAY CODE

Issued by the Minister of Transport
with the authority of Parliament in
pursuance of Section 45 of the
Road Traffic Act, 1930.

LONDON :

PRINTED AND PUBLISHED BY HIS MAJESTY'S STATIONERY OFFICE
To be purchased directly from H.M. Stationery Office at the following addresses:
Admiral House, Kingsway, London, W.C.2; 120, George St., Edinburgh;
York Street, Manchester; 1, St. Andrew's Crescent, Cardiff;
15, Donegall Square West, Belfast;
or through any bookseller.

1931.

Price 1d. net.

55-166

What are the Issues?

- ▶ Who owns our data?
- ▶ Is it 'finders keepers'?
- ▶ Does ownership proliferate?
- ▶ What does data protection offer?
- ▶ Who has the right to share our data?
- ▶ Can we withdraw this right?

Moral Panics: Perhaps Rightly

The image shows a screenshot of a web browser displaying a BBC News article. The browser's address bar shows the URL www.bbc.co.uk/news/health-27069553. The page features the BBC logo and navigation links for News, Sport, Weather, iPlayer, TV, Radio, and More. The main headline is "NHS Care.data information scheme 'mishandled'", dated 18 April 2014, by Chris Vallance. A photograph shows a hand writing on a document. The article text discusses the mishandling of patient information. A sidebar on the right lists "Top Stories" and "Features".

18 April 2014 Last updated at 17:00

NHS Care.data information scheme 'mishandled'

By Chris Vallance
PM, BBC Radio 4



The chair of the panel set up to advise the NHS and ministers on the governance of patient information has told the BBC the Care.data programme was mishandled.

Under the scheme, GP records in England will be put on a database and combined with other data to improve care.

Top Stories



- Sarkozy placed under investigation
- Met 'deleted discrimination record'
- PM warns of antibiotic resistance
- Israel vows to find teens' killers
- Facebook faces UK probe over study

Features

- It's good to talk**
The fightback against text and email 'conversations'
- Hazardous waters**
The swimming lessons blighted by sewage and corpses
- Mutant insects**
The butterflies with wings made of Sim cards
- End of the road?**
Police suspicion squeezes Syria aid convoys
- Control issues**

Related Stories

- Care.data: How did it go so wrong?
- Giant NHS database

 **Listen to the Story**  + Playlist
Morning Edition 3 min 56 sec + Download
Transcript



Cyclists look at a Ferrari parked illegally and blocking the bicycle lane off a main road in Beijing, on April 7, 2011.

Frederic J. Brown/AFP/Getty images

Share

54 Comments

Before it became China's capital in 1949, Beijing was a fairly provincial little city of 2 million people.

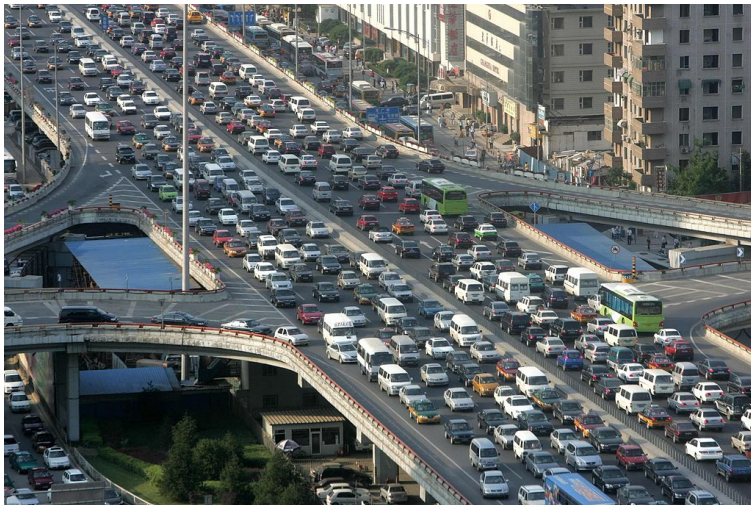


Image from [Wikimedia Commons](#)

What's Changed (Changing) for Medical Data?

- ▶ Genotyping.
- ▶ Epigenotyping.
- ▶ Transcriptome: detailed characterization of phenotype.
 - ▶ Stratification of patients.
- ▶ Massive unstructured data sources.

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.

UK Government Stipulation on Data Availability

Patients will view their NHS records online in three years

www.telegraph.co.uk/health/healthnews/9673802/Patients-will-view-their-NHS-records-online-in-three-years.html

Thursday 14 November 2013

Home News World Sport Finance Comment Culture Travel Life Women Fashion Luxury Tech Dating Offers Jobs

Women Men Motoring Health Property Gardening Food History Relationships Expat Puzzles Announcements Shop

Health News Health Advice Diet and Fitness Wellbeing Expat Health Pets Health

HOME • HEALTH • HEALTH NEWS

Patients will be given online access to their health records in the next three years under plans to be announced by the Government today.



The move for online health records comes despite the decision by Andrew Lansley, the previous health secretary, to cancel a massive NHS national database. Photo: ALAMY

By Robert Winnett, Political Editor
7:00AM GMT 13 Nov 2012

Print this article

Share 143

Facebook 76

26 Comments

Steen_Doctor_a...jpg Minority_Repor...jpg pl8vk.jpg pl8tg.jpg Show all downloads...

Search - enhanced by Google

Telegraph Politics

Bupa Care Services

Promotions

Health Most Viewed

TODAY PAST WEEK PAST MONTH

1. Millions more told to take statins
2. 111 line 'doing more bad than good'
3. Children's heart rates on the rise
4. NHS faces ruin and it will take brave decisions to save it
5. 'Only one region can cope with baby boom'

Patient Online: Roadmap

The screenshot shows a web browser window displaying the RCGP website. The address bar shows the URL: www.rcgp.org.uk/news/2013/march/patient-online-launch-with-secretary-of-state.aspx. The page features the RCGP logo and navigation menus. The main content area highlights a news article titled "Patient Online launch with Secretary of State" published on 06 March 2013. The article includes a photograph of six individuals holding copies of a document. A sidebar on the right contains a search box and a "Find courses & events" section. The browser's taskbar at the bottom shows several open files.

RCGP
Royal College of
General Practitioners

Search GO

Shop | Empty

e-Learning Revalidation ePortfolio Trainee ePortfolio

Revalidation CPD Exams Policy Clinical Membership RCGP near you


RCGP in the news
About the RCGP Press team
News archive

Home > News > Patient Online launch with Secretary of State

Patient Online launch with Secretary of State

Publication date: 06 March 2013

Patient Online: the route to electronic access



(Left-right) Dr Peter Short, GP; Chris Ghush, Head of RCGP CIRC; Dr Imran Raff, Chair of CIRC, who led the programme; Jeremy Hunt, Secretary of State for Health; Dr Clare Gerads, Chair of RCGP; and Dr Arvind Madan, GP

Patient Online: the route to electronic access

New guidance to support GP practices in providing online access for

Find courses & events

Enter keyword(s)

Topic

Region

From Date

To Date

Advanced search

Find

RSS News Feed

Steen_Doctor_a_...jpg | Minority_Repor...jpg | pl8vk.jpg | pl8tg.jpg | Show all downloads...

RCGP rules out full online access to GP records for most patients

Thursday 14 November 2013

Home | News | Views | Clinical | **Your Practice** | Commissioning | CPD on Pulse Learning | About Pulse | Jobs | Sponsored Information

Articles | News | Practice topics | Video | Comment | Resources | Dilemmas | Working Life | Special reports | Finance diary

Home > Your Practice > Practice topics > IT

RCGP rules out full online access to GP records for most patients

6 March 2013 | By Maden Davies

Print | Email | Like | Tweet | +1 | Comments (16) | Save

GP's should not be forced to give patients retrospective access to information in their medical records as it would pile work on practices and risk destabilise the doctor-patient relationship, says the RCGP in its road-map for Government plans for online records access by 2015.

The Department of Health-commissioned report says that online access to records should be 'prospective' by default and that practices should be able to set an 'access from' date for all records.

The college recommends practices assess whether access to information entered prior to this date should be allowed for patients with complex diseases and only on a 'case-by-case' basis.

The report, *Patient Online: The Road Map*, also warns of the 'unintended consequence' of an increase in queries from patients when allowing access to patient records online. It also recommends practices should be able to specify which patients were able to see their test results before the GP had reviewed them.

The GPC said the report was a blow to the Department of Health's plans to give all patients online access to their full patient record by 2015.

According to the report, while 75% of practices have the capability to provide electronic access to medical records, less than 1% of practices-63 in total-have

Mountain medicine
How to combine a love for expeditions with a career as a GP

Search Pulse
Enter your search term

Sign In | Forgotten password

Click here to manage your newsletter subscription for PulseToday and Pulse Learning

MOST POPULAR

- BMA calls for practices to work together to extend GP access in blueprint on future of general practice
- GP takes 'unlawful' decision to opt patients out of care.data programme
- Offer more same-day phone consultations for urgent patients to reduce burden on 'creaking' A&E, GPs told
- Monitor calls for greater competition between GPs and walk-in centres
- Husband appeals for safe return of missing GP

MOST COMMENTED

RELATED ARTICLES

- GP's given slice of £2.4m Government funding for online access

04 September 2013

Steen_Doctor_a_...jpg | Minority_Repor...jpg | pl8vk.jpg | pl8tg.jpg | Show all downloads...

EMIS Patient Access

The screenshot shows a web browser window with the URL <https://patient.uservice.com/knowledgebase/articles/214226-how-do-i-view-my-medical-record>. The page title is "How do I view my medical record?". The Patient.co.uk logo is at the top left. The main content area includes a breadcrumb trail "Patient Access - Medical Record", a paragraph explaining that users need to sign in to view their records, and three sections: "Appointments" (with a "Book an appointment" link), "Medical Record" (with a "View your medical record" link), and "Repeat Prescriptions" (with links for "Make a request", "See your repeat prescriptions", and "See requests detail"). A video player is partially visible at the bottom of the main content. On the right side, there is a sidebar with a search bar, a "Give feedback" link, a "Knowledge Base" section with a list of articles (including "Patient Access - Getting Started", "Patient Access - Registering", "Patient Access - Signing In and User Details", "Patient Access - Appointments", "Patient Access - Prescriptions", "Patient Access - Message Service", "Patient Access - Medical Record", "Patient Access - Reporting Issues", "Patient Access - Videos", "MyHealth", "Advertising", and "All articles"), and the Patient.co.uk logo.

How do I view my medical record?

← Patient Access - Medical Record

New and returning users may [sign in](#)

If your practice offers this service there is a link called [View your medical record](#) in the Medical Record section of the home page after you have signed in. This area of the site requires an extra sign in process so you will have to request details from your practice to gain access.

Appointments [Book an appointment](#)

Date	Time	Clinician	Location	Action
You have no appointments booked				

Medical Record

[View your medical record](#)

This link will open in another window and you will need to sign in there to view your record. Use your Access user details and security word. When finished remember to sign out and close the window.

Repeat Prescriptions [Make a request](#) [See your repeat prescriptions](#) [See requests detail](#)

Watch a video on how to view your medical record

Patient Access - Medical Record

How do I view my medical record?

Why can't I view my medical record?
Why do I have to sign in again to view my medical record?

[Give feedback](#)

Knowledge Base

- [Patient Access - Getting Started](#) 5
- [Patient Access - Registering](#) 12
- [Patient Access - Signing In and User Details](#) 13
- [Patient Access - Appointments](#) 15
- [Patient Access - Prescriptions](#) 13
- [Patient Access - Message Service](#) 4
- [Patient Access - Medical Record](#) 3**
- [Patient Access - Reporting Issues](#) 2
- [Patient Access - Videos](#) 1
- [MyHealth](#) 1
- [Advertising](#) 4
- [All articles](#)

[Patient.co.uk](#)

Steen_Doctor_a...jpg | Minority_Repor...jpg | pl8vk.jpg | pl8tg.jpg | [Show all downloads...](#)

Personal data - Providing - x

https://www.gov.uk/government/policies/providing-better-information-and-protection-for-consumers/supporting-pages/personal-data

Introducing Wa... LastPass - Dow... Getting Started My Boosters Add to Tri... Proverbi napol... IEEE Xplore - On... Other Bookmarks

GOV.UK

Search

Departments Topics Worldwide How government works Get involved
Policies Publications Consultations Statistics Announcements

GOV.UK uses cookies to make the site simpler. [Find out more about cookies](#)

Policy

Providing better information and protection for consumers

Organisation: Department for Business, Innovation & Skills
Page history: Updated 23 September 2013, see all updates
Topic: Consumer rights and issues
Minister: The Rt Hon Dr Vince Cable MP

Policy Detail Latest

Personal data

Community buying

Consumer rights bill

Misleading and aggressive selling

Implementing the Consumer Rights Directive 2011/83/EU

Consumer and competition landscape

Supporting detail:

Personal data

The midata project works with businesses to give consumers better access to the electronic personal data that companies hold about them.

It also aims to give consumers greater control of their data.

Give people greater access to electronic records of their past business and

Steen_Doctor_a...jpg Minority_Repor...jpg pl8vk.jpg pl8tg.jpg Show all downloads...

Outline

Data Heterogeneity

Deep Learning

Not the Scale it's the Diversity

The screenshot shows a web browser window with the URL `dataconomy.com/big-data-proving-to-be-a-real-challenge-for-data-scientists/`. The page features the Dataconomy logo and navigation menu (NEWS, EVENTS, OPINION, START UPS, INDUSTRY, RESOURCES, ABOUT, JOBS). The article title is "Big Data Proving to Be A Real Challenge for Data Scientists" by Furhaad Shah, dated July 2, 2014. The main image is a silhouette of a person looking at a starry sky with a circular data visualization overlay. The article text discusses the challenge of diverse data types rather than just volume, quoting Marilyn Matz, CEO of Paradigm4. A quote at the bottom reads: "The increasing variety of data sources is forcing data scientists into shortcuts that leave data and money on the table," said Marilyn Matz, CEO of Paradigm4. "The focus on the volume of data hides the real challenge of data analytics today. Only diverse types of data will we be able to unlock the enormous potential of analytics."

Category: Data Science, News, [permalink](#)

Tagged under: Big Data, Data Scientist, survey

[in](#) [twitter](#) [facebook](#)

In a recent survey conducted by [Paradigm4](#), a computational database company, it was revealed that big data was proving to be a challenge for data scientists – but not because of the amount, or volume, of data being produced, but rather the variety and diverse types of data these professionals have to handle.

"The increasing variety of data sources is forcing data scientists into shortcuts that leave data and money on the table," said Marilyn Matz, CEO of Paradigm4. "The focus on the volume of data hides the real challenge of data analytics today. Only diverse types of data will we be able to unlock the enormous potential of analytics."

[dataconomy.com/big-data-proving-to-be-a-real-challenge-for-data-scientists/](#)

Follow @DataconomyMedia

Top Stories

[Predicting the World Cup with Big Data](#)

[Kreditech Raises \\$40 Million at \\$190 Million Valuation](#)

Privacy & Cookies Policy

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,

$$p(y_1, y_2) = \int p(y_1, y_2, y_3, \dots, y_p) dy_3, \dots, dy_p$$

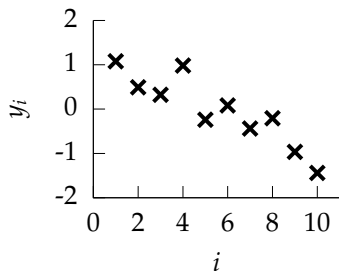
- ▶ Where y_3, \dots, y_p contains:
 1. all tests not applied to this patient
 2. all tests not yet invented!!

Magical Marginalization in Gaussians

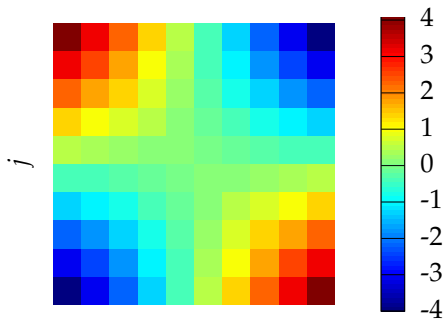
Multi-variate Gaussians

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

Gaussian Marginalization Property



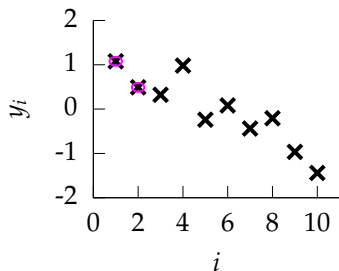
(a) A 10 dimensional sample



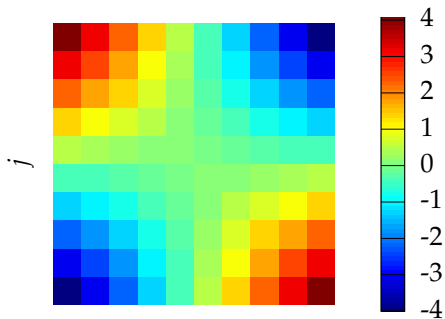
(b) colormap showing covariance between dimensions.

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

Gaussian Marginalization Property



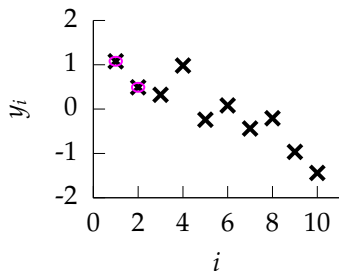
(a) A 10 dimensional sample



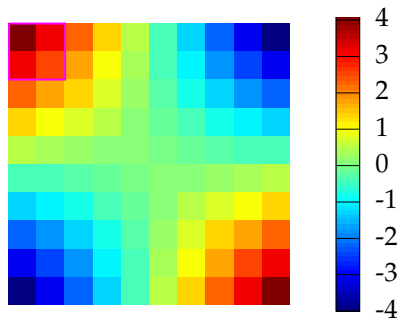
(b) colormap showing covariance between dimensions.

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

Gaussian Marginalization Property



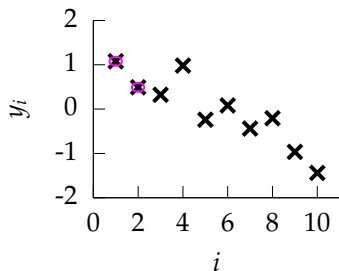
(a) A 10 dimensional sample



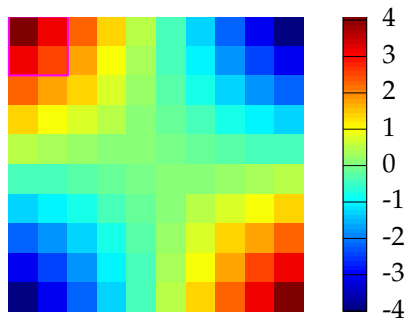
(b) colormap showing covariance between dimensions.

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

Gaussian Marginalization Property



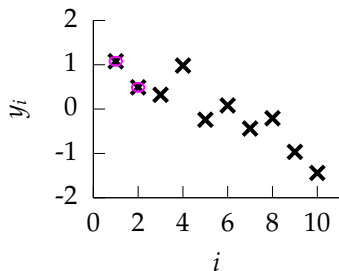
(a) A 10 dimensional sample



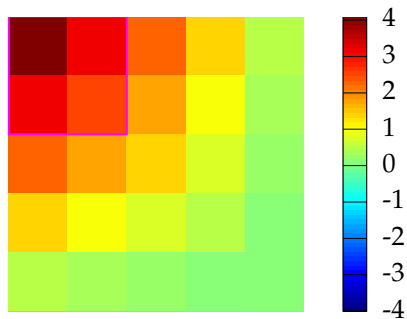
(b) colormap showing covariance between dimensions.

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

Gaussian Marginalization Property



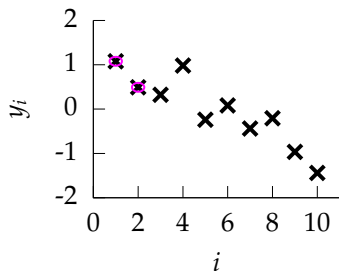
(a) A 10 dimensional sample



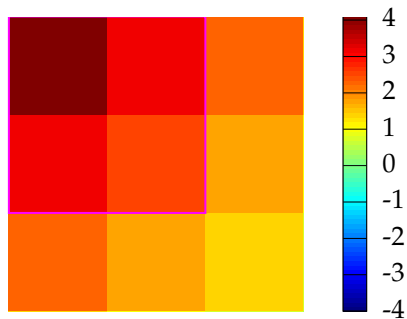
(b) colormap showing covariance between dimensions.

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

Gaussian Marginalization Property



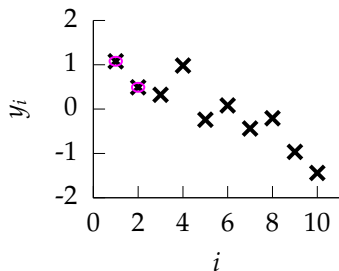
(a) A 10 dimensional sample



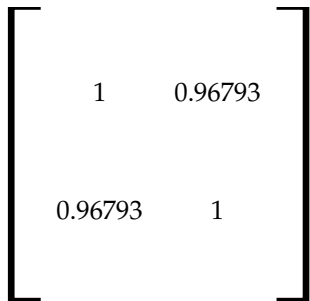
(b) colormap showing covariance between dimensions.

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

Gaussian Marginalization Property



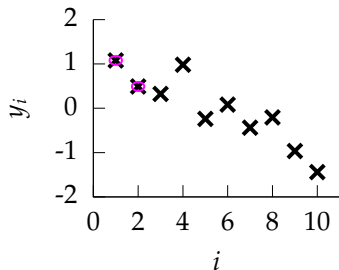
(a) A 10 dimensional sample



(b) colormap showing covariance between dimensions.

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

Gaussian Marginalization Property



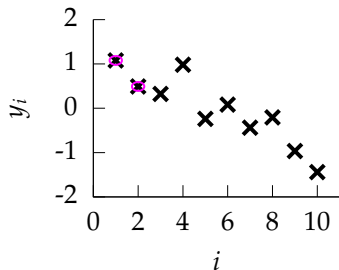
(a) A 10 dimensional sample

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

(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 recent submission with Nicolás 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

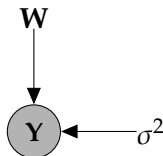
$$\mathbf{y}_{i,:} = \mathbf{W}\mathbf{x}_{i,:} + \boldsymbol{\epsilon}_{i,:},$$

where

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

Linear Latent Variable Model II

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



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

Linear Latent Variable Model II

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

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

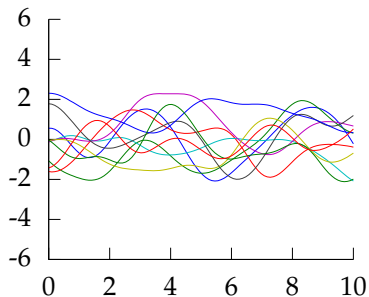
$$\log p(\mathbf{Y}|\mathbf{W}) = -\frac{n}{2} \log |\mathbf{C}| - \frac{1}{2} \text{tr}(\mathbf{C}^{-1}\mathbf{Y}^\top\mathbf{Y}) + \text{const.}$$

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

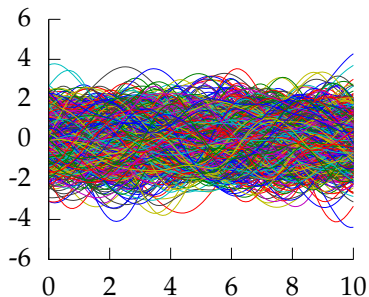
$$\mathbf{W} = \mathbf{U}_q\mathbf{L}\mathbf{R}^\top, \quad \mathbf{L} = (\Lambda_q - \sigma^2\mathbf{I})^{\frac{1}{2}}$$

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

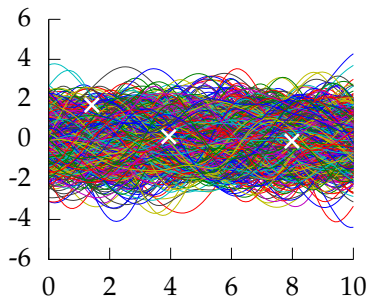
Gaussian Processes: Extremely Short Overview



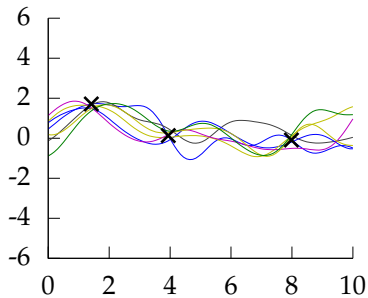
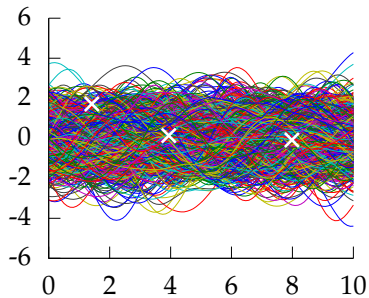
Gaussian Processes: Extremely Short Overview



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

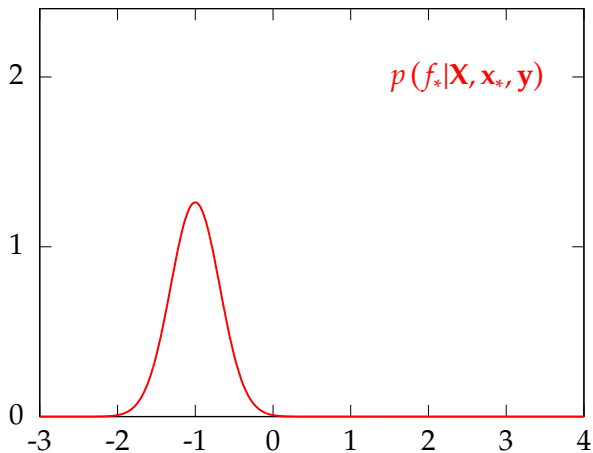


Figure : Inclusion of a data point with Gaussian noise.

Gaussian Noise

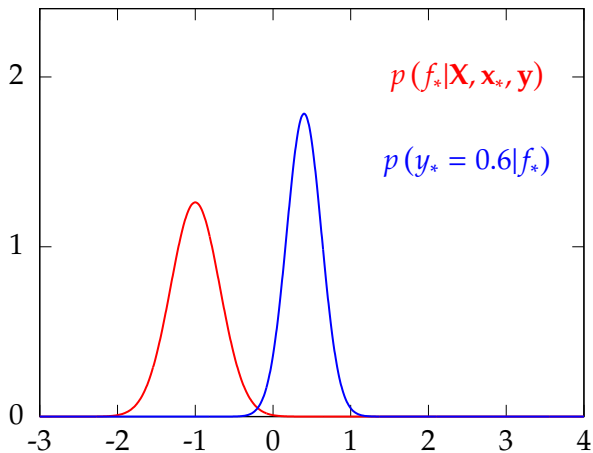


Figure : Inclusion of a data point with Gaussian noise.

Gaussian Noise

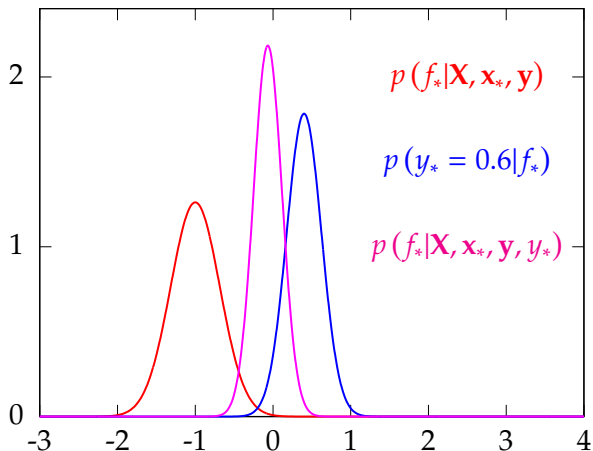


Figure : Inclusion of a data point with Gaussian noise.

Classification Noise Model

Probit Noise Model

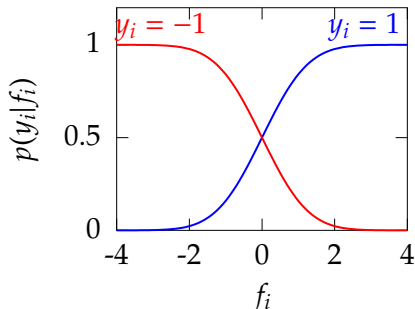


Figure : The probit model (classification). The plot shows $p(y_i|f_i)$ for different values of y_i . For $y_i = 1$ we have

$$p(y_i|f_i) = \Phi(f_i) = \int_{-\infty}^{f_i} \mathcal{N}(z|0, 1) dz.$$

Classification

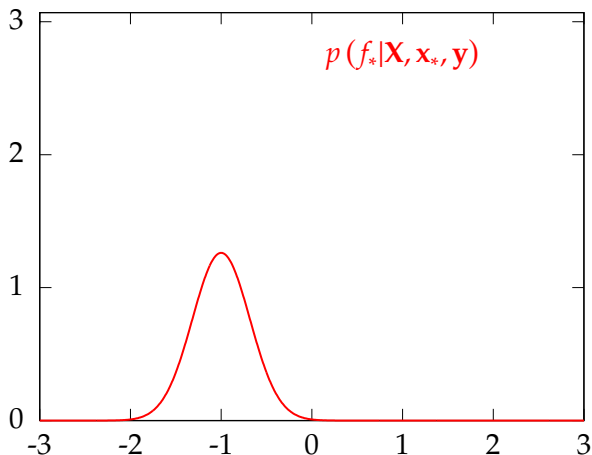


Figure : An EP style update with a classification noise model.

Classification

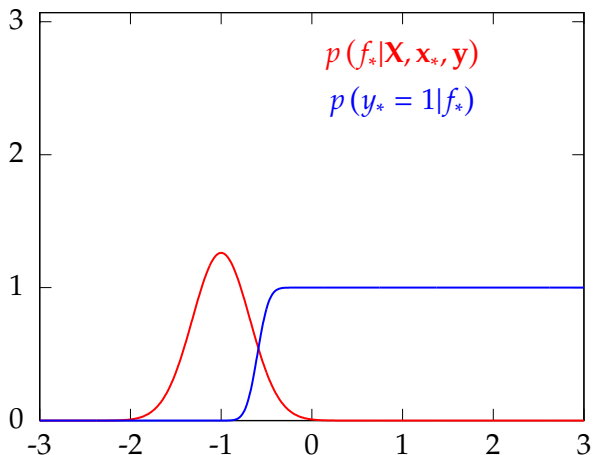


Figure : An EP style update with a classification noise model.

Classification

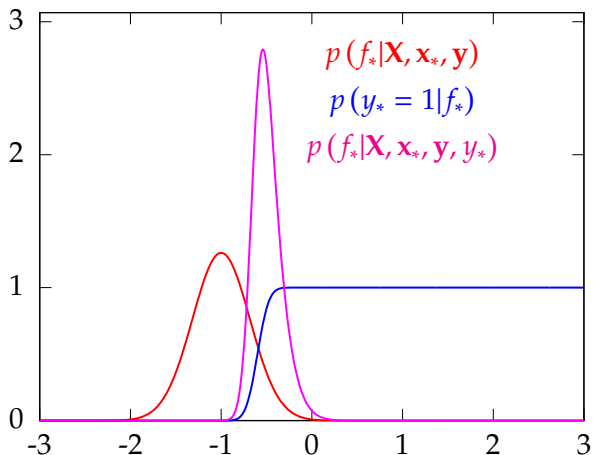


Figure : An EP style update with a classification noise model.

Classification

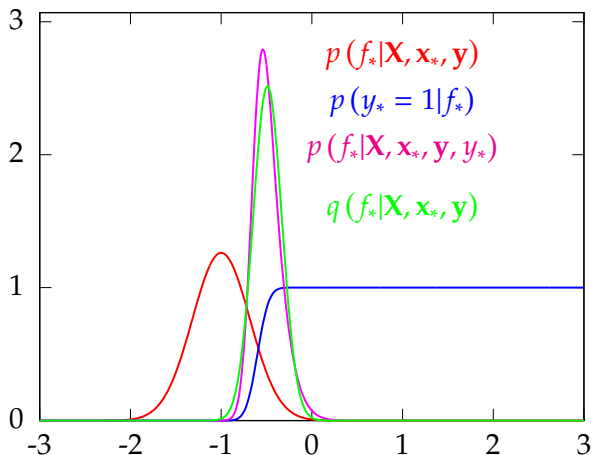


Figure : An EP style update with a classification noise model.

Ordinal Noise Model

Ordered Categories

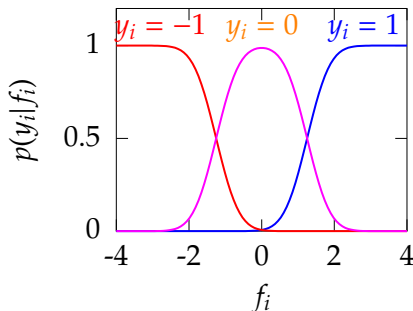


Figure : The ordered categorical noise model (ordinal regression). The plot shows $p(y_i|f_i)$ for different values of y_i . Here we have assumed three categories.

Other Challenges



- ▶ Spatial Data (workshops in November 2013 and January 2014 with 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).

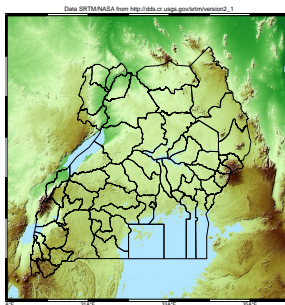
Other Data

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

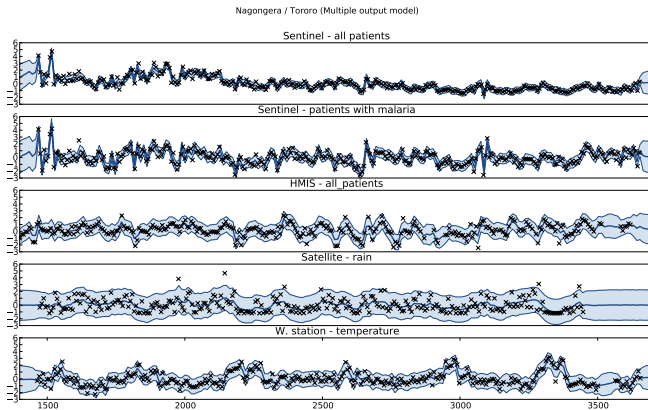
Example: Prediction of Malaria Incidence in Uganda

- ▶ Work with John Quinn and Martin Mubaganzi (Makerere University, Uganda)
- ▶ See <http://cit.mak.ac.ug/cs/aigroup/>.

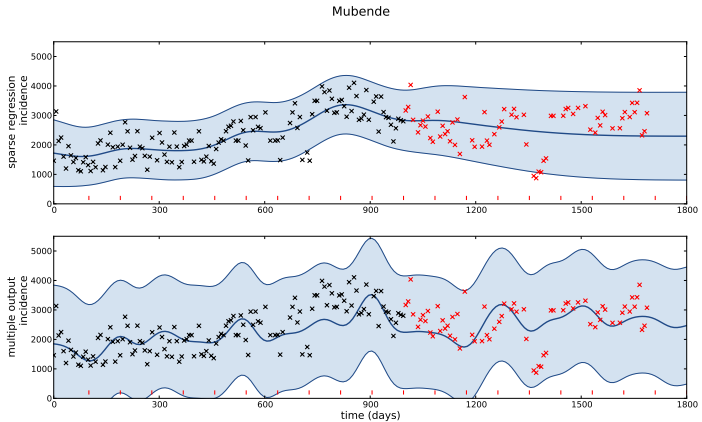
Malaria Prediction in Uganda



Malaria Prediction in Uganda



Malaria Prediction in Uganda



Visit to Uganda



Outline

Data Heterogeneity

Deep Learning

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)

Scientists See Advances in Deep Learning, a Part of Artificial Intelligence

mes.com/2012/11/24/science/scientists-see-advances-in-deep-learning-a-part-of-artificial-intelligence.html

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

HOME PAGE TODAY'S PAPER VIDEO MOST POPULAR U.S. Edition

Subscribe: Digital / Home Delivery Log In Register Now Help

The New York Times **Science** Search All NYTimes.com

WORLD U.S. N.Y. / REGION BUSINESS TECHNOLOGY SCIENCE HEALTH SPORTS OPINION ARTS STYLE TRAVEL JOBS REAL ESTATE AUSTRALIA ENVIRONMENT SPACE & COSMOS

Scientists See Promise in Deep-Learning Programs



From [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

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.

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

Log in to see what your friends are sharing on [Facebook](#) | [Log In With Facebook](#)
[nytimes.com](#) | [Privacy Policy](#) | [What's This?](#)

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

- WELL**
Lost Sleep Can Lead to Weight Gain
- THIS LIFE**
The Stories That Bind Us
- WELL**
A New Approach to Hip Surgery
- Unwanted Electronic Gear Rising in Toxic Piles
- DAVID BROOKS**
The Progressive Shift
- CONTINUING EDUCATION SPECIAL SECTION**
A Gray Jobs Market for All Ages
- Vatican's Bureaucracy Tests Even the Infallible**

FACEBOOK TWITTER GOOGLE+ SAVE E-MAIL SHARE PRINT

点击查看本文中文版

Connect With

Is "Deep Learning" a Revolution in Artificial Intelligence? | Scientists See Advances in Deep Learning | graphics - Is it possible to...

newyorker.com/online/blogs/newsdesk/2012/11/is-deep-learning-a-revolution-in-artificial-intelligence.html

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



THE NEW YORKER

SUBSCRIBE and Save up to 85%

- SUBSCRIBE
- RENEW
- ORDER A GIFT
- INTERNATIONAL ORDERS
- ONLINE ARCHIVE




SUBSCRIBE THIS WEEK'S ISSUE NEWS CULTURE POLITICS BOOKS BUSINESS CARTOONS HUMOR ARCHIVE

DOUBLE TAKE | PHOTO BOOTH | DAILY SHOUTS | PAGE-TURNER | DAILY COMMENT | AMY DAVIDSON | JOHN CASSIDY | BOROWITZ | RICHARD BRODY

THIS WEEK TOPICS | ONLINE ONLY

NEWS DESK

Reporting the latest on Washington and the world.



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

NOVEMBER 29, 2012

IS "DEEP LEARNING" A REVOLUTION IN ARTIFICIAL INTELLIGENCE?

POSTED BY GARY MARCUS

f Share 877 | Tweet 361 | +1 | COMMENTS

PRINT | MORE

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



WELCOME SIGN IN | HELP | REGISTER

Search Web site Find

MOST POPULAR | MOST E-MAILED | THIS ISSUE

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

THE NEW YORKER DIGITAL

CLICK HERE

SUBSCRIBE TODAY!

THE NEW YORKER DIGITAL

f | t | + | s | p

TABLET, MOBILE, AND MORE

Newsletter sign-up: Enter e-mail address Submit

Google To Expand Knowledge Graph Through Hire Of Geoffrey Hinton

Mar 14, 2013 • 8:23 am | (10)

by [Betsy 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.

[◀ PREV STORY](#) [NEXT STORY ▶](#)

49 10 16
[Tweet](#) [+1](#) [Like](#)
[SHRE](#) [D](#) [R](#)

SUBSCRIBE [t](#) [f](#) [g+](#) [r](#)

Enter Email Address [Subscribe Now](#)

[SUBSCRIBE OPTIONS >](#)

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 Hinton | 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

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


FOLLOW WIRED ENTERPRISE [Twitter](#) [Facebook](#) [RSS](#)


MOST RECENT WIRED POSTS

 **Jawbone's Up Fitness Band Is Now Android-Compatible**

 **Review: Ecovacs Winbot, a Window-Cleaning Robot**

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

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



Google Hires Geoffrey Hinton - Google+ | 102889418997957626067/posts

Home Search Images Maps Play YouTube News Gmail Drive Calendar More

Geoffrey Hinton

1,734 have this in circles

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.

+1 415 167

64 comments

Reza Samahri 10 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 Hartzmann Add
- Naveed Jaffry Add

23 IN COMMON WITH YOU

1,734 HAVE

Chat

facebook's 'Deep Learnin... x

www.wired.com/wiredenterprise/2013/12/facebook-yann-lecun-qa/

Introducing Wa... LastPass - Dow... Getting Started My Boosters Add to Tri... Proverbi napol... IEEE Xplore - On... Other Bookmarks

ENTERPRISE research | [intelligence](#)

Facebook's 'Deep Learning' Guru Reveals the Future of AI

BY CADE METZ 12:12:13 6:30 AM

[Follow @cademetz](#)


[Share](#) 452

[Tweet](#) 332

[+1](#) 115

[Share](#) 49

[Print](#)



Yann LeCun Photo: WIREDMAN/REUTERS

FOLLOW WIRED [Twitter](#) [Facebook](#) [RSS](#)

ENTERPRISE

MOST RECENT WIRED POSTS

[Everything That Happened in 2013 in One Image](#)

[Build a Bot](#)

[Facebook Forces Video Ads on You Because Marketers Told It To](#)

[Obama Taps Ex-Microsoft Exec to Fix Site That's Ruining His Presidency](#)

[WIRED Space Photo of the Day: Vesta's Aella](#)

[What Improves Every Party's Giant](#)

Are you an expert in machine learning? Facebook is hiring

theconversation.com/are-you-an-expert-in-machine-learning-facebook-is-hiring-21438

Introducing Wa... LastPass - Dow... Getting Started My Boosters Add to Tri... Proverbi napol... IEEE Xplore - On... Other Bookmarks

Edition: AU | UK Newsletter

Become an author Sign up as a reader Sign in

THE CONVERSATION

Academic rigour. Journalistic fear.


Business + Economy Environment + Energy Health + Medicine Politics + Society **Science + Technology**

Follow Topics Nelson Mandela Hard Evidence Digital economy DNA sequencing H2O Explainer NHS Media

12 December 2013, 2:41pm GMT


Are you an expert in machine learning? Facebook is hiring

AUTHOR

 **Neil Lawrence**
Professor of Machine Learning and Computational Biology at University of Sheffield

DISCLOSURE STATEMENT


Neil Lawrence does not work for, consult to, own shares in or receive funding from any company or organisation that would benefit from the article, and has no relevant affiliations.



Do you know anything about machine learning? [View on YouTube](#)

“Move fast and break things.” That is the Facebook motto plastered all over their California headquarters to remind engineers never to stop innovating. This week, the company

Sign in to Favourite
Post a Comment
Republish

 The University Of Sheffield

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. Tesco's 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.

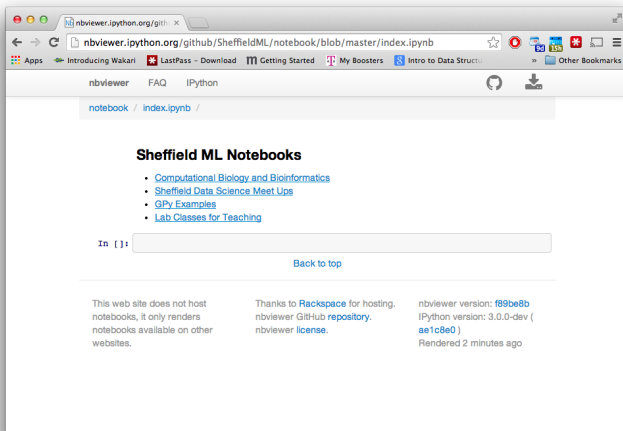
Open Data Science: A Magic Bullet?

- ▶ 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



The screenshot shows a web browser window with the URL `nbviewer.ipynb.org/gist/...` and `nbviewer.ipynb.org/github/SheffieldML/notebook/blob/master/index.ipynb`. The browser's address bar and tabs are visible. The page content includes a navigation bar with links for `nbviewer`, `FAQ`, and `IPython`. Below the navigation bar, the breadcrumb `notebook / index.ipynb /` is shown. The main heading is **Sheffield ML Notebooks**, followed by a bulleted list of links: [Computational Biology and Bioinformatics](#), [Sheffield Data Science Meet Ups](#), [GPY Examples](#), and [Lab Classes for Teaching](#). Below the list is a search bar with the text `In []:` and a `Back to top` link. At the bottom of the page, there are three columns of text: a disclaimer about rendering notebooks, a thank you message to Rackspace for hosting, and technical details about the nbviewer version (`f89be8b`), IPython version (`3.0.0-dev (ae1c8e0)`), and the rendering time (`2 minutes ago`).

nbviewer FAQ IPython

notebook / index.ipynb /

Sheffield ML Notebooks

- [Computational Biology and Bioinformatics](#)
- [Sheffield Data Science Meet Ups](#)
- [GPY Examples](#)
- [Lab Classes for Teaching](#)

In []:

[Back to top](#)

This web site does not host notebooks, it only renders notebooks available on other websites.

Thanks to [Rackspace](#) for hosting. nbviewer GitHub [repository](#). nbviewer [license](#).

nbviewer version: `f89be8b`
IPython version: 3.0.0-dev (`ae1c8e0`)
Rendered 2 minutes ago

Gaussian Processes Summer School

ml.dcs.shef.ac.uk/gpss/past_meetings.html

GAUSSIAN PROCESS SUMMER SCHOOL

IUDICIUM POSTERIUM DISCIPULUS EST PRIORIS

Philosophy | Target | History | News

You are here: Machine Learning Group / Gaussian Process Summer School

- Home
- Accommodation
- Getting There
- Registration
- History
- Past Meetings
- Facebook Page

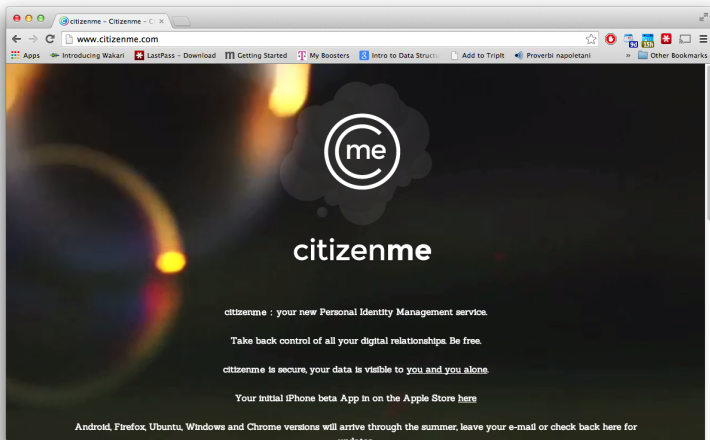
Past Meetings

- [Gaussian Process Road Show, Pereira, Colombia, February 2014](#)
- [Workshop on Spatiotemporal Modelling with Gaussian Processes, Sheffield, January, 2014](#)
- [Gaussian Process Winter School, Sheffield, January, 2014](#)
- [Gaussian Process Road Show, Kampala, Uganda, August 2013](#)
- [Latent Force Model Workshop, Sheffield, June, 2013](#)
- [Gaussian Process Summer School, Sheffield, June, 2013](#)
- [Gaussian Processes in Practice, 2009](#)
- [Gaussian Process Round Table, 2005](#)

This document last modified Tuesday, 18-Mar-2014 06:17:51 GMT

But we need to do much more! (Dan Cornford's User Groups)

Digital Identity and Data Ownership



The image shows a browser window displaying the website www.citizenme.com. The browser's address bar shows the URL, and the page content features a dark background with a bokeh light effect on the left. The logo, consisting of the letters 'me' inside a circle, is enclosed in a thought bubble shape. Below the logo, the text 'citizenme' is displayed in a large, white, sans-serif font. Underneath, there are several lines of text: 'citizenme : your new Personal Identity Management service.', 'Take back control of all your digital relationships. Be free.', 'citizenme is secure, your data is visible to you and you alone.', and 'Your initial iPhone beta App in on the [Apple Store here](#)'. At the bottom, a line of text states: 'Android, Firefox, Ubuntu, Windows and Chrome versions will arrive through the summer, leave your e-mail or check back here for updates.'

citizenme - Citizenme - C X

www.citizenme.com

Apps Introducing Wakari LastPass - Download Getting Started My Boosters Intro to Data Struct Add to TripIt Proverbi napoletani Other Bookmarks

me

citizenme

citizenme : your new Personal Identity Management service.

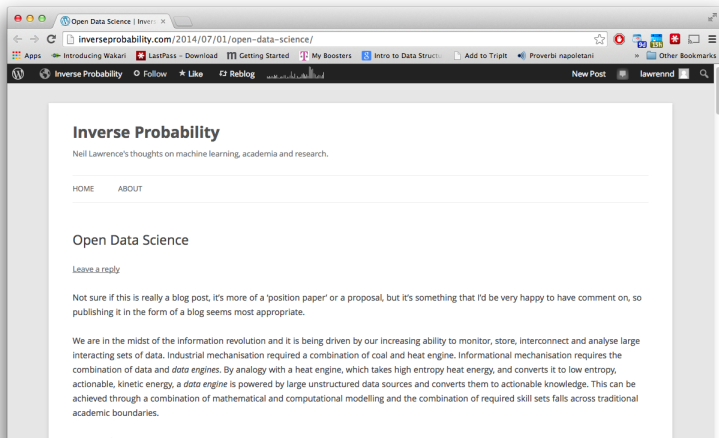
Take back control of all your digital relationships. Be free.

citizenme is secure, your data is visible to you and you alone.

Your initial iPhone beta App in on the [Apple Store here](#)

Android, Firefox, Ubuntu, Windows and Chrome versions will arrive through the summer, leave your e-mail or check back here for updates.

Blog Post



The screenshot shows a web browser window with the address bar displaying `inverseprobability.com/2014/07/01/open-data-science/`. The page content includes:

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.

Summary

- ▶ The Challenges of Modern Big Data are Radically Different
- ▶ statistics + computer science = data science
- ▶ Need to change the way in which we do science.
- ▶ Major methodological difficulties, computational difficulties and accessibility difficulties.
- ▶ Open Data Science provides and Answer.

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 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. [\[PDF\]](#).
- 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. [\[PDF\]](#).
- G. E. Hinton and S. Osindero. A fast learning algorithm for deep belief nets. *Neural Computation*, 18:2006, 2006.
- P. S. Laplace. *Essai philosophique sur les probabilités*. Courcier, Paris, 2nd edition, 1814. Sixth edition of 1840 translated and reprinted (1951) as *A Philosophical Essay on Probabilities*, New York: Dover; fifth edition of 1825 reprinted 1986 with notes by Bernard Bru, Paris: Christian Bourgeois Éditeur, translated by Andrew Dale (1995) as *Philosophical Essay on Probabilities*, New York:Springer-Verlag.
- 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. [\[PDF\]](#).
- 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.
- S. T. Roweis. EM algorithms for PCA and SPCA. In M. I. Jordan, M. J. Kearns, and S. A. Solla, editors, *Advances in Neural Information Processing Systems*, volume 10, pages 626–632, Cambridge, MA, 1998. MIT Press.
- 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.
- M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society, B*, 6(3):611–622, 1999. [\[PDF\]](#). [\[DOI\]](#).
- A. G. Wilson, D. A. Knowles, and Z. Ghahramani. Gaussian process regression networks. In J. Langford and J. Pineau, editors, *Proceedings of the International Conference in Machine Learning*, volume 29, San Francisco, CA, 2012. Morgan Kauffman.