

---

# Research Journal

---

Jerome Wynne

University of Bristol

19th June 2017 - Present



# Contents

<b>Thursday, 1st June 2017</b>	<b>1</b>
1 Summary . . . . .	1
2 Placement objectives . . . . .	1
3 Reading a paper critically . . . . .	1
4 Structuring your time . . . . .	2
5 Figuring out what to write about . . . . .	2
6 Structuring your work . . . . .	3
7 How to use this book . . . . .	3
<b>Monday, 19th June 2017</b>	<b>5</b>
1 Summary . . . . .	5
2 Prince: <i>Computer Vision</i> (Ch. 1, Introduction) . . . . .	6
3 Prince: <i>Computer Vision</i> (Ch. 2, Introduction to Probability) . . . . .	7
<b>Tuesday, 20th June 2017</b>	<b>8</b>
1 Summary . . . . .	8
2 Prince: <i>Computer Vision</i> (Ch. 3, Common Probability Distributions) . . . . .	8
3 Prince: <i>Computer Vision</i> (Ch. 4, Fitting Probability Models) . . . . .	10
<b>Wednesday, 21st June 2017</b>	<b>11</b>
1 Summary . . . . .	11
2 PV Panels . . . . .	12
3 PV Panels: Crack detection . . . . .	12
<b>Thursday, 22nd June 2017</b>	<b>14</b>
1 Summary . . . . .	14
2 PV Panels: Crack Detection . . . . .	14
3 Prince: <i>Computer Vision</i> (Ch. 5/6) . . . . .	15
4 Prince, <i>Computer Vision</i> (Ch. 7, Modeling complex data densities) . . . . .	17
<b>Friday, 23rd June 2017</b>	<b>19</b>
1 Prince, <i>Computer Vision</i> (Ch.7 Modeling Complex Data Densities) . . . . .	20
2 PV Panels: Crack detection . . . . .	21
<b>Saturday, 24th June 2017</b>	<b>23</b>
1 Fulbright Scholarship . . . . .	23
2 VEST Scholarship . . . . .	26

## Contents

<b>Sunday, 25th June 2017</b>	<b>27</b>
1 Prince, <i>Computer Vision</i> (Ch. 7, Modeling Complex Data Densities)	29
2 Prince, <i>Computer Vision</i> (Ch. 8, Regression models)	30
<b>Monday, 26th June 2017</b>	<b>31</b>
1 Summary	31
2 Prince, <i>Computer Vision</i> (Ch. 8, Regression models)	32
<b>Tuesday, 27th June 2017</b>	<b>35</b>
1 Summary	35
2 Boyd, <i>Convex Optimisation</i> (Lecture 2)	38
<b>Thursday, 29th June 2016</b>	<b>40</b>
<b>Friday, 30th June 2016</b>	<b>41</b>
1 Meeting with MM	41

# Thursday, 1st June 2017

## 1 Summary

- Set up this labbook.
- Wrote down what I'd like to achieve during my placement.
- Listed a bunch of classic *Computer Vision* papers.

## 2 Placement objectives

During my 8-week placement I would like to:

- Apply and understand the most useful *Computer Vision* algorithms.
- Write either a statistics, machine learning, or *Computer Vision* research paper worthy of publication.
- Develop a systematic workflow to problems involving data, and demonstrate this workflow in publicly available scripts or notebooks.

## 3 Reading a paper critically

- Motivations for the problem posed
- The choices made in finding a solution
- The assumptions behind the solution
- The validity of those assumptions
- Whether assumptions can be removed without invalidating the approach
- What was accomplished
- Future research directions

Thursday, 1st June 2017

## 4 Structuring your time

I think it's realistic to aim for 12 hours of productive work each day. This includes reading, studying, and running my own experiments. To begin with, I think I should spend at least three-quarters of this time learning. Maybe by week four I will shift the ratio.

I will see how it goes, but I suspect I will prefer to work from the University rather than DNV-GL's offices.

To make this placement a success, I will need to be disciplined about how I use my time. I know that if I get up early and immediately go to work, I can easily crack out four hours without breaking a sweat. After this time I can take at least a couple of hours - go to the gym, read, walk, or - of course - eat. After that I'll get back on that horse for another four or five hours, before taking another short break, then do a few more hours.

Having lots of sleep is important for my mental health and emotional wellbeing, so I should aim to get at least eight hours. If I get up at six, this means that I should go to bed at half-nine. I can work from my flat, University buildings, or the DNV-GL offices. It would be a good idea to mix the places I study up - I know this has helped me keep focused in the past. I need to be careful to provide some time for myself too, so that I can recuperate: I think two hours at the end of the day, eight until ten, will be enough.

## 5 Figuring out what to write about

1. Narrow down your field of study
2. Define what to investigate
3. Establish a thesis or an argument

I am studying *Computer Vision* and statistics. This is because I want to build robust algorithms to understand visual information, which in turns makes it easier to automate difficult or tedious tasks, such as watching CCTV cameras or spotting damage to structures in video footage. I am doing this so that we will know more about how patterns in visual information can be found, and so that we can exploit these patterns to automate economically and socially beneficial tasks. I am also interested in how we can represent visual information to make it easier to work with and to understand.

## 6 Structuring your work

1. Find data that poses an unsolved problem, or find a problem that needs data to be solved.
2. Find the data, or pose the problem.
3. Review the relevant literature.
4. Propose a solution, then run experiments and conduct analysis to test that solution.
5. Write up the results.

## 7 How to use this book

I should use this book to record what I'm doing, ideas and conversations I have, experiments I run, papers and books to read, things I understand and don't understand - everything related to my research. It only takes a few minutes to put a screenshot in this document - remember this!

As a habit, at the beginning of each day I will write down what I plan to do. At the end of the day I'll review what I've done. I should also review the book each week, to get an idea what I've been up to and where I'm headed.

# Papers

Title	Authors	Year	Topic	Read
Theory of Edge Detection	D. Marr, E. Hildreth	1980	Computer vision	No
A Computational Approach to Edge Detection	J. Canny	1986	<i>Computer Vision</i>	No
Determining optical flow	B.K.P. Horn, B. Schunk	1981	Computer vision	No
An iterative image registration technique with an application to stereo vision	B. Lucas, T. Kanade	1981	Computer vision	No
Snakes: Active contour models	M. Kass, A. Witkin, D. Terzopoulos	1988	<i>Computer Vision</i>	No
Eigenfaces for recognition	M. Turk, A. Pentland	1991	<i>Computer Vision</i>	No
Shape and motion from image streams under orthography: a factorization method	C. Tomasi, T. Kanade	1992	<i>Computer Vision</i>	No
Texture features for browsing and retrieval of image data	B. Manjunath, W. Ma	1996	<i>Computer Vision</i>	No
Conditional density propagation for visual tracking	M. Isard, A. Blake	1998	<i>Computer Vision</i>	No
Normalized cuts	J. Shi, J. Malik	2000	<i>Computer Vision</i>	No
Non-parametric model for background subtraction	A. Elgammal, D. Harwood, L. Davis	2000	<i>Computer Vision</i>	No
Distinctive image features from scale-invariant keypoints	D. Lowe	2004	<i>Computer Vision</i>	No



# Monday, 19th June 2017

## 1 Summary

- Arrived at DNV offices @ 10:00.
- H & S induction. Remember to look at the green lights above the doors!
- Met Elizabeth Traiger:
  - Solar panel cracks toy problem.
  - DNV are keen to use Python.
  - Approx. 1500 minutes of turbine data available.
  - We need a method for labeling the data.
- IT was not set up - she will email me tomorrow when it is ready.
- DNV use Microsoft Azure services internally.
- The company is open to using TensorFlow.
- We agreed that weekly meetings would be appropriate. She is away this Friday (in London, speaking about using satellite imagery of windfarms to monitor fish populations (?), in association with the European Space Agency) and next week (site visit?).
- Various ideas came up in our conversation: super-resolution methods, using simulation data, creating synthetic data, thinking about how to segment the images.
- Footage from two turbines is available this week. More footage will be available next week.
- I can VPN into DNV's network using the laptop that they provided me.
- I left the DNV offices after lunch (14:00), then worked on the Design Project briefs until 16:30.
- Liz planned to book a meeting with the inspections team to discuss how they look for damage. Possible questions:
  - What does damage look like?
  - Over what time periods does it develop?

Monday, 19th June 2017

- What causes the damage?
- What mistakes do you make when looking for damage?
- What type of damage is easiest/most difficult to spot?
- What do you do when you find damage?
- What problems do you think there are with using drones to identify damage?

In the afternoon I read the introduction to Canny's paper on edge detection. I learnt that there is a tension between edge detection and localization (the distance between the detected edge and the true edge). Apparently the paper determines the optimal operator for various common edges.

I also read the introduction and first chapter of S.J. Prince's book on computer vision. The introduction outlined the book's structure and the first chapter was a quick review of basic probability theory.

- Read Canny paper.
- Read the Chapter 1, 3, and 6 of S.J. Computer Vision. Make notes and complete the associated exercises.
- Apply what was learnt in this chapter using Python.
- Outline the structure of compute vision / machine learning.

## 2 Prince: Computer Vision (Ch. 1, Introduction)

The book's structure was outlined. It consists of six sections:

1. Probability.
2. Machine learning for machine vision.
3. Graphical models (principled ways to simplify the relationship between image data and estimated properties) for machine vision.
4. Image preprocessing.
5. Geometric machine vision - generating 3D models based on image data and localizing a camera based on its view.
6. Families of machine vision models for common problems.

It took me about ten minutes to read through this first chapter of five pages. Relevant papers are listed at the end of each of the book's sections.

### 3 Prince: Computer Vision (Ch. 2, Introduction to Probability)

The second chapter took about an hour to read and make notes for. Almost all of it was review material: random variables, distributions, joint distributions, Bayes' rule, conditionality, and expectation. I did learn, however, two useful things:

- A Hinton diagram can be used to visualize probability mass functions. A square is associated with each point; the square's area relative to the total area of all squares corresponds to that value's relative probability. Figure 1 shows a Hinton diagram in two variables.

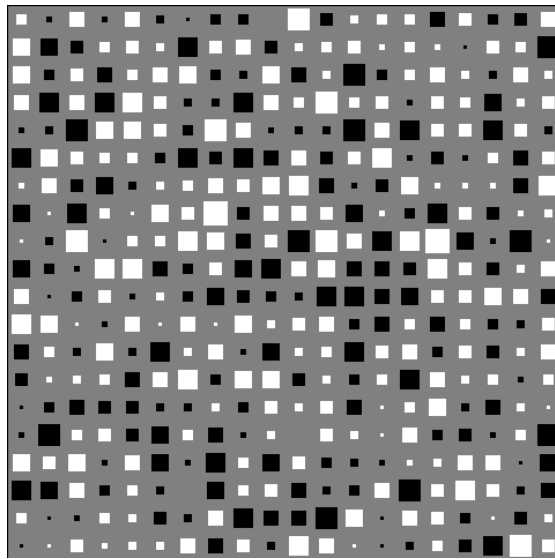


Figure 1: A Hinton diagram can be used to visualize probability mass functions in one or two variables.

- A useful identity: if two random variables are independent, then the product of their expectation is equivalent to the expectation of their product. More generally:

$$E[f(x) \cdot g(y)] = E[f(x)] \cdot E[g(y)] \quad (0.1)$$

# Tuesday, 20th June 2017

## 1 Summary

I woke at six (I'm getting the hang of it). For the first hour of the day I reviewed the project brief MZ had commented on and sent it off to PH. PH mentioned I could use his office for the next two weeks while he was away - hopefully I can pick up the key from him this afternoon.

Today I would like to complete Chapters 3 and 4 of Prince's *Computer Vision*, do some modeling in Python, and begin the CVX course. I should also get an overview of the main *Computer Vision* tools for Python.

I made notes on Chapter 3 and completed five of its exercises, which related to conjugacy, the exponential family, and the modes of the distributions studied. PH said to pick up the key to his office from the porter. ET emailed me to say that IT have set up my account - I agreed to go in tomorrow morning to meet her and pick up my laptop. I should also ask her when we plan to meet the wind turbine inspections team.

I intend to complete Chapter 4 this afternoon. I am now going to read Canny's paper. I tried to read Canny's paper but kept zoning out. I did understand, however, that he formulated an optimization problem by expressing detection performance as signal-to-noise ratio and localization performance as a similar quantity but involving gradients.

I spent fifty minutes reading through Chapter 4. I made notes on maximum likelihood, maximum a posteriori, and bayesian methods for optimizing parameter estimates. I read through an example wherein the mean and variance of a univariate normal distribution were estimated using a normal-scaled inverse gamma prior; as it turns out, the benefits of using the conjugate here are twofold - not only does the posterior have a closed-form solution, but so also does the posterior predictive!

## 2 Prince: Computer Vision (Ch. 3, Common Probability Distributions)

I started on Chapter 3 of Prince's book yesterday. It covers eight probability distributions that will be useful for machine vision purposes. Four of these are for modelling image data or world properties; the other four are for modelling the parameters of these four.

I made notes, but have yet to complete the exercises. The distribution pairs covered were:

- Bernoulli - Beta
- Categorical - Dirichlet
- Univariate normal - Normal-scaled inverse gamma
- Multivariate normal - Normal inverse Wishart

The four distributions on the left were familiar to me, as was the beta distribution (however it was useful to note that the expectation of  $\text{Beta}[\alpha, \beta]$  is defined by  $\frac{\alpha}{\alpha+\beta}$ ). The Dirichlet, normal-inverse scaled gamma, and normal inverse wishart distributions were new to me.

The Dirichlet distribution is effectively a generalization of the beta distribution to cover more than one variable. It's used to model the  $K$  parameters of a categorical distribution. As such,  $\sum_k \lambda_k = 1$  at all points in the Dirichlet distribution. Its hyperparameters  $\alpha_1, \dots, \alpha_K$  describe the categorical distribution's expectations over each of its variables  $E[\lambda_1], \dots, E[\lambda_K]$ ; the magnitude of the  $\alpha$  values determines the Dirichlet's dispersion.

The Normal-scaled inverse gamma distribution is used to model uncertainty in the mean and variance of a univariate normal distribution. It has four parameters. In a similar vein, the normal inverse Wishart distribution describes uncertainty in the mean vector and covariance matrix of a multivariate normal distribution.

I'm now going to complete three questions from the exercises, then go to the barbers. When I get back I will complete three more exercises, before continuing to read the Canny paper.

I completed three questions. The first asked for the mode of the beta distribution (found by maximizing the pdf w.r.t. the variable). The second pointed out that all of the distributions in the chapter were members of the exponential family; it then requested the beta pdf in exponential family form. The third related a normal prior and restricted normal likelihood to a normal posterior (i.e. it was a conjugacy problem). These questions together took me about forty-five minutes to complete.

This is not the first time the exponential distribution has been mentioned - why is understanding multiple distributions as variants on one distribution useful?

I completed two more questions: showing that the normal-scaled inverse gamma is the conjugate prior to the univariate normal distribution; showing that the Dirichlet distribution is the conjugate prior to the categorical distribution. They were mostly exercises in algebra, but it did drive home why the Dirichlet and normal-scaled inverse gamma distributions are relevant, and what they look like (particularly w.r.t. the Dirichlet-beta similarities).

Tuesday, 20th June 2017

### 3 Prince: Computer Vision (Ch. 4, Fitting Probability Models)

The chapter began with maximum likelihood, maximum a posteriori, and bayesian methods for optimizing parameter estimates. I read through an example in which the mean and variance of a univariate normal distribution were estimated using a normal-scaled inverse gamma prior; as it turns out, the benefits of using the conjugate here are twofold - not only does the posterior have a closed-form solution, but so also does the posterior predictive! I think I would benefit from writing out the Bayesian solution, and possibly running a simulation in Python. Maybe I'll go through the other example before building the simulation.

I made notes on both worked examples. The second example involved inferring posterior and posterior predictive distributions from a categorical likelihood and its conjugate (a Dirichlet prior). As in the case of the univariate normal - normal-scaled inverse gamma problem, both distributions had neat solutions. The posterior was another Dirichlet distribution (with correspondingly updated parameters  $\tilde{\alpha}_j = \alpha_j + N_j$ ) and the posterior predictive turned out to be a function of beta functions.

I'm going to write two R simulations, one for each example.

I built an R simulation for the first example.

I attempted Q4.6, but fell short (the algebra became untangled). I will try again tomorrow morning; perhaps doing Q4.5 first will make it slightly easier. Once I've done those two questions, I should like to do Q4.7-4.10. Worked 4 - 7:30 on Com. Vis. Chapter 4.

This evening I watched the first lecture from the CVX course. It explained what convex optimisation is, common variants of it (least squares and linear programming), discussed application areas, defined a convex function, provided an overview of the course, and gave some history on the topic.

# Wednesday, 21st June 2017

## 1 Summary

Today I plan to:

- Complete the exercises from Ch. 4 of Computer Vision
- Walk through one of the examples available at PyImageSearch
- Read Ch. 5 of *Computer Vision* and complete the associated exercises
- Pick up Paul Harper's office key
- Meet ET at 10:00 to have IT induction
- Explore solar panel dataset
- Investigate Python tools for tagging images
- Read Chapter 1 of Boyd's Convex Optimisation

(07:00 - 09:50) I spent the first hour and forty-five minutes of the day completing Exercises 4.5 - 4.10 of Computer Vision. I then made notes on Chapter 5, which is focused on the normal distribution. It began by showing the difference between spherical, diagonal, and full covariance matrices - spherical matrices have circular isodensity contours, diagonals have ellipsoidal ones whose axes are aligned with coordinate axes, and full covariance matrices have ellipsoidal contours in any orientation.

As it happens, it's possible to rotate a distribution's coordinate system to permit a full covariance matrix to be reexpressed as a diagonal one. The rotation matrix is deduced using the singular value decomposition (something I should probably know, but don't). I'm about to head off to DNV, but I intend to watch Gilbert Strang's lecture on the singular value decomposition when I get back. I think I could also do with inspecting matrix inverses more closely.

I spent the afternoon (10:30 - 16:00) in the DNV offices getting set-up with IT and so on. I also went through the two datasets made available to me so far. In the evening I wrote a Python script to resize and subset the PV panel images. I also skimmed Ch. 6 of Computer Vision.

Tomorrow morning I should label a few hundred of these images and try out the algorithms described in Ch. 6 of Com. Vis. For the first three hours of the day I should

Wednesday, 21st June 2017

like to study from the book. The middle hours can be spent labelling and coding, before returning again to studying. I will read through Ch. 1 of S. Boyd's book on convex optimization also.

I did find a tool for annotating images - it's called Fiji (<https://fiji.sc/>). I should really start time-stamping my entries on this...

I also read a paper [*Detection and Classification of Surface Defects on Gun Barrels Using Computer Vision and Machine Learning*, Sanmuhamani et al.] on gun barrel damage classification. They had success using an extended maxima transform was used to segment their images. They used sequential forward-feature selection. The six features that presented themselves as most useful (in a texture-based classification problem, fed into a radial-basis SVM) were:

- Mean
- Variance
- Smoothness
- Skewness
- Uniformity
- Entropy

I think it would be worthwhile getting the definitions for these should any texture-based classification work come up.

I need to start logging to-dos using the Office 365 Planner app.

## 2 PV Panels

The first dataset is a 50 scans of PV panels. Possible objectives with these panels are:

- Detecting cracks.
- Detecting pit marks.
- Detecting blown cells.

An example image is shown in Figure 1. I made a repo for the PV project.

## 3 PV Panels: Crack detection

The objective of this task is to label images as according to whether or not they contain a crack - I don't think bounding boxes are appropriate. I'm going to begin by:



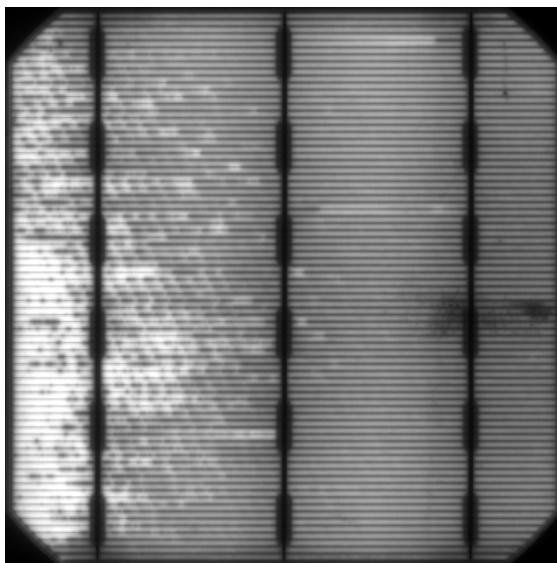


Figure 1: A sample image from the PV panels dataset.

- Rescaling the images to a constant size.
- Labeling the panels according to the damage they've incurred.
- Sorting the panels according to their type.

I created `panel-resizer.ipynb` to rescale the panel images to a fixed size<sup>1</sup>. This took me about an hour. While doing this, I realized that I could cluster the panels according to their metropolis distance from one another, generate an 'undamaged' panel for each class, then use subtraction to identify damaged panels. I think the biggest weakness with such an approach, however, is the need for an 'undamaged' panel image. How is easy would this be to generate synthetically?

1. Label images (crack/no crack).
2. 70/30 split of training/test data.
3. Benchmark 'dumb' classifiers.
4. Select the best dumb classifier and optimize it.

---

<sup>1</sup>I deleted one duplicated image (52) and removed three smaller images

# Thursday, 22nd June 2017

## 1 Summary

Today I planned to:

- Review what I covered yesterday.
- Read the blog post “Classifying plankton with deep neural networks”.
- Label at least 50% of the PV panel data.
- Set up a LaTeX report for the PV panel data.
- Update my current strategy for the PV data on Office 365.
- Complete problem exercises from Ch. 6 of *Computer Vision*.
- Implement the algorithms presented in Ch. 6 of *Computer Vision* on a dummy dataset.
- Benchmark 3 dumb classifiers against the labeled PV panel data.
- Read Ch. 1 of Boyd’s *Convex Optimisation*.

If there is time I will also begin reading Ch. 7 of *Computer Vision*.

(07:30 - 08:00) I planned my approach to the PV panel crack detection task on Office 365.

(08:00 - 11:00)

## 2 PV Panels: Crack Detection

To-dos:

- Label subset images
- Build augmented images dataset
- Set up report
- Benchmark simple classifiers against augmented images

- Optimize best simple classifier
- Record simple classifier performances in report
- Set up Azure instance and train convnet against augmented images
- Upload augmented images dataset
- Investigate clustering of panels

### 3 Prince: Computer Vision (Ch. 5/6)

Yesterday my work from this book constituted reviewing the many properties of the multivariate normal distribution, among which were:

- The marginal distribution of a subset of a vector of normal r.v.s is also normal.
- The conditional distribution of a subset of vector of normal r.v.s, given the other r.v.'s values, is also a normal distribution.
- Its covariance matrix can be diagonalized using the singular value decomposition, which forms a rotation matrix that twists the reference frame to align with the axes of the distribution's ellipsoidal isodensity contours.
- It can have one of three types of covariance matrix - spherical, diagonal, or full - which determines how elliptical its isodensity contours are and what its orientation relative to the coordinate axes is.

In the evening I began work on Ch. 6, *Learning and Inference in Vision*. It defined the basic task of machine vision as inferring world properties based on image data. It also outlined a framework for solving this problem consisting of three components:

- A model relating image data to world properties.
- A learning algorithm for adjusting that model's parameters based on observations.
- An inference algorithm for returning a predicted world state based on input image data.

The model can be either discriminative or generative. A discriminative model describes  $\Pr(\mathbf{w}|\mathbf{x})$ , the world state based on the image data. By contrast, a generative model instead represents the likelihood  $\Pr(\mathbf{x}|\mathbf{w})$  - what the image data should be for a given world state.

(08:20 - 9:20) Made notes on modeling a univariate continuous response with univariate continuous inputs using discriminative and generative methods. Also described how to model a binary response in terms of continuous image data, again via both generative and discriminative methods.

- Univariate continuous data  $x$ , univariate continuous world state  $w$

Thursday, 22nd June 2017

- Discriminative approach: Model  $\Pr(w|x)$  using  $\text{Norm}_w[\phi_0 + \phi_1 x, \sigma^2]$ . Fit  $(\phi_0, \phi_1, \sigma^2)$  based on training set  $\{x_i, w_i\}_{i=1}^I$ .
- Generative approach: Model  $\Pr(x|w)$  using  $\text{Norm}_x[\phi_0 + \phi_1 w, \sigma^2]$ , and  $\Pr(w)$  using  $\text{Norm}_w[\mu_p, \sigma_p^2]$ . Fit  $(\mu_p, \sigma_p^2)$  using the available world states  $\{w_i\}_{i=1}^I$ . Fit  $(\phi_0, \phi_1, \sigma^2)$  using the training data  $\{x_i, w_i\}_{i=1}^I$ . Compute the posterior using Bayes' rule.
- Univariate continuous data  $x$ , binary world state  $w$ .
  - Discriminative approach: Model  $\Pr(w|x)$  using a Bernoulli distribution parameterized by  $\lambda = \text{sig}[\phi_0 + \phi_1 x] = \frac{1}{1 + \exp(-\phi_0 - \phi_1 x)}$ . Fit  $(\phi_0, \phi_1)$  using the training data.
  - Generative approach: Set up a normal distribution over the data  $\Pr(x|w) = \text{Norm}_x[\phi_0 + \phi_1 w, \sigma^2]$  such that its parameters switch according to the world state  $w \in \{0, 1\}$ . Use a Bernoulli distribution for the world state prior  $\Pr(w) = \text{Bern}_w[\lambda_p]$  - fit  $\lambda$  using the observed world states. For inference compute the posterior at the chosen set of inputs.

(09:30 - 10:10) Read through skin detection and background subtraction examples and the chapter summary. The skin detection algorithm set up a generative model describing pixel intensity based on skin/non-skin status via a multivariate normal distribution. The prior was a Bernoulli distribution on the world states. By contrast, the background subtraction example set up a likelihood function wherein the pixel intensities of foreground objects were described using a uniform distribution (i.e. all combinations of  $(x_R, x_G, x_B)$  were deemed equally likely) and the intensity at each background pixel in each scene was expressed in terms of a multivariate normal distribution.

(10:20 - 11:45) Completed Ch.6 exercises 1, 3, 5, 7. Related modelling a continuous response against a binary world state to quadratic discriminant analysis and linear discriminant analysis (I ended up seeing this when comparing the for the decision boundary formed by two normal distributions with equiprobable associated classes with the decision boundary derived from a logistic regression classifier).

(12:50 - 13:45) Labeled PV panels dataset.

(14:00 - 16:05) Read (and made notes on) Ch. 7 of *Computer Vision*.

(16:05 - 16:45) Typed up lab notes from this afternoon's work on hidden variables and the EM algorithm.

(19:15 - 22:15) Augmenting PV panel images. Image augmentations:

- vertical flip
- rescaling
- adding noise

## 4 Prince, Computer Vision (Ch. 7, Modeling complex data densities)

This chapter frames the methods introduced by considering the shortcomings of attempting to build a generative model for a face classifier. Given a worldstate corresponding to face/no face,  $w \in \{0, 1\}$ , what is the probability density function for each pixel's intensity in an image of a fixed size? We might choose to model this likelihood using a normal distribution whose parameters are fit using training data:

$$\Pr(\mathbf{x}|w) = \text{Norm}_{\mathbf{x}}[\mu, \Sigma]$$

Notice that given  $w = 1$ , the parameters will be  $\{\mu_1, \Sigma_1\}$ , and given  $w = 0$  the parameters will be  $\{\mu_0, \Sigma_0\}$ . By defining a prior over the world state (a Bernoulli distribution whose parameter  $\lambda$  is fit using training data), we can get a hold of a posterior describing the probable world state, given a particular image's set of pixel intensities,  $\Pr(w|\mathbf{x})$ .

This modeling strategy has several shortcomings:

- The normal distribution is sensitive to outliers - one extreme observation will set estimates of its mean and covariance way out of kilter.
- A unimodal distribution doesn't describe what pixel intensities are probable if someone's face is in an image - people have different skin, eye, and hair colours.
- An outrageous number of parameters needs to be estimated. Given a  $30 \times 30$  RGB image, 2700 mean intensity values and  $\frac{2700 \times 2701}{2}$  covariance values need to be estimated. To make this estimate unique then, about three-and-a-half million training examples are needed.

To get around these problems we can - respectively - use t-distributions, mixture models, and subspace models. All of these tools exploit hidden variables. A joint distribution over the variable being modeled  $\mathbf{x}$  and a hidden variable  $\mathbf{h}$  is defined such that when  $\Pr(\mathbf{x}, \mathbf{h})$  is marginalized over  $\mathbf{h}$ , a complex density function of  $\mathbf{x}$  emerges. As with any distribution, this joint distribution is parameterized by some set of values  $\theta$  which are chosen to maximize the likelihood function:

$$\hat{\theta} = \arg \max_{\theta} \left[ \sum_{i=1}^I \log \left[ \int \Pr(\mathbf{x}_i, \mathbf{h}_i | \theta) d\mathbf{h}_i \right] \right]$$

The expectation maximization algorithm is used to maximize the likelihood. This algorithm proceeds by iteratively raising a lower bound on the likelihood function. The lower bound is a function of the parameters  $\theta$  and density distributions over the hidden variables. On the  $t^{\text{th}}$  iteration of the algorithm:

1. The boundary is raised as far as possible by setting the density functions  $q_i(\mathbf{h}_i)$  to be  $\Pr(\mathbf{h}_i | \mathbf{x}_i, \theta^{[t-1]})$ .

*Thursday, 22nd June 2017*

2.  $\theta^{[t]}$  is chosen to maximize the log-likelihood for the given density functions.

The first step is called the E-step (for expectation) and the second step is called the M-step (for maximization). By executing this algorithm repeatedly, it is possible to converge upon a local maximum in the likelihood function (which may also be the global maximum).

Use various scales of panel images to generate more data.

# Friday, 23rd June 2017

(08:00 - 13:30) Read and made notes on Ch.7 of *Computer Vision*. Covered hidden variables, expectation maximization, mixture of gaussians, the t-distribution, and factor analysis. Spent three hours building a mixture of gaussians model using World Happiness Report data. Figure 1 shows a fit being made with four component distributions. It may not look like much, but I learnt a lot from doing it! (16:00 - 17:45) Made notes on

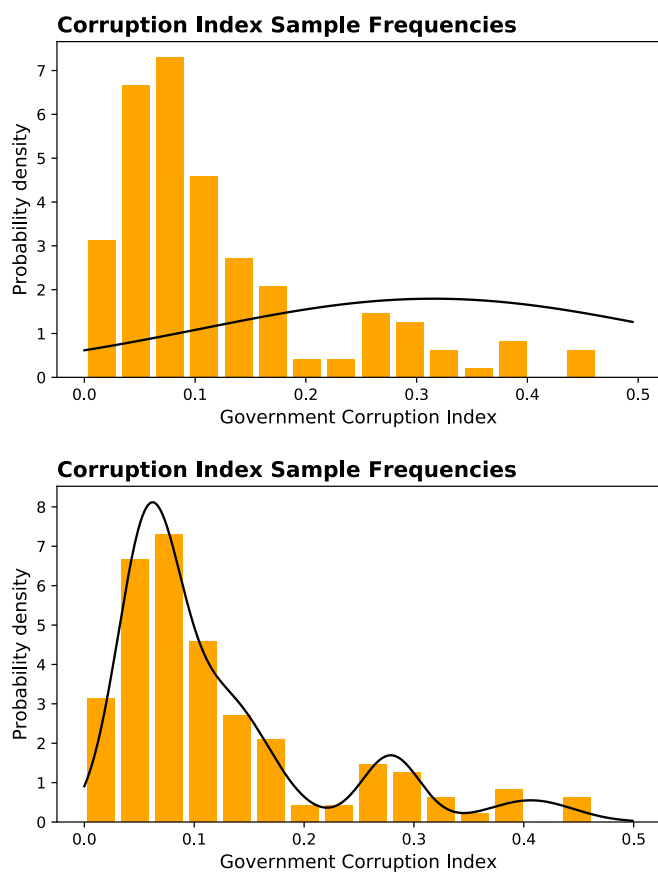


Figure 1: Fitting a four-component mixture of gaussians model using expectation maximization.

factor analyzers.

Friday, 23rd June 2017

(17:45 - 20:00) Ironed out some bugs in the PV panel organizer. Ran a random forest and an rbf SVM against augmented training data (they did all right, 80% accuracy). I'll type out what I did in more detail later.

## 1 Prince, Computer Vision (Ch.7 Modeling Complex Data Densities)

So today I covered three things, all of in this chapter, all bearing a single commonality. They were introduced as tools with which to address weaknesses of using a normal model for the likely intensities of a pixel (or some other continuous feature) for a given world state. In particular, a single peak probably doesn't represent reality very well, the parameter fits are very sensitive to outliers, and a vast number of parameters need to be determined as a result of the covariance matrix. The solutions to these problems, respectively, were:

- Using a mixture of gaussians instead of one.
- Using a t-distribution rather than a normal distribution.
- Reducing the number of parameters to be determined by only modeling the co-variation in directions where it is strongest.

Neatly, it turns out that all of these model variations can be represented in terms of hidden variables (which I described yesterday). A mixture of gaussians uses a categorically distributed hidden variable to weight the various normal distributions, each of which have their own covariance and mean. A variable with a t-distribution can actually be thought of as a normally distributed variable whose variance is parameterised by a scalar gamma-distributed variable. That is,

$$\begin{aligned}\Pr(\mathbf{x}|\mathbf{h}) &= \text{Norm}_{\mathbf{x}_i}[\mu, \Sigma/h] \\ \Pr(\mathbf{h}) &= \text{Gamma}_h[\nu/2, \nu/2]\end{aligned}$$

While I had previously appreciated that the t-distribution's greater spread is a consequence of uncertainty w.r.t. its variance, this formulation made it especially clear what the role played by the degrees of freedom is.

The third solution - factor analyzers - is something I'm still wrapping my head mathematically but I think I understand in principle. Sample a vector  $\mathbf{h}$  from a multivariate standard normal distribution then use it to adjust the mean of a normal distribution to somewhere in the subspace described by its factors. Sample  $\mathbf{x}$  from this shifted normal distribution. The hidden variable here is the sampled vector: the distribution of samples from the shifted normal distribution turns out to be normal also, but with a covariance such that:

$$\Pr(\mathbf{x}|\mathbf{h}) = \text{Norm}_{\mathbf{x}_i}[\mu, \phi\phi^T + \Sigma]$$



Where  $\phi$  is a  $D \times K$  matrix whose columns are the basis vectors (factors) of the subspace and  $\Sigma$  is a  $D \times D$  diagonal (or spherical) covariance matrix. The factors are chosen to describe as much of the covariation in the full columnspace as possible, with the diagonal covariance matrix capturing the meat of the variance: a factor analyzer attempts to capture as much of the variation in the data's full  $D$ -dimensional columnspace in terms of an  $\mathbb{R}^K$ -dimensional subspace. One of the things I'm still trying to understand is why  $\phi\phi^T$  represents covariation in the subspace - I plan to look at this today (I wrote the follow-up to the 23<sup>rd</sup>'s work on the 24<sup>th</sup>). As an aside, the author mentioned that when  $\Sigma$  is a spherical covariance matrix (that is, a scalar multiple of the identity matrix), this method is called principal component analysis. The benefit of using a spherical rather than a diagonal matrix is that it reduces further still the number of parameters to be fit (hence its widespread use when wrangling wide data).

The reason the author framed the t-distribution, mixture of gaussian, and factor analysis model in terms of hidden variables is because density functions that can be expressed as a marginalization over a hidden variable (Equation 1) lend themselves to optimization via expectation maximization.

$$\Pr(\mathbf{x}|\theta) = \int \Pr(\mathbf{x}, \mathbf{h}|\theta) \cdot d\mathbf{h}$$

The model parameters  $\theta$  can be fit by iterative maximization of a bound on the l.h.s.'s likelihood function. This bound is a function of the r.h.s.; EM raises this boundary by iteratively:

- updating density functions over the hidden variables,
- maximizing the boundary's value by adjusting the values of the parameters in  $\theta$ .

Eventually, the boundary converges on the log-likelihood function such that a set of locally optimal parameter values are found.

As I said, I think there are still things about factor analyzers I'd like to understand; I would quite like to build a dummy model tomorrow to get more of a feel for how they work. I also need to look at the proof the boundary that the EM algorithm exploits and understand what its limitations are.

## 2 PV Panels: Crack detection

As I briefly mentioned in today's summary, I fixed some bugs in the `panel-organizer` set of scripts and ran some simple classifiers against them, namely a random forest and an rbf SVM. Like I said, they both achieved accuracies of  $\approx 80\%$ . Given that these models don't exploit the spatial information available in the images (I downsampled them from  $159 \times 159$  to  $60 \times 60$ , then flattened them), I'm optimistic about what we can achieve. I am aware that there might still be a bug w.r.t. class balances (I think I may have generated too many augmented crack images) and that only augmenting cracked panels

*Friday, 23rd June 2017*

is methodologically flawed (augmentation introduces distinctive features to the images), problems I intend to resolve tomorrow. I would also like to experiment with capturing spatial data (e.g. HOGs or moments, I feel that reading would also help here). It may also be the case that segmenting the images might help, as might clustering models of panel. I also need to set up a report for the panel work and conceive of a system to log classifier/preprocessing performance (or borrow a method from someone smarter than me!).

# Saturday, 24th June 2017

I woke at 06:45 this morning to swim for an hour and a quarter (I did 64 lengths of a 30m pool - 1920m). I had never swum more than 2 or 3 lengths in a pool before: my form was terrible and for the first five lengths I had difficulty breathing. The breathing trouble made me slightly panicked, which made it even harder to breathe (and so on). I pushed through however, and by the end I was breast-stroking like an enthusiastic amateur. I also felt pretty great. I swam the lengths as part of a relay effort parallel to a friend's Channel attempt (1000 lengths, 30km) and felt that doing the lengths helped me to better understand their achievement.

I spent a couple of hours with MZ discussing various things, most notably our 4<sup>th</sup> year project, a book on scarcity he is reading, and the VEST and Fulbright scholarships. I intend to apply for both of these scholarships, so I should probably capture the information here.

(12:00 - 14:40) I spent the first hour-and-a-half reviewing what I worked on yesterday and basically attempting to orient myself for the forthcoming week. I then an hour understanding the Fulbright and Vest scholarships. With regards to these, at the beginning of August I need to begin drafting my university and funding applications. I should also make a schedule of application deadlines and requirements. Register for ACT/SAT tests. This evening and tomorrow morning I should make time to plan my research placement more thoroughly.

(14:40 - 16:40) Detailed notes on the EM algorithm (yay understanding). Prove Jensen's inequality.

(16:40 - 19:10) Wrote source code for identifying faces using a factor analyzer. Found dataset (<http://cswww.essex.ac.uk/mv/allfaces/faces94.html>), imported and labeled the data.

Today I would like to neaten up my understanding of the EM algorithm, build a factor analyzer, and read the first chapter of Boyd's *Convex Optimization*.

## 1 Fulbright Scholarship

Fulbright scholarships are awarded by the US Bureau of Educational and Cultural affairs under the Department of State. Approximately 800 scholarships are awarded each year to international students (Visiting Scholars). Emphasises leadership, commitment to

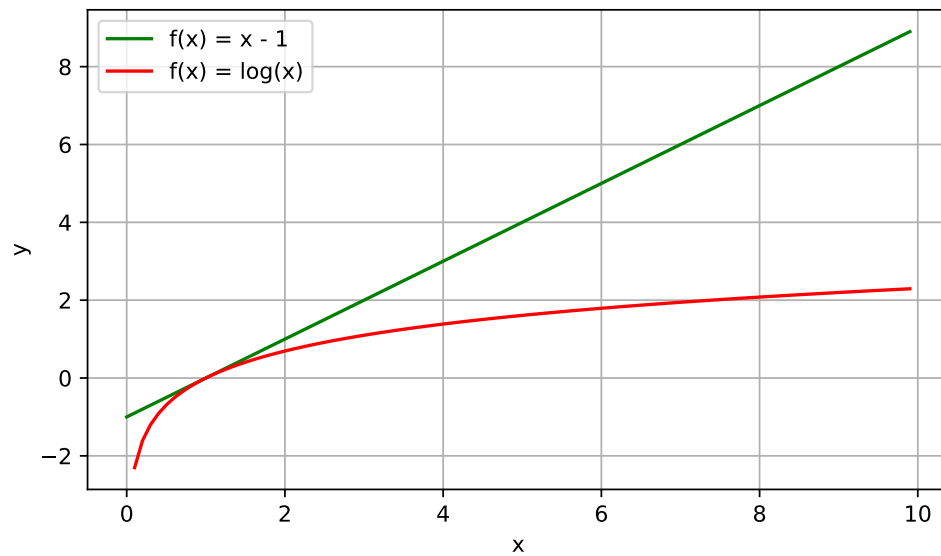


Figure 1:

international engagement, and cultural interaction. Fulbrighters seek to create a more peaceful and prosperous world. Funding is available for research and/or lecturing.

Benefits of studying in the USA:

- Quality of research
- Flexible and interdisciplinary
- International experience
- Funding

Fulbright postgraduate scholarships (<http://www.fulbright.org.uk/going-to-the-usa/postgraduate>): any master's/doctoral degree at any graduate programme at a US university. The scholarship offers:

- Financial contributions towards the first year of study
- Sickness and accident coverage
- J-1 visa sponsorship and pre-departure orientation
- Support whilst in the USA
- Access to the Fulbright alumni network

The programme aims to foster cultural understanding between the US and UK - little previous experience with the USA is desirable. Awards available:

- Brown University: PhD in any discipline.
- The New School: Master's degree in any discipline.
- Lloyd's of London: Master's degree in any discipline related to risk.
- University of South Florida: PhD in any discipline.
- Fulbright: Master's/PhD/research at any university.
- Elsevier Data Analytics: Postgraduate study requiring data analysis or research at any university.

"We are looking for applicants who can engage with the USA in an atmosphere of openness, academic integrity and intellectual freedom, thereby promoting mutual understanding." Desirable applicants would be able to demonstrate:

- Strong academic proposals and/or research projects
- Ambassadorial skills
- Intercultural sensitivity
- Genuine desire to learn aspects of American culture
- Extracurricular activities
- Community involvement
- Leadership potential
- Plans to further the Fulbright mission and give back to the UK upon returning

Applications process:

1. Read the UK-specific instructions.
2. Applications open: link on the bottom of this page (<http://www.fulbright.org.uk/going-to-the-usa/postgraduate/fulbright-postgraduate-scholarships/how-to-apply>). This requires bio information, academic/professional accomplishments, Fulbright project plans, personal statement, details of the US institutions that you intend to apply to, and three reference nominations.
3. Upload the US-UK Fulbright Commission Form to this applications.
4. Upload other information (passport, CV, transcripts, US university admission letter, portfolio)

Applications to US universities must be made separately and in good time. If shortlisted, the interview will be in January or February. Status is notified in March. Applications open in August. The British Fulbright Commission provide lots of information regarding

*Saturday, 24th June 2017*

American University's application procedures (see here: <http://www.fulbright.org.uk/going-to-the-usa/postgraduate/educationusa-advice>).

## **2 VEST Scholarship**

The Charles M. Vest NAE Grand Challenges for Engineering International Scholarship Program provides an opportunity for graduate students at international universities to pursue research addressing a global Grand Challenge at leading United States university. The goals of the Vest Scholarships are to provide a platform to exchange ideas, share problem-solving skills and strengthen international relationships in order to advance progress in some of the most critical global challenges in the twenty-first century. The scholarship offers;

- The opportunity to spend a year at a leading US engineering school pursuing research.
- Living and travel expenses and tuition covered by the host institution for the 12-month duration of the scholarship.
- An opportunity to perform research in the laboratory of a leading faculty scholar.
- Access to engineering classes and academic credit towards graduate degrees.
- 10 scholarships available - one at each partner institution.

Applicants must be enrolled in a graduate-level engineering program at the time of application and during the scholarship year. Applicants will be asked to submit a personal statement proposing a one-year research and study plan that addresses one of the 14 Grand Challenges for Engineering. Winning statements will be those that present the most compelling and potentially impactful proposal, and that map best onto Grand Challenge activities at the participating institutions. Academic aptitude and potential for success will be an important part of the selection process. Applications open August 1 and close November 1. The winners are contacted on February 15, 2018.

The Grand Challenges relevant to me are:

- Engineer the tools of scientific discovery (Caltech, Duke, Washington)
- Advance health informatics (MIT, North Carolina, University of Minnesota)
- Advance personalized learning (MIT, Olin)

It is only possible to apply under one Challenge. Read more about each challenge and institution at <https://vestscholars.org/research-challenges>.

# Sunday, 25th June 2017

(08:45 - 09:45) I overslept this morning (it's a Sunday), but I can pick up the hours this afternoon/evening. Yesterday I built part of a factor analyzer, produced detailed notes on the EM algorithm - namely a proof for the bound used and a proof for the choice of density function in the E-step - and read about the Fulbright and Vest scholarships.

Today I would like to:

- Finish the factor analyzer
- Email MM and ET with a review of what I've been up to this week:
  - Reading: generative models of data using hidden variables and the EM algorithm (esp. mixture of gaussians and factor analyzers). Built a univariate MoG model and a factor analyzer for face identification. Read Canny *A Computational Approach to Edge Detection*. Read paper *Detection and classification of surface defects of gun barrels using computer vision and machine learning*. I started a MOOC focused on convex optimization.
  - DNV: IT induction. Gained access to wind turbine and PV panel datasets. Began work on PV panel problem to detect cracked panels. I built several simple classifiers (a random forest, a boosted forest, and a radial basis SVM) against a subset of the data.
  - NW Reading: Bayesian linear, nonlinear, and sparse linear regression models (2 days). Bayesian logistic, nonlinear logistic, and boosting models (2 days). Graphical models (2 days).
  - NW DNV: If Azure access is available, train a convnet against the solar panel dataset. Set up a dummy generative adversarial model for synthetic data generation.
- Read the GAN paper.
- Fix the size of the augmented dataset. Change the programme such that it augments both cracked and non-cracked panels (add Bernoulli variable to speckling).
- Build a HOG feature generator.
- What's SIFT?
- Use `skimage`'s tools to segment the panels.

Sunday, 25th June 2017

- Email family describing what you've been up to this week.

There are 7 weeks left as of right now:

- Week 2:
  - Read five papers. (9 hours)
  - Complete two sets of lectures from *Convex Optimization* (12.5 hours).
  - Complete Chapters 8, 9, 10 of *Computer Vision* (25 hours).
  - Complete as much of the TensorFlow Udacity course as you can. (25 hours).
  - Complete and report the PV panel crack detection problem (12.5 hours).
  -
- Week 3:
  - Read five papers. (9 hours)
  - Complete two sets of lectures from *Convex Optimization* (12.5 hours).
  - Complete Chapters 11, 12, 13 of *Computer Vision* (25 hours).
  - Finish the TensorFlow Udacity course. (25 hours).
  - Choose a research problem to solve.
  - Complete and report the PV panel crack detection problem (12.5 hours).
- Week 4:
  - Read five papers. (9 hours)
  - Complete two sets of lectures from *Convex Optimization* (12.5 hours).
  - Complete Chapters 14, 15, 16 of *Computer Vision* (25 hours).
  - Work on the research problem (37.5 hours).
  - Set a reminder to register for NIPS.
  - Set a reminder to start investigating postgraduate degrees at universities (and schedule submission dates, exams and so on).
- Week 5:
  - TBD

(12:30 -14:00 ) Reviewed yesterday's work, made notes on the applications of the models discussed in Ch.7 of *Computer Vision*. These included regression onto missing data (described in the context of reorienting faces), using t-distributions to model patches of images for object classification, identifying faces using factor analysis, and understanding transformations as hidden variables.



(14:00 - 16:25) Building factor analyzer for face recognition (likelihood fitter)

(18:00 - 21:00) Fixed the fitting part of the factor analyzer (yay)

Hours worked this week:

- Monday: Untracked.
- Tuesday: Untracked.
- Wednesday: 8.5hrs
- Thursday: 11hrs
- Friday: 7.75hrs
- Saturday: 7hrs
- Sunday: 8hrs

## 1 Prince, Computer Vision (Ch. 7, Modeling Complex Data Densities)

Yesterday I finished this chapter by re-covering factor analyzers - generative models that use a subspace to quantify covariance rather than the full feature space - and tying up the loose ends in my understanding of the EM algorithm. The particular loose ends covered were:

- Proof that the boundary function is less than the log-likelihood. That is, we proved that:

$$\sum_{i=1}^I \int q_i(h_i) \log \left[ \frac{\Pr(h_i, x_i | \theta)}{q_i(h_i)} \right] dh_i \leq \sum_{i=1}^I \log \left[ \int q_i(h_i) \frac{\Pr(h_i, x_i | \theta)}{q_i(h_i)} dh_i \right]$$

Or, in a form that gives a better insight into what's actually going on:

$$\sum_{i=1}^I E \left[ \log \left[ \frac{\Pr(x_i, h_i | \theta)}{q_i(h_i)} \right] \right] \leq \sum_{i=1}^I \log \left[ E \left[ \frac{\Pr(x_i, h_i | \theta)}{q_i(h_i)} \right] \right]$$

The proof for this inequality relies on Jensen's inequality, which states that:

$$E[\log y] \leq \log[Ey]$$

I haven't proven this, but logarithms tend to squish big numbers, hence shrinking their expectations. I think it's the case that this form of boundary is chosen because it makes it clearer what the appropriate density functions  $q_i(h_i)$  to maximize the boundary are.

Sunday, 25th June 2017

- In particular, we proved that the choice of  $q_i(h_i)$  in the E-step should be the hidden variable's posterior given the data and parameters at that step:

$$\hat{q}_i(h_i) = \Pr(h_i|x_i, \theta^{[t]})$$

We did this by rearranging the optimization problem of the E-step to bring out the Kullback-Leibler divergence:

$$q_i(h_i) \cdot \log \left[ \frac{\Pr(h_i|x_i, \theta)}{q_i(h_i)} \right]$$

which can be shown to always be negative, further meaning that the boundary-maximizing choice of  $q_i(h_i)$  is indeed the aforementioned posterior.

- On a high level, I understood why the boundary function is always less than or equal to the log-likelihood of the joint distribution of  $\Pr(x_i, h_i|\theta)$ , and what choice of  $q_i(h_i)$  satisfies the equality of the two for a given choice of model parameters  $\theta$ . The EM algorithm is useful because it allows nonlinear functions to be optimized, provided that they can be expressed in terms of a hidden variable.

After making notes on this, I started building a factor analyzer to identify faces from images. I wrote out the source code; today I intend to create a functional programme. I think I'll also go through the applications, since these are what have really been allowing me to convert theoretical knowledge to practical utility.

Idea: regressing cracks onto the rest of their panels, then conditioning over an incomplete uncracked panel to generate a cracked panel image (See pp. 104 of *Computer Vision*).

I think I would benefit most from Chapter 7's problems if I were to do them in the middle of next week (say Thursday) rather than right now.

## 2 Prince, Computer Vision (Ch. 8, Regression models)



Figure 1: The mean (left) and single factor (right) of a grayscale image from the factor analyzer.

# Monday, 26th June 2017

## 1 Summary

(08:00 - 11:00) Prince's *Computer Vision*, Ch. 8. Linear regression, Bayesian linear regression, nonlinear regression (inc. kernels, radial basis functions, and gaussian process regression), sparse linear regression.

(11:15 - 13:00) Tensorflow lectures - received first assignment.

(15:15 - 16:45) Sent a weekly update to ET and MM. It was as follows:

Hello Professor Mirmehdi,

I hope you're well. Paul Harper introduced us several months ago: I'm an undergraduate researcher in machine vision working at the University/DNV-GL. Pleased to meet you (again).

My reason for emailing is that since you are my point of contact with the University and are an expert in machine vision, I thought it would be a good idea if gave you a weekly update of what I've been up to. This is so that:

- (a) you know I'm spending my placement time productively
- (b) I can benefit from your guidance as to whether I'm following an unpromising path / could make use of some theory/tools in particular.

I appreciate that you are busy and will try to keep these emails succinct - if you'd rather not hear from me, please let me know. Anyway, last week I:

- Spent a lot of time reading about how to use the expectation maximisation algorithm. I built a factor analyser to identify faces in this dataset (<http://cswwww.essex.ac.uk/mv/allfaces/index.html>) and a Gaussian mixture model against another dataset. I know that these are quite old-school methods, but I think understanding them will aid my general grasp of machine learning and vision.
- Began working on classifying grayscale images of cracked PV panels for DNV. I've cleaned, subset, augmented and split a dataset that was originally of about 50 unique panels. I also ran a random forest and a radial basis function support vector machine against the training set, without any preprocessing (apart from standardisation). I was planning to use a histogram of oriented gradients as the classifier's input (possibly in addition to the pixel intensities?). This is because the cracks seem to be distinguished by their orientation. I may try a convnet, but think it will probably overfit.

Monday, 26th June 2017

- Started a course on numerical methods for convex optimisation. This is in an effort to understand how to phrase optimisation problems in a sensible way. I think the first few weeks focus on theory, after which there's a strong emphasis on applications.
- Became more interested in methods for generating synthetic data. I was thinking of building a generative adversarial network (GAN, a pair of competing neural networks, one of which attempts to generate fake images, the other which tries to spot images that are fake) and looking at using generative models of cracked PV panels to transplant fake cracks onto images of uncracked panels

This week I was going to finish the PV panel classifier and summarise its performance in a report. I have also allocated a big chunk of time to reading, some of which will go towards reading about and building GANs.

Do you have time to meet this week? I was had a few questions about choice of model and preprocessing procedure for the PV panel classification task. Also, your suggestions of what I should read about/work on to make the placement worthwhile would be really useful to me.

Thank you and have a great week,

Jerome

(16:45 - 17:45) Tensorflow assignment.

(17:55 - 18:40) Read Ch. 1 of *Convex optimization*. The course begins by focusing on the maths of convex sets, functions, and optimization, before moving onto applications and algorithms. A convex function is a function that is less than or equal to a line between two points in its

(18:45 - 19:45) Finished the Tensorflow assignment. It focused on data cleaning and fitting a logistic regression classifier. It didn't actually involve using TF! I used the package `tmmd`, which made it easy to associate loading bars with `for` loops. My memory of pickling was also refreshed - it is a pretty convenient way to store large Python structures on disk for access at short notice.

## 2 Prince, Computer Vision (Ch. 8, Regression models)

After the previous chapter's focus on various generative models, which modeled the data's distribution for given world states, this chapter focuses on discriminative methods for regression. It began by introducing a simple linear model for a univariate continuous world state  $w$  and continuous multivariate vector  $\mathbf{x}$ :

$$\Pr(\mathbf{w}|\mathbf{X}, \phi) = \text{Norm}_w[\mathbf{X}^T \phi, \sigma^2]$$

Where  $\mathbf{X}$  is a  $D \times I$  matrix of  $I$  examples, each of  $D$  dimensions. In other words, we assume that the mean response is a linear function of the inputs  $\mathbf{x}$ , with constant

variance. The model's parameters can be fit using maximum likelihood, maximum a posteriori, or bayesian analysis. The ML estimate of the parameters is:

$$\phi = (\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}^T\mathbf{w}$$

$$\sigma^2 = \frac{(\mathbf{w} - \mathbf{X}^T\phi)^T(\mathbf{w} - \mathbf{X}^T\phi)}{I}$$

Notice that here the likelihood being maximized is  $\Pr(w|x, \theta)$ , whereas when fitting the generative methods we maximized the likelihood  $\Pr(x|w, \theta)$ .

Several flaws in the simple linear model were pointed out:

- There is no embedded representation of the uncertainty associated with extrapolation - the model is as confident where there have been many observations as it is where there are none.
- A linear function may not be representative of the true relationship between the inputs and mean response.
- Many of the input features may be redundant or of little predictive power, but will be assigned tiny coefficients upon fitting the model - this makes the model needlessly complex.

To represent the uncertainty in the model's coefficients more accurately, we can model the parameters as normally distributed variables:

$$\Pr(\phi) = \text{Norm}_\phi[\mathbf{0}, \sigma_p^2]$$

Where  $\sigma_p^2$  is a hyperparameter (we'd usually fix it to a large value to represent our initial uncertainty). Then, when fitting the model, we can use Bayesian analysis rather than maximum likelihood:

$$\Pr(\phi|\mathbf{X}, \mathbf{w}) = \frac{\Pr(\mathbf{w}|\mathbf{X}, \phi) \cdot \Pr(\phi)}{\Pr(\mathbf{w}|\mathbf{X})}$$

This yields a normal posterior distribution over the parameter values with a covariance that increases with  $\mathbf{x}$ . To make predictions using this distribution, compute the posterior predictive and evaluate its value at the query world state and data.

Several matrix identities were introduced in this section with the intention of making matrix inverse calculations less expensive. One of these was the Sherman-Morrison-Woodbury identity, which can be used to speed up posterior and posterior predictive inference in the aforementioned Bayesian model.

To work around the issue that the expected response and the inputs are unlikely to have a linear relationship, the model was generalized to nonlinear functions in neat way: simply replace the raw input vector  $\mathbf{x}$  with its value under nonlinear transformation:

$$\mathbf{z} = \mathbf{f}[\mathbf{x}]$$

Monday, 26th June 2017

Then model the relationship between the world state and input as a linear one:

$$\Pr(\mathbf{w}|\mathbf{Z}, \phi) = \text{Norm}_{\mathbf{w}}(\mathbf{Z}^T \phi, \sigma^2)$$

Apply the equivalent versions of the ML estimates or the Bayesian analysis methods mentioned earlier to determine the parameters, then evaluate this expression at a new query for inference. The parameters of the nonlinear functions can be defined beforehand, or they can be fit. To fit them, we need to marginalize the above equation over  $\phi$ :

$$\Pr(\mathbf{w}|\mathbf{Z}) = \int \Pr(\mathbf{w}|\mathbf{Z}, \phi) \cdot \Pr(\phi) \cdot d\phi$$

We can then maximize this - the evidence - against the nonlinear function's parameters  $\lambda$ :

$$\hat{\lambda} = \arg \max_{\lambda} [\Pr(\mathbf{w}|\mathbf{Z})]$$

Use a nonlinear optimization procedure (i.e. local search with multiple instantiations, global search) to solve this.

If  $\mathbf{z}$  is very high-dimensional, we can replace the dot products  $\mathbf{z}_i^T \mathbf{z}_j$  in the ML and Bayesian models with the kernel function  $k[\mathbf{x}_i, \mathbf{x}_j]$ .

The final area that I began to work on was sparse linear regression, which pushes coefficient values towards zero. By using prior distributions that favour zero-valued coefficients, it's possible to do exactly this.

# Tuesday, 27th June 2017

Today I would like to:

- Summarize yesterday's work
- Finish the Ch. 8 of *Computer Vision* (2 hours)
- Build at least one of the regression models described in this chapter (2 hours)
- Spend two hours on the Tensorflow lectures (2 hours)
- Spend four hours on the PV panel classifier: (4 hours)
  - Fix the augmenter balance.
  - Augment undamaged panels.
  - Build a gradient descriptor.
  - Document progress so far.
- Print and read GAN papers

## 1 Summary

(07:00 - 08:00) Summarized yesterday's work. (08:00 - 11:00) Sparse regression, dual linear regression.

Dual linear regression expresses the gradient vector in terms of the training data points. It's appropriate for situations in which the data has many more dimensions than examples.

Sparse linear regression places a t-distributed prior on each of the gradient values to represent the fact that many of the gradient elements may not be useful in predicting the response. To fit the mean and variance of the parameter vector of the sparse regression model, it is necessary to maximize an approximation of the likelihood function. To make this approximation, a function must be maximized with respect to the hidden variables of the t-distributions. In reality, the likelihood function and the parameters that maximize it are optimized simultaneously.

Tuesday, 27th June 2017

Relevance vector regression combines the principles of sparse and dual linear regression to produce an efficient regressor: a prior is placed on each element of the dual vector (which correspond to weightings of each observation) that pushes it towards zero, then the mean and variance of this dual vector's posterior is computed while simultaneously approximating the likelihood function by maximizing a function w.r.t. the t-distribution's hidden variables. Relevance vector machines describe predictive models to be described in terms of just a few datapoints, as opposed to attempting to weight high-dimensional input vectors or use all of the datapoints available (which makes optimization expensive).

(11:00 - 13:00) Swept the steps. Took Lunch.

(13:00 -14:00) Bio for The Bakery.

(14:00 - 15:00) More on sparse linear regression. I've found it quite difficult to grasp, and still feel as though I'm missing something. Say we have a linear model:

$$\Pr(w|X, \sigma^2, \phi) = \text{Normal}_w[X^T \phi, \sigma^2 I]$$

$X$  is a  $D \times I$  matrix of inputs,  $\sigma^2$  is common variance,  $w$  is a vector of  $I$  responses. A non-sparse Bayesian linear model would assume a normal prior distribution  $\text{Norm}_\phi[\mathbf{0}, \sigma_p^2]$  over the weights, where  $\sigma_p^2$  is large. In combination with the normal likelihood above, this would yield a normal posterior distribution:

$$\Pr(\phi|w, X, \sigma^2) = \frac{\Pr(w|X, \phi, \sigma^2) \cdot \Pr(\phi)}{\Pr(w|X, \sigma^2)}$$

Since the posterior and the likelihood were normal, the posterior predictive had a closed-form solution (yet another normal distribution). It was also possible to compute the posterior directly.

In a sparse linear model, the prior term is adjusted to weight zero-valued elements of  $\phi$  more heavily. The purpose of this is to identify input that have little predictive value. This is achieved using a multivariate  $t$ -distribution:

$$\Pr(\phi) = \prod_{d=1}^D \text{Stud}_{\phi_d}[0, 1, \nu]$$

Problematically, this means that there is no analytical expression for the posterior distribution over the weights  $\phi$ . We still need an expression for this posterior, however, to be able to express the posterior predictive distribution. The book asserts that we can estimate the parameters  $\mu$ ,  $\Sigma$  of a normal distribution that approximates the weight's true posterior by maximizing an approximation to the model's evidence,  $\Pr(w|X, \sigma^2)$ , with respect to the t-distribution's hidden variables and the model's common variance  $\sigma^2$ . I am puzzled about:

- (a) Where the normal posterior came from to approximate the product of a normal and t-distribution.

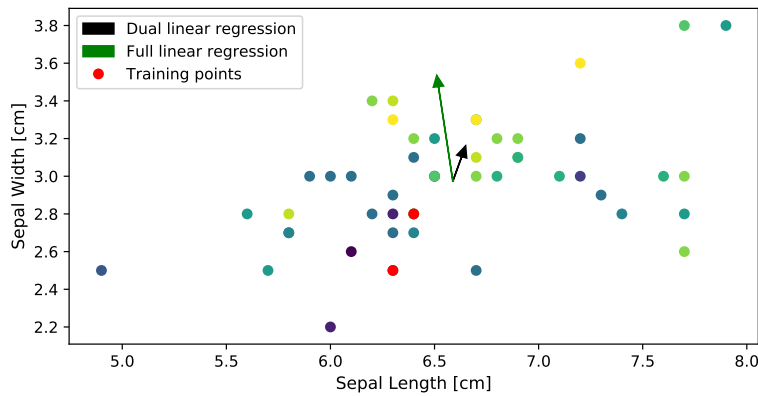


(b) How the values for the hidden variables influence the choice of weights.

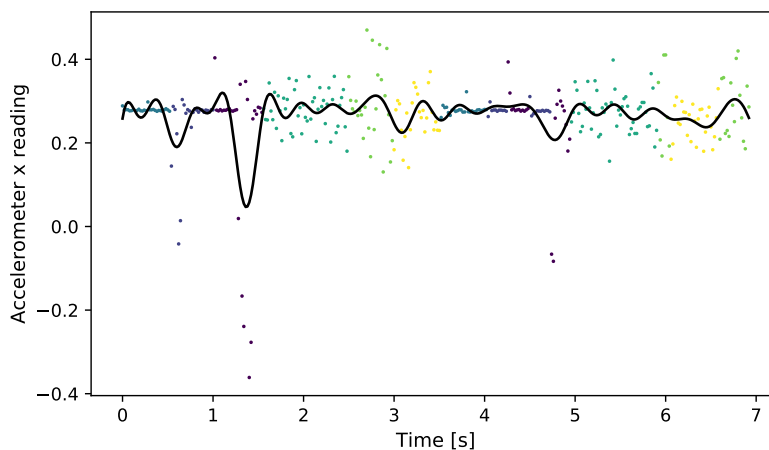
The former point I think I can appreciate in an unrigorous sense, but the latter needs more thought. The hidden variables of the t-distribution affects the distribution's variance - smaller values of  $h_d$  suggest a larger variance, further implying a more probable non-zero value for  $\phi_d$ . So the idea with solving for values of the hidden variables is to stretch the prior in the directions that are likely to have non-zero values. By contrast,  $\sigma^2$  has to be updated alongside the hidden variables because when the uncertainty in the values of the hidden variables changes, so too will the uncertainty in the response. I think I'm getting closer with this - maybe a few more readings...

(15:00 - 16:15) Wrote the above on sparse regression. Replied to John Ryan.

(16:15 - 17:50) Built a dual regression model on Iris dataset.



(17:50 - 19:45) Built a Gaussian nonlinear regression model.



Tuesday, 27th June 2017

labday Wednesday, 28th June 2017 Yesterday I built a nonlinear regression model and a dual linear regression model.

The nonlinear model has a simple premise - perform a nonlinear transformation of the inputs, then perform a linear regression of the transformed inputs onto the response. Kernels can be used to make this process much more efficient/feasible.

The dual model assumes that the gradient vector is in the subspace described by the observations. It's useful when the dimensionality of the data exceeds the number of observations. In such a situation, linear regression is not possible since its model would be under-constrained. Why is it valid to assume that the gradient vector lies in the subspace described by the observations? In short, the assumption is justified by the observation that if input variation is not observed in a given direction, then there is no information available that can be used to compute a gradient in that direction.

(09:30 - 12:30) *Convex Optimization* Lecture 2 - convex sets.

(13:45 - 17:30) Tensorflow work. Moved from the logistic regression model to neural networks by introducing nonlinear transfer functions (specifically, RELUs). Built a two-layer network for 10-way classification. Understood the TensorFlow workflow: set up a graph, pass the graph to a session. Reviewed stochastic and minibatch gradient descent, momentum (taking a running average of the loss gradient and using it to choose the direction of optimization for the current iteration), l2 regularization (penalizing large weights - why?), dropout (to encourage redundancy in the network), learning rate decay (to permit better overall convergence), and early termination (to avoid overfitting).

(19:00 - 21:15) Prince, *Computer Vision* (Ch. 9, Classification models). Logistic regression, Newton's method for making an ML fit of its parameters. Laplace approximation for a distribution. Integrating a multivariate normal distribution over one dimension instead of many via a linear projection. Highlighted the weaknesses of logistic regression:

- Using maximum likelihood to fit the parameters whitewashes the actual uncertainty in a class prediction (resolved via Bayesian logistic regression).
- Linear decision boundaries are sometimes inadequate.
- The model is expensive to fit in many dimensions, and is prone to overfitting in wide situations.

## 2 Boyd, Convex Optimisation (Lecture 2)

The lecture covered various basic types of convex sets, and operations that preserve convexity. Simple convex sets that he brought up included:

- The basic convex set:  $S : \{x_1, x_2 | \theta x_1 + (1 - \theta)x_2 \in S, \theta \in [0, 1]\}$

- Affine sets:  $\{x|\theta x_1 + (1 - \theta)x_2, \theta \in \mathbb{R}\}$
- Hyperplanes: constraints on single elements from an affine set;  $\{x|a^T x = b, a \in \mathbb{R}^n, b \in \mathbb{R}\}$
- Convex hulls: the set of all convex combinations of points in a set:  $x = \theta_1 x_1 + \dots + \theta_k x_k$  with  $\sum \theta_i = 1, \theta_i \geq 0 \forall i$ .
- Halfspaces: like a hyperplane, but with an inequality.  $\{x|a^T x \leq b\}$ .
- Convex cones: any linear combination of two points, provided that the coefficients are positive.  $C : \{x|x = \theta_1 x_1 + \theta_2 x_2, \theta_1, \theta_2 \geq 0\}$
- Euclidean balls and ellipsoids (linear transformations of a unit ball).  $\{x|(x - x_c)^T P^{-1} (x - x_c) \leq 1\} = \{x_c + Au | \|u\|_2 \leq 1\}$ . A euclidean ball is the set of points that satisfy  $\{x|\|x - x_c\|_2 \leq r\}$ . We also defined a norm as a function that satisfies various conditions.
- Polyhedra: A set of points satisfying finitely many inequalities (halfspaces) and equalities (hyperplanes).  $\{x|Ax \preceq b, Cx = d\}$ .

He emphasized that these basic convex sets in combination with convexity-preserving operations are useful for demonstrating a set's convexity. Alternatively, we can use the basic definition of convexity to show that a set is convex.

The second half of the lecture focused on these convexity-preserving operations, which were the:

- Intersection
- Affine function
- Perspective function - the division of the first  $n$  elements in a vector by the vector's  $n$ th and final element.

$$P(x, t) = \frac{x}{t}, \quad \text{dom} P = \{(x, t) | t > 0\}$$

- Linear-fractional function - the perspective function after an affine transformation.

$$f(x) = \frac{Ax + b}{c^T x + d}, \quad \text{dom} f = \{x | c^T x + d > 0\}$$

## Thursday, 29th June 2016

(07:10 - 12:50) PV panel classifier work

(14:00 - 15:20)(16:45 - 18:45) Computer Vision exercises (Ch. 5, ). Distribution of a multivariate normal r.v. under an affine transformation. Marginalizing a normal distribution using an affine transform. Standardizing a multivariate normal distribution. Proving the Schur complement identity. (Up to 5.5) (1, 2, 3, 4, 7, 8).

(18:45 - 20:45) Computer Vision exercises (Ch. 7) Describing generative models for a continuous input vector and a binary response. Described how the t-distribution can be used to make this type of model robust to outliers (e.g. mislabelled samples). Derived the component weight updates in the M-step of the EM algorithm when fitting a Gaussian mixture model. Learnt how to use Lagrange multipliers to solve constrained optimization problems (tangency between the constraint and objective function gradients!).

# Friday, 30th June 2016

(09:00 - 11:00) Solving problems from Ch. 8 of *Computer Vision*. Deriving the t-distribution from the gamma and normal distributions. Deriving the M-step values for the mean and covariance values in a Gaussian mixture model. (11:00 - 12:40) Nonlinear logistic regression - as with linear regression, we simply apply the nonlinear transformation to the inputs. Fitting the model requires a nonlinear optimization method - fitting these models is usually done using maximum likelihood, since prior modeling over both the gradient vector  $\phi$  and the parameters of the nonlinear function requires a lot of effort.

I also made notes on the dual logistic regression classifier, and how it can be fit using either maximum likelihood or Bayesian inference. The latter requires the Laplace approximation - approximating a distribution using a normal distribution with equal mean and second derivative at the mean to the actual distribution being modeled. Fitting logistic regression or dual logistic regression models in a Bayesian manner requires that the true parameter ( $\phi$  or  $\varphi$ ) posterior distribution's parameters be approximated by their equivalents in the maximum a posteriori fit.

## 1 Meeting with MM

Topics of conversation:

- PV panel classification problem : identifying cracks in scans of PV panels.
- Questions:
  - Very small dataset problem: 50:50 train/test, split the training data 70:30 into training/validation sets (i.e. I have very poor resolution on what the generalization error of the models I'm fitting is likely to be).
  - What preprocessing can I use to enhance the cracks?
    - \* The images have been equalized, resized for intensity classification.
    - \* Augmented the training images to expand the dataset size.
    - \* Did not augment testing images.
    - \* I've looked at using HOGs

*Friday, 30th June 2016*

- \* Other edge/feature detectors?
- \* Possible idea: Cluster panels according to model, subtract an undamaged panel from the input?
- Are there any papers that he would recommend?
- Other methods for expanding the amount of data available?
- Wind turbines:
  - \* Meeting the inspections team next week
  - \* ET wants me to focus on using visible spectrum data
  - \* Use stills, as opposed to sequenced videos
  - \* Need to label the data - recommendations on how to do this?
  - \* I will email him images of the turbines
- General placement enquiries:
  - \* What can I do that would be useful/what problems would be tractable for the placement, but would still be interesting/novel?
  - \* Does he think that doing an (admittedly small) piece of original research is possible in this timeframe?
  - \* I'm interested in synthetic data generation - does this sound like a suitable area to focus my efforts?